An Experimental Evaluation of Inversion- and Transposition-Based Genomic Distances through Simulations

Moulik Kothari[1] and Bernard M.E. Moret[1,2,3]

Abstract— **Rearrangements of genes and other syntenic blocks have become a topic of intensive study by phylogenists, comparative genomicists, and computational biologists: they are a feature of many cancers, must be taken into account to align highly divergent sequences, and constitute a phylogenetic marker of great interest. The mathematics of rearrangements is far more complex than for indels and mutations in sequences. Inversions have been well characterized through 20 years of work, but transpositions still await comparable results. We can compute inversion and DCJ (a combination of inversions and block exchanges) distances, and bounds on the transposition distance. The first has been extensively used in comparative genomics and phylogenetics, the second is quite new, and the third has not seen significant use to date.**

We present here a detailed experimental study of these three distance measures within the context of genome comparison (pairwise distances) and phylogenetic reconstruction. We used data generated through simulated evolution along various trees, using various evolutionary rates and various mixes of inversions and transpositions. Our main finding is that inversion and DCJ measures return very similar results even on data generated using only transpositions, while the measure based on Hartman's bound is often too loose to provide comparable accuracy in genomic comparisons or phylogenetic reconstruction.

I. Introduction

Rearrangements of genes and other syntenic blocks have become a topic of intensive study by phylogenists, comparative genomicists, and computational biologists. Rearrangements are a key feature of many cancers, must be taken into account to align sequences that have seen high divergence, and constitute in themselves a phylogenetic marker of great interest. Unfortunately, the mathematics of rearrangements is far more complex than that of simple indels and mutations in sequences. The rearrangement known as inversion has been well characterized through nearly 20 years of work, but the equally important transposition still awaits comparable results. At present, we have the means of computing inversion distances and so-called DCJ distances (a combination of inversions and block exchanges), as well as to derive fairly tight bounds on the transposition distance. The first has been extensively used in both comparative genomics and phylogenetics, but the second is quite new and the third has not seen significant use to date. In particular, nothing is known on the importance of taking into account in some explicit manner the possibility that a sequence of rearrangements may include transpositions and not just inversions.

We therefore set out to investigate the effect of mixing transpositions and inversions, up to the extreme of using only transpositions, on the usefulness of these three measures for phylogenetic reconstruction. For this purpose, we generated simulated data in the accepted manner (birth-death trees and published trees, various deviations from ultrametricity, various tree diameters, and various ratios of inversions to transpositions, including zero), then computed pairwise distances among the resulting gene orders, and finally computed several measures on these gene orders. These measures included the quality of trees reconstructed from the pairwise distances using neighbor-joining, the sum of the branch lengths of these trees, the sum of the pairwise distances, and the monotonicity of the distance estimates (as defined by the number of inverted pairs of values in a sorted list of these pairwise distances).

Since the DCJ measure uses only block interchanges (including transpositions) in a transposition-only scenario, whereas the inversion measure suffers from a complete mismatch of operations, one might expect the DCJ measure and the transposition bound to outperform the inversion measure in such a scenario. Somewhat surprisingly, we found that inversion and DCJ distances are nearly indistinguishable under our various measures and that the transposition bound is only slightly different. On the other hand, since all three measures are based (loosely) on the same construct (the breakpoint graph), the similarity might have been expected. Note that our results do not imply that solving the problem of sorting by transpositions (or by transpositions and inversions) would yield little or no benefit: we looked only at the effect of the choice of distance computation on various measures related to tree reconstruction; should one desire to reconstruct ancestral arrangements, for instance, the distinction between inversions and transpositions would become crucial.

II. Preliminaries

We provide a very short introduction to gene rearrangements and their uses in comparative genomics and phylogenetic reconstruction. Further information on the biological motivation for studying rearrangement can be found in the following: the surveys of Moret *et al.* [1], [2] address the uses of rearrangements in phylogenetic reconstruction; papers by Pevzner and colleagues [3], [4], [5], [6] illustrate the uses of computational

[1]Department of Computer Science, University of New Mexico, Albuquerque, NM 87131, USA

[2]School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

[3]Swiss Institute of Bioinformatics, CH-1015, Lausanne, Switzerland

approaches to rearrangements in comparative genomics; many other researchers are developing methods based on gene orders to assign gene orthologies (cf. [7], [8], [9]); and studies of rearrangements in many types of cancer (thyroid, breast, etc.) abound.

A. Gene-order data

With the advent of high-throughput sequencing and thanks to the large number of researchers engaged in gene hunting and annotation, the complete ordering of genes on each chromosome of a large number of organisms has become available; many more (especially bacteria and archea) will become available shortly. These data have much to offer to the biomedical community, but tools for analyzing them lag behind tools for sequence analysis, in part of course because of nearly 20 years' worth of additional efforts on sequence-based tools, but also because of the intrinsically more complex mathematics governing these new data.

In a typical setup, homologous gene sequences (or other syntenic blocks of interest) have been identified by conventional methods and gene families can thus be set up, with homologous families in other organisms. If we index all distinct (non-homologous) gene families from 1 to m, then each gene can be denoted simply by the index of its family, so that the list of genes along a chromosome becomes a list of indices, each with a value between 1 and m[1]; if genes are on different strands, we give a positive sign to the indices of genes read from the 3' to 5' end and a negative sign to those on the other strand. In the simplest case (unlikely to be encountered even in simple organelles), each gene family has a single member and then each chromosome becomes a signed permutation of a subset of the indices in $\{1, \ldots, m\}$; as a further simplification, we could also assume that each genome contains exactly the same genes and has a single chromosome, in which case each chromosome is now a signed permutation of the set $\{1, \ldots, m\}$. If we assume that the indices are such that the last common ancestor (LCA) of our samples was characterized by the identity permutation $1, 2, \ldots, m$, the most basic problem becomes "how did the new signed permutations come into being?"

B. Rearrangements

We can focus on the outcomes or on the mechanism. Researchers first focused on outcomes: synteny, for instance, is a simple measure of the result of evolution [10]—we test whether elements i and $i+1$ remain on the same chromosome on which they were located in the LCA. Adjacency is a more demanding version of the same idea—if elements i and $i+1$ are adjacent in one genome, we test whether they are also adjacent in another. The loss of adjacency manifests itself by the presence of a *breakpoint*—two indices adjacent in our modern genome that were not adjacent in the LCA. Sankoff [11], [12] advocated the use of breakpoints in evaluating rearrangements, characterized many of their aspects, and provided early software for their use. To move from results to explanations, however, we need models of evolution in terms of rearrangement operations.

Rearrangement operations may move chromosome segments within the same chromosome or among chromosomes. In the former category, the simplest operations are inversions, which excise a segment and put it back in the position, but in the reverse direction, and transpositions, which move a segment from one position to another along the chromosome. Hannenhalli and Pevzner [13], [14] provided a framework in which to study inversions and characterized the shortest sequence of inversions between two arbitrary signed permutations, giving a polynomial-time algorithm to compute it; later, Bader et al. [15] gave a linear-time algorithm to compute the resulting edit distance. While Bafna and Pevzner [16] gave a similar framework in which to study transpositions as well as a 1.5-approximation algorithm, the problems of finding a shortest edit sequence or of computing the edit distance remain open to this day. The best results to date are due to Hartman, who gave a simplified 1.5-approximation algorithm [17], later improved it to a 1.375-approximation [18]. The more general operation of block interchange, which exchanges two non-overlapping segments (and of which transposition is a special case) has also been studied and can be solved in polynomial time [19].

In the second category are operations that alter the gene content of chromosome and may even alter the number of chromosomes. These operations including fusion (of two chromosomes into one), fission (of one chromosome into two), and translocation (of a segment of one chromosome to some location of another). Recently, Yancopoulos et al. [20] proposed a unifying operation, the double-cut-and-join or DCJ; their work was further elaborated by Bergeron et al. [21]. A DCJ operation, as its name indicates, makes a pair of cuts (which can be anywhere, including one cut each on two different chromosomes) and proceeds to reglue cut ends; the repair can be done in several ways and can yield an identity, an inversion, a fission (one linear chromosome into two chromosomes, one linear and one circular, for instance), a fusion (a circular chromosome and another fused), and a translocation (exchanging tails between two linear chromosomes). Combining two DCJ operations can create a block interchange and thus, in particular cases, a transposition. While there is no direct biological evidence for DCJ operations in their full generality, the unifying model they present, the natural way in which they equate the cost of one transposition with that of two inversions (a valuation long preferred by the community, but always artificially imposed), and the fact that computations with DCJ are even simpler than computations with just inversions, have all combined to make this model very attractive.

C. Distance measures

While one would like the actual amount of evolution (in terms of the rearrangements operations permitted under the

[1]This indexing assumes that genes are linearly ordered along a chromosome, which is not always the case—in fact, one gene can be nested within another; we assume the simpler, linearly ordered case, which is why it may make more sense to think of these orderings as orderings of blocks of consecutive codons.

model) that separates two genomes, one has to settle for computing metrics such as the smallest number of permitted operations that can transform one genome into the other, known as an edit distance. These edit distances naturally underestimate the amount of evolution, especially for more divergent genomes—a problem common to such computations for sequence data as well. Thus various corrections, which apply model statistics to produce an expected value for the amount of evolution indicated by the edit distance, have been proposed (see, e.g., [22], [23]) and shown to improve the accuracy of phylogenetic reconstruction. However, such corrections have so far only been proposed for inversion metrics; the complexity of the DCJ model makes it difficult to produce model statistics (we would need to know more about the parameters of the various rearrangements it encompasses) and the elasticity of the transposition bounds quickly magnify the principal drawback of such corrections, which is the large increase in variance for the expectation produced. Therefore our study focussed on uncorrected distances only. We used the 1.5-approximation of Hartman, rather than his later 1.375-approximation, in part due to the simplicity of the former, but also because the fractions only denote a worst-case behavior and there was little reason to assume that the second algorithm would do better on average than the first.

III. Experimental Design

Our aim is to evaluate the relative merit of these three distance estimates (inversion, DCJ, and transposition, the latter through Hartman's bound) in phylogenetic and comparative work.

A. Comparison strategies

We know that all three measures (certainly the first two, which are true edit distances) will typically underestimate the true evolutionary distance, but we are interested in possible differences in these underestimates. On the other hand, because rearrangement paths using only transpositions necessarily differ greatly from rearrangement paths using only inversions, there is no point in comparing reconstructed ancestral genomes in phylogenies. Thus our experiments focus on the relative performance of the three methods in estimating distances and the differences observed in the quality of reconstructed trees.

We chose four measures to compare the three distance estimates.

- The sum of all pairwise distances. We can compare this quantity to the sum of all true pairwise distances (known from the simulation) and thus see whether the underestimate is more pronounced with some of the distances than with others.
- The relative ordering of the pairwise distances. We can sort the true pairwise distances and, for each of the three distance estimates, compare its sorted order with the true one. We used the number of pairwise index inversions (as needed in a bubblesort) as the indicator.
- The sum of all tree branch lengths in a tree reconstructed from the distance matrix by the neighbor-joining (NJ) method [24]. This measure avoids the need for ancestral reconstructions, yet gives us a quick estimate of how much overall evolution is implied for the dataset under each distance estimate and how it compares to the true total amount of evolution (known from the simulation).
- The Robinson-Foulds (RF) distance [25] between the model tree and the tree reconstructed under each distance estimate with the NJ method. This measure of topological error is the most widely used in phylogenetic work and, within the limits of the chosen reconstruction method, tells us how well each distance estimate lends itself to phylogenetic analysis.

B. Generating sample data

Because DCJ distances reduce to inversion distances on data generated using only inversions and because transpositions remain the unknown factor, we generated data with large numbers of transpositions; indeed, most of our tests used data generated with transpositions only. On such data, DCJ methods will not postulate any inversions, Hartman's bounds will be well matched, and inversion methods will presumably be at their worst—although earlier studies indicated that inversion distances performed well under most mixes of inversions, transpositions, and inverted transpositions [26], [27].

We first generate sample model trees with evolutionary distances on all branches, using the standard `r8s` software [28]. The edge lengths are set such that the diameter of the tree, the largest distance between any two leaves on their connecting path in the tree, is a predetermined factor of the number of genes. Since these trees are ultrametric (the path length from root to leaf is the same for all leaves), we alter branch lengths to deviate the tree from ultrametricity by a prescribed factor, following a standardized procedure (see, e.g., [27]). For each branch in the model tree, we pick a random number s, where $\ln(s)$ is drawn from the interval $[-c, c]$, where c is the chosen deviation from ultrametricity, set to 5 in our experiments (a rather severe deviation). We then multiply the expected branch length by s to obtain a new branch length for the true tree; finally, in order to avoid branches of zero length (which could artificially confuse the NJ algorithm and would, in any case, force it to return false positive edges), we add 2 to the length of every resulting edge. For each of the model trees, we generate 25 replicates (which will have identical edge length distributions, but resampled). We generated in this manner trees of 50, 100, and 200 leaves.

On each tree thus obtained, we generate gene orders (permutations) for the leaves, which are then stored in a FASTA file for processing by our distance estimators. In all our experiments, we used artificial genomes of 400 genes: the number is large enough to accommodate larger trees and to present complex interactions among rearrangements, yet small enough to allow for large-scale experimentation. Leaf genomes are generated by starting with the identity permutation at the root and traversing down to the root, generating each child of a node by applying to that node's genome a number of rearrangement operations equal to the length of the edge to the child. The rearrangement operations are selected in the chosen proportion of inversions to transpositions (we show only

results for 100% transpositions) and each operation is specified independently at random by picking two (for inversions) or three (for transpositions, with the third value outside the range defined by the first two) integers in the range from 1 to 400.

C. Running the experiments

The outcome of the previous step is a collection of true trees (with edge lengths) and of FASTA files (which contain the leaf genomes). We then compute a pairwise distance matrix for each FASTA file with each distance estimate and run the NJ algorithm on that matrix, recording the resulting tree and its edge lengths, and calculating the four measures described in Section III-A. Because DCJ could use block exchanges rather than actual transpositions (there is no way to forbid block exchanges without losing the advantages of the DCJ model), we also measured the percentage of block exchanges that were actually transpositions in the DCJ distance calculations; ideally, this percentage should be 100, since there is no biological evidence for block exchanges rather than transpositions at the genomic level. Finally, we averaged the results over the 25 replicates of each model tree.

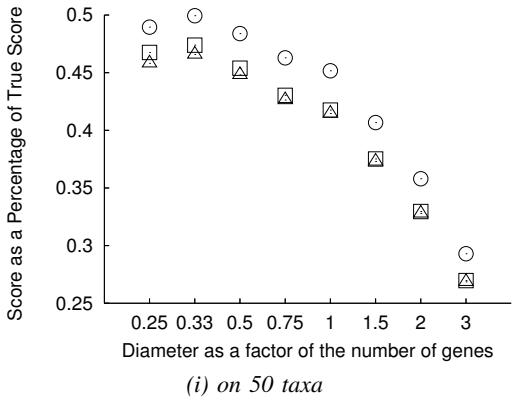
IV. Results and Discussion

We present and discuss results under each of the four measures of Section III-A in turn. Lack of space prevents us from showing all of our results; moreover, many are repetitive, so we present only a sample.

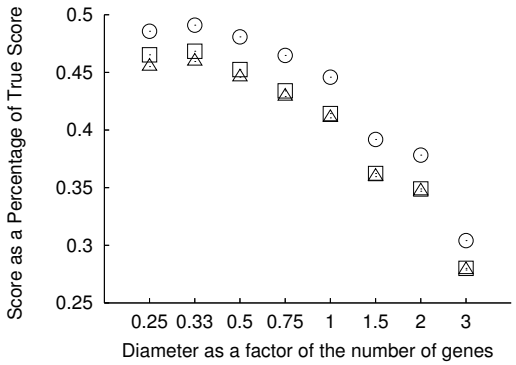
A. Sum of all pairwise distances

Figure 1 shows the sum of all pairwise distances returned by each distance estimate (as a percentage of the true sum) plotted against increasing tree diameters. In this and all plots (except for Figure 2), the data are shown for trees of 50 taxa on 400 genes, with tree diameters varying from 100 to 1200 rearrangements, and inversion estimates are shown with squares, Hartman's bound with circles, and DCJ estimates with triangles. Predictably, as the tree diameter increases, the underestimation gets worse. The three methods give very tightly clustered values, with DCJ slightly lower and, in most cases, Hartman's bound slightly higher than inversion. This tight clustering further supports the assumption that a transposition should be given the weight of two inversions.

We carried out a series of experiments for genomes of varying sizes, from 50 to 400 genes, in order to determine where the three distance estimators first began to underestimate the true distance—the beginning of saturation for our distance estimates. With most distance measures, reconstruction is far more accurate below saturation than past saturation; NJ reconstruction from these distance matrices should be no exception. Figure 2 shows plots for genomes of 50, 150, and 200 genes. In these plots, the true distance corresponds to the diagonal, shown as a thin line. Hartman's bound unsurprisingly overestimates true distances for small values, then closely follows true distances up to almost $\frac{n}{3}$ transpositions, where n is the number of genes. Both inversion and DCJ distances closely follow true distances up to about $\frac{n}{5}$ transpositions. These results are similar to those obtained by Moret et al. [22] for inversion distances under various scenarios.



(i) on 50 taxa



(ii) on 100 taxa

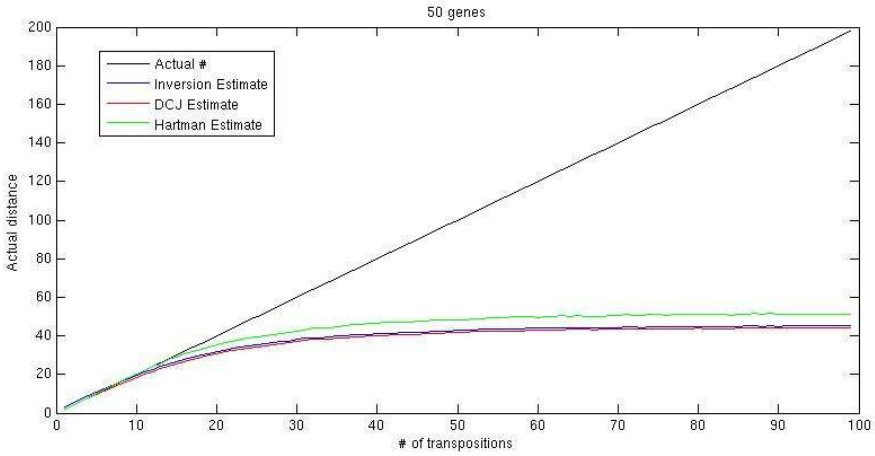
Fig. 1. Sum of all pairwise distances as a percentage of the true sum on 50 (top) and 100 (bottom) taxa and for various true tree diameters.

B. Bubblesort exchanges

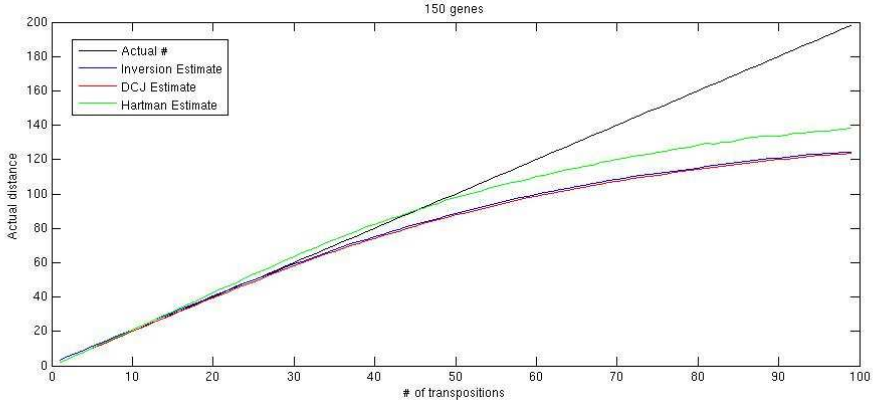
Figure 1 shows that the three methods perform very closely when we look at the entire tree. However, we also wanted to check how uniform the difference in distance estimation is between the true distances and the method being considered. We looked at the number of bubblesort exchanges it takes to reorder the sorted pairwise distances returned by one method so as to agree with the order obtained from the true pairwise distances. Figure 3 shows the results; to make comparisons easier, the results in the plot are normalized by the number of taxa. Again, the three methods perform very closely with DCJ marginally below inversion and Hartman's approximation usually above. What is more interesting is that the ordering is less sensitive to the diameter than the sum of all pairwise distances, hardly deviating from the true ordering up to almost $\frac{n}{2}$ transpositions. This is good news for tree reconstruction, as the relative ordering of pairwise distances is more important for accuracy than the absolute values.

C. Sum of tree edges

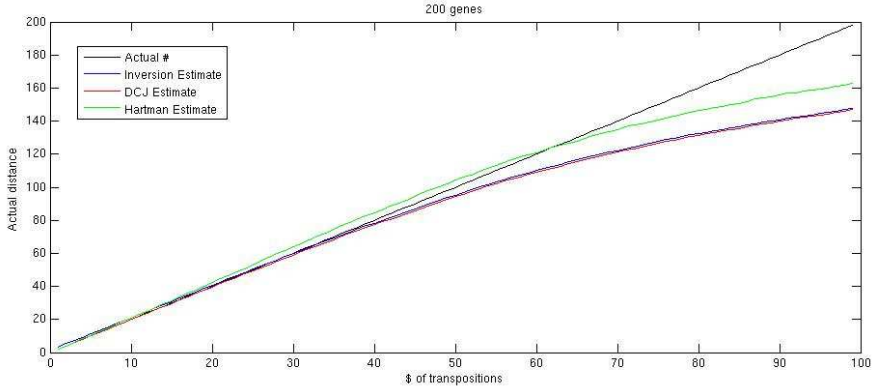
The next measure we look at is the sum of all edge lengths for the trees generated by NJ from the distance matrices computed under each distance estimate. Once again, we expect that measure to underestimate the sum of the true edge lengths of the true tree. Figure 4 shows these values as percentages of the true sum. The same pattern observed earlier continues, with all three distance estimates yielding very close values



(i) on 50 genes

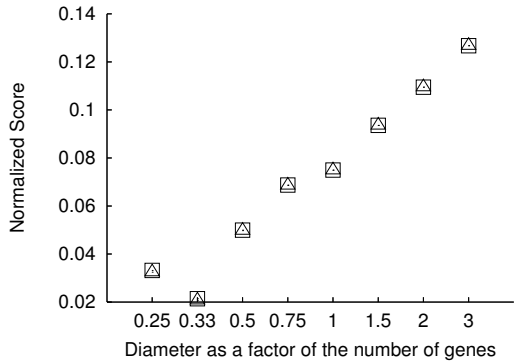


(ii) on 150 genes

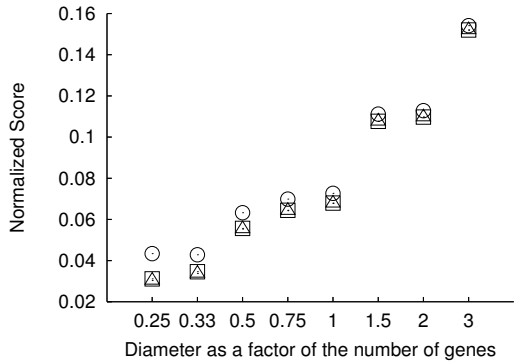


(iii) on 200 genes

Fig. 2. Pairwise inversion, DCJ, and Hartman's distance estimates vs. true distances for genomes of 50, 150, and 200 genes (from top to bottom).

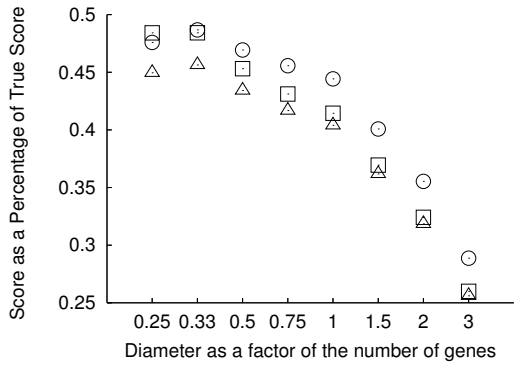


(i) on 50 taxa

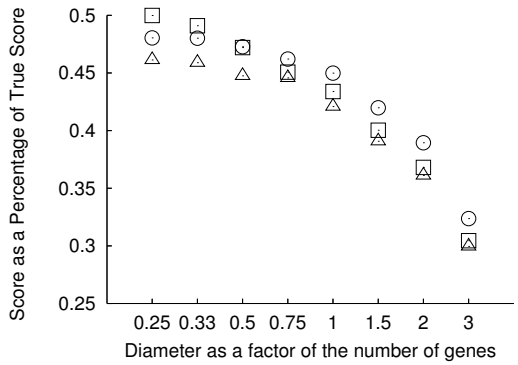


(ii) on 100 taxa

Fig. 3. Normalized bubblesort exchanges on 50 (top) and 100 (bottom) taxa and for various true tree diameters.



(i) on 50 taxa



(ii) on 100 taxa

Fig. 4. Sum of all branches for trees returned by NJ on 50 (top) and 100 (bottom) taxa and for various true tree diameters.

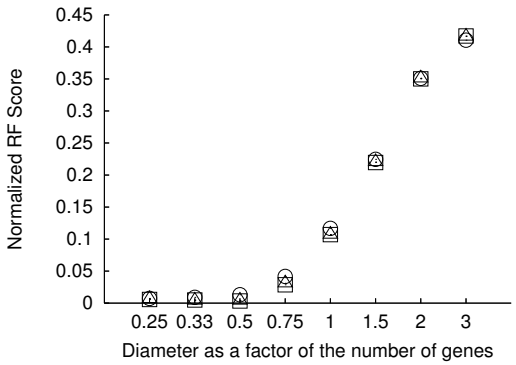
and with DCJ marginally lower than inversion and Hartman's approximation slightly larger than the other two.

D. Robinson-Foulds metric

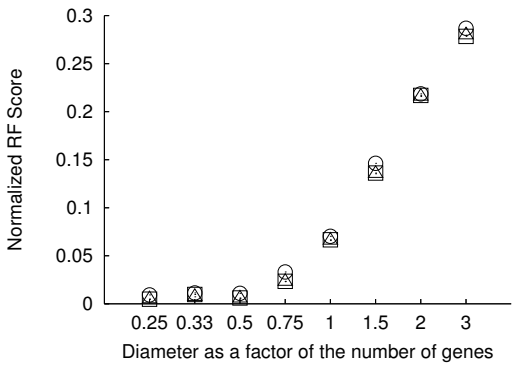
From what we have seen so far, it is clear that the DCJ edit distance tends to yield the smallest estimate, closely followed by the inversion edit distance. Now, inversion distances have been used successfully in tree reconstruction, so we should expect that DCJ distances would be equally successful. We now compare the tree structures returned by NJ run on the pairwise distance matrices compared to the true tree, using the Robinson-Foulds metric. Normalized RF values are the sums of the number of false positive edges and the number of false negative edges divided by (twice) the number of edges in the true tree and thus vary from 0 (perfect) to 1 (completely wrong). Figure 5 plots normalized RF scores as a function of tree diameter for 50 and 100 taxa. As expected from our results on the relative ordering of pairwise distances, the reconstructions are quite accurate up to a diameter of about $\frac{n}{2}$ transpositions and acceptable to a diameter of n. Once again, the results for the three distance estimates are barely distinguishable.

E. Transpositions vs. block interchanges

The DCJ algorithm first attempts to find possible inversions, but only if oriented edges can be found in the breakpoint graph. Since data generated using only transpositions will produce breakpoint graphs without any oriented edges, DCJ will not



(i) on 50 taxa



(ii) on 100 taxa

Fig. 5. Normalized RF scores on 50 (top) and 100 (bottom) taxa and for various true tree diameters.

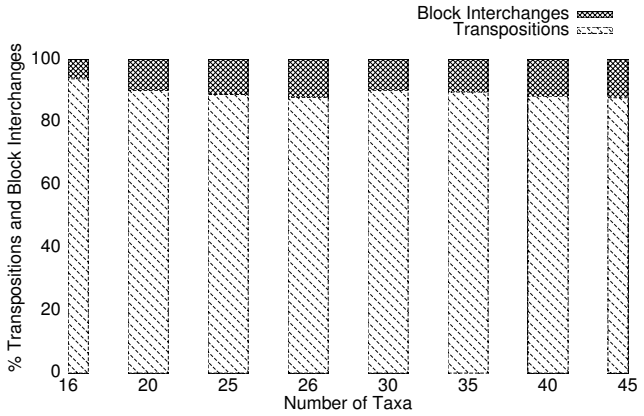


Fig. 6. Percentage of transpositions and block interchanges in DCJ.

infer any inversions in a sorting sequence for such data. However, it will infer both general block interchanges and transpositions, the special case in which one of the two blocks has zero length. Since transpositions are well documented in genomics, but general block interchanges are not, we would prefer to see the DCJ method infer as many transpositions as possible. Figure 6 shows the proportions of transpositions and block interchanges in the sum of all edge lengths (all operations, times two) for each tree. It is gratifying to see that between 85% to 95% of block interchanges are actually transpositions.

F. General discussion

The striking similarity of the three distance estimates under the four measures we used confirms the widely held view that the "cost" of a transposition or block exchanges should be taken as twice that of an inversion. That is because we used such a cost for block exchanges and transpositions in our calculations and the (very different) sorting paths for inversions and for transpositions came to almost exactly the same cost. The fact that good reconstructions (under NJ) are possible for diameter of up to n confirms findings from our group regarding inversion distances: gene-order data is generally more "resistant" to saturation than sequence data, due to the huge number of character states available. Finally, the fact that the DCJ approach inferred a sorting sequence composed mostly of pure transpositions for data generated with transpositions only shows that, except in order to reconstruct ancestral sequences, we may not need to "solve" the transposition problem in order to produce accurate phylogenies.

V. Concluding Remarks

Because DCJ distances and sorting sequences are as easy to compute as inversion distances and sorting sequences, yet DCJ operations are more general, applying as they do to multichromosomal genomes, it makes sense to replace inversion computations by DCJ computations. However, there remain a number of significant issues.

First, the issue of distance correction: as shown by Moret et al. [22], correcting inversions distances yields considerably more accurate results in phylogenetic work; there is no doubt that DCJ distances would benefit just as much from a correction, but this correction will be much more complex to design and would appear to require a large number of assumptions about basic parameters of the evolutionary process.

Second, the assumption of equal gene content and no duplicates is clearly unrealistic, especially for larger genomes. Some significant advances have been made with unequal gene content and duplications in an inversion context [29], [30], [31], but, once again, the large variety of operations possible with DCJ will make similar work very challenging.

Finally, the elusive goal of ancestral reconstruction will not be made any more reachable by the DCJ model: what has prevented us so far from reconstructing good ancestral gene orders under inversion only is the lack of biological knowledge to constrain what appears to be an enormous choice of equally good possibilities. The more flexible DCJ model can only make that choice even larger. The various tantalizing issues arising with inversions (such as the apparent prevalence of short over long inversions [32], [33] and the apparent presence of hotspots in certain chromosomal loci [5]) are likely to grow in number and complexity with a DCJ model.

VI. Acknowledgments

This work supported in part by the US National Science Foundation under grants EF 03-31654 (the CIPRES project), IIS 01-21377, and DEB 01-20709.

References

[1] B. Moret, J. Tang, and T. Warnow, "Reconstructing phylogenies from gene-content and gene-order data," in *Mathematics of Evolution and Phylogeny*, O. Gascuel, Ed. Oxford University Press, 2005, pp. 321–352.

[2] B. Moret and T. Warnow, "Advances in phylogeny reconstruction from gene order and content data," in *Molecular Evolution: Producing the Biochemical Data, Part B*, ser. Methods in Enzymology, E. Zimmer and E. Roalson, Eds. Elsevier Publishers, 2005, vol. 395, pp. 673–700.

[3] G. Andelfinger, C. Hitte, L. Etter, R. Guyon, G. Bourque, G. Tesler, P. Pevzner, E. Kirkness, F. Galibert, and D. Benson, "Detailed four-way comparative mapping and gene order analysis of the canine ctvm locus reveals evolutionary chromosome rearrangements," *Genomics*, vol. 83, pp. 1053–1062, 2004.

[4] P. Pevzner and G. Tesler, "Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution," *Proc. Nat'l Acad. Sci., USA*, vol. 100, no. 13, pp. 7672–7677, 2003.

[5] ——, "Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution," *Proc. Nat'l Acad. Sci., USA*, vol. 100, no. 13, pp. 7672–7677, 2003.

[6] G. Tesler, "Efficient algorithms for multichromosomal genome rearrangements," *J. Comput. Syst. Sci.*, vol. 65, no. 3, pp. 587–609, 2002.

[7] G. Bourque, Y. Yacef, and N. El-Mabrouk, "Maximizing synteny blocks to identify ancestral homologs," in *Proc. 3rd RECOMB Workshop Comparative Genomics (RECOMBCG'05)*, ser. Lecture Notes in Computer Science, vol. 3678. Springer Verlag, 2005, pp. 21–34.

[8] X. Chen, J. Zheng, Z. Fu, P. Nan, Y. Zhong, S. Lonardi, and T. Jiang, "Computing the assignment of orthologous genes via genome rearrangement," in *Proc. 3rd Asia-Pacific Bioinformatics Conf. (APBC'05)*. Imperial College Press, London, 2005, pp. 363–378.

[9] K. Swenson, N. Pattengale, and B. Moret, "A framework for orthology assignment from gene rearrangement data," in *Proc. 3rd RECOMB Workshop Comparative Genomics (RECOMBCG'05)*, ser. Lecture Notes in Computer Science, vol. 3678. Springer Verlag, 2005, pp. 153–166.

[10] D. Sankoff and J. Nadeau, "Conserved synteny as a measure of genomic distance," *Disc. Appl. Math.*, vol. 71, no. 1–3, pp. 247–257, 1996.

[11] M. Blanchette, G. Bourque, and D. Sankoff, "Breakpoint phylogenies," in *Genome Informatics*, S. Miyano and T. Takagi, Eds. Tokyo: Univ. Academy Press, 1997, pp. 25–34.

[12] D. Sankoff and M. Blanchette, "Multiple genome rearrangement and breakpoint phylogeny," *J. Comput. Biol.*, vol. 5, pp. 555–570, 1998.

[13] S. Hannenhalli and P. Pevzner, "Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals)," in *Proc. 27th Ann. ACM Symp. Theory of Comput. (STOC'95)*. ACM Press, New York, 1995, pp. 178–189.

[14] ——, "Transforming mice into men (polynomial algorithm for genomic distance problems)," in *Proc. 36th Ann. IEEE Symp. Foundations of Comput. Sci. (FOCS'95)*. IEEE Press, Piscataway, NJ, 1995, pp. 581–592.

[15] D. Bader, B. Moret, and M. Yan, "A fast linear-time algorithm for inversion distance with an experimental comparison," *J. Comput. Biol.*, vol. 8, no. 5, pp. 483–491, 2001.

[16] V. Bafna and P. Pevzner, "Sorting permutations by transpositions," in *Proc. 6th Ann. ACM/SIAM Symp. Discrete Algs. (SODA'95)*. SIAM Press, Philadelphia, 1995, pp. 614–623.

[17] T. Hartman, "A simpler 1.5-approximation algorithm for sorting by transpositions," in *Proc. 14th Ann. Symp. Combin. Pattern Matching (CPM'03)*, ser. Lecture Notes in Computer Science, vol. 2676. Springer Verlag, 2003, pp. 156–169.

[18] I. Elias and T. Hartman, "A 1.375-approximation algorithm for sorting by transpositions," in *Proc. 5th Int'l Workshop Algs. in Bioinformatics (WABI'05)*, ser. Lecture Notes in Computer Science, vol. 3692. Springer Verlag, 2005, pp. 204–215.

[19] D. Christie, "Sorting permutations by block-interchanges," *Inf. Process. Lett.*, vol. 60, no. 4, pp. 165–169, 1996.

[20] S. Yancopoulos, O. Attie, and R. Friedberg, "Efficient sorting of genomic permutations by translocation, inversion and block interchange," *Bioinformatics*, vol. 21, no. 16, pp. 3340–3346, 2005.

[21] A. Bergeron, J. Mixtacki, and J. Stoye, "A unifying view of genome rearrangements," in *Proc. 6th Int'l Workshop Algs. in Bioinformatics (WABI'06)*, ser. Lecture Notes in Computer Science, vol. 4175. Springer Verlag, 2006, pp. 163–173.

[22] B. Moret, J. Tang, L.-S. Wang, and T. Warnow, "Steps toward accurate reconstructions of phylogenies from gene-order data," *J. Comput. Syst. Sci.*, vol. 65, no. 3, pp. 508–525, 2002.

[23] L.-S. Wang and T. Warnow, "Distance-based genome rearrangement phylogeny," in *Mathematics of Evolution and Phylogeny*, O. Gascuel, Ed. Oxford University Press, 2005, pp. 353–383.

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

[25] D. Robinson and L. Foulds, "Comparison of phylogenetic trees," *Mathematical Biosciences*, vol. 53, pp. 131–147, 1981.

[26] B. Moret, A. Siepel, J. Tang, and T. Liu, "Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data," in *Proc. 2nd Int'l Workshop Algs. in Bioinformatics (WABI'02)*, ser. Lecture Notes in Computer Science, vol. 2452. Springer Verlag, 2002, pp. 521–536.

[27] L.-S. Wang, R. Jansen, B. Moret, L. Raubeson, and T. Warnow, "Distance-based genome rearrangement phylogeny," *J. Mol. Evol.*, vol. 63, no. 4, pp. 473–483, 2006.

[28] M. Sanderson, "r8s: inferring absolute rates of evolution and divergence times in the absence of a molecular clock," *Bioinformatics*, vol. 19, pp. 301–302, 2003.

[29] J. Tang, B. Moret, L. Cui, and C. dePamphilis, "Phylogenetic reconstruction from arbitrary gene-order data," in *Proc. 4th IEEE Symp. on Bioinformatics and Bioengineering BIBE'04*. IEEE Press, Piscataway, NJ, 2004, pp. 592–599.

[30] M. Marron, K. Swenson, and B. Moret, "Genomic distances under deletions and insertions," *Theor. Computer Science*, vol. 325, no. 3, pp. 347–360, 2004.

[31] K. Swenson, M. Marron, J. Earnest-DeYoung, and B. Moret, "Approximating the true evolutionary distance between two genomes," in *Proc. 7th SIAM Workshop on Algorithm Engineering & Experiments (ALENEX'05)*. SIAM Press, Philadelphia, 2005.

[32] J.-F. Lefebvre, N. El-Mabrouk, E. Tillier, and D. Sankoff, "Detection and validation of single gene inversions," in *Proc. 11th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'03)*, ser. Bioinformatics, vol. 19. Oxford U. Press, 2003, pp. i190–i196.

[33] D. Sankoff, "Short inversions and conserved gene cluster," *Bioinformatics*, vol. 18, no. 10, p. 1305, 2002.