# On the tandem duplication-random loss model of genome rearrangement

Kamalika Chaudhuri * Kevin Chen † Radu Mihaescu ‡ Satish Rao §

November 1, 2005

**Abstract**

We initiate the algorithmic study of a new model of genome rearrangement, the *tandem duplication-random loss model*, in which a genome evolves via successive rounds of tandem duplication of a contiguous segment of genes, followed by the loss of one copy of each of the duplicated genes. This model is well-known in the evolutionary biology literature, where it has been used to explain many of the known rearrangements in vertebrate mitochondrial genomes. Based on the model, we formalize a notion of distance between two genomes and show how to compute it efficiently for two interesting regions of the parameter space. We then consider *median problems* (i.e. finding the point which minimizes the sum of distances to a given set of points under some distance function) in the context of maximum parsimony phylogenetic reconstruction for these two special cases. Surprisingly, one of them turns out to correspond to the well-known *rank aggregation* problem, while the other corresponds to the biologically interesting case of *whole genome duplication and loss*, and we give an $O(\log \log n)$ *additive* approximation algorithm for the latter.

## 1 Introduction

The growing availability of complete genome sequences has made it possible to study molecular evolution at the level of large-scale genome rearrangements. There is now a sizeable body of algorithmic work on computing distances between genomes under various models of genome rearrangement including inversions, translocations and chromosome fusions/fissions, and using them to reconstruct phylogenetic trees (see [1, 2] for excellent introductions to the field of gene order phylogeny).

Several attempts have been made to incorporate gene duplications and losses into these models [3, 4, 5, 6], but the resulting combinatorial problems have typically been quite difficult. Our approach is to isolate the duplication-loss problem by concentrating on the combinatorial properties of a model involving *only* duplications and losses. In the *tandem duplication-random loss model*, or simply the *duplication-loss model*, a genome evolves via the tandem duplication of a contiguous segment of genes (i.e. the duplicated copy is inserted *immediately* after the original copy), followed by the loss of one copy of each of the duplicated genes. In most, though not all, cases, this process will result in a genome rearrangement (Figure 1).

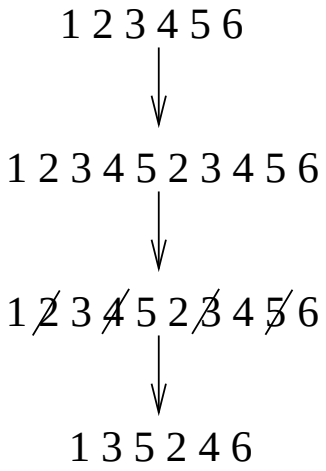


Figure 1: Example of a genome rearrangement caused by one round of tandem duplication and random loss.

There are good theoretical and empirical reasons to model the gene loss events following a tandem duplication as occuring virtually instantaneously (at least on an evolutionary time-scale), so we can think of the duplication and gene losses together as a single atomic event (readers interested in the biological basis of the model are referred to [7, 8]). Although clearly an approximation to reality, this model nonetheless gives us a clean framework in which to study the combinatorics of gene duplication and loss. In addition, the model has been well-studied in the biology literature, where it has been shown by Boore, Brown and others to be perhaps the most important rearrangement process in the case of animal mitochondrial genomes [7, 8, 9, 10, 11, 12], and has also been discussed in the computer science literature [13]. Furthermore, it is a rich model that

*CS Division, UC Berkeley. Email: kamalika@cs.berkeley.edu
†CS Division, UC Berkeley. Email: kevinc@cs.berkeley.edu
‡Dept. of Mathematics, UC Berkeley. Email: mihaescu@berkeley.edu
§CS Division, UC Berkeley. Email: satishr@cs.berkeley.edu

contains the special cases of *whole-genome duplication and loss*, which is of special interest in the study of genome evolution [14, 15, 16], small inversions and small transpositions.

In our model, we adopt the usual convention of defining a genome to be a (linear, unsigned) permutation of the integers 1 to $n$. Although it seems intuitively clear that the cost of a duplication should be some non-decreasing function of the length of the duplication, it is not clear exactly what functional form this cost function should take. In this paper, we propose a geometrically increasing cost function as one possibility: the cost of a duplication of a segment of $k$ genes is $\alpha^k$ for some constant parameter $\alpha \geq 1$. The physical intuition for this cost function is a model in which each gene following the initiation of duplication is the termination point of the duplication with uniform constant probability. Certainly other cost functions, such as an affine function, can be considered, but we do not do so here.

The first problem we consider is that of computing the edit distance: a minimum cost sequence of duplications and losses required to transform one permutation into another.

DEFINITION 1.1. *Given a permutation $\pi$ and a parameter $\alpha$, find a minimum cost sequence of duplication-loss steps required to transform the identity permutation, $\pi_I$, into $\pi$. We denote this distance $d_\alpha(\pi_I, \pi)$ or more simply, $d_\alpha(\pi)$.*

Our first main result is a solution to this problem for two interesting regions of the parameter space.

THEOREM 1.1. *Given a parameter $\alpha \geq 1$ and a permutation $\pi$,*

1. $\alpha = 1$

   *Define $\rho(\pi)$ to be the number of maximal increasing substrings in $\pi$ (e.g. $\rho(142563) = 3$). Then*

   $$d_\alpha(\pi) = \lceil \log_2 \ \rho(\pi) \rceil$$

   *The distance can be trivially computed by inspection in linear time, and it is easy to see that the case $\alpha = 1$ (i.e. all duplications equi-probable) reduces to the whole-genome duplication case.*

2. $\alpha \geq 2$

   *The Kendall-Tau distance (the bubblesort distance) is defined as*

   $$d_{KT}(\pi) = |(i,j) : i > j \text{ and } \pi(i) < \pi(j)|$$

   *Then $d_\alpha(\pi)$ is exactly the Kendall-Tau distance and can be computed in $O(n \log n)$ time.*

We prove this theorem in Section 2. In Section 3, we extend our techniques to the problem of phylogenetic tree reconstruction under our new distance measure. For the case $\alpha \geq 2$, the Kendall-Tau distance is a metric, so efficient distance-based methods, such as neighbor joining (e.g. [17]), can be directly applied. However, for the case $\alpha = 1$, we show that the distance measure is strongly asymmetric in a certain sense, making this approach infeasible. Our distance measure is the only asymmetric distance in the gene order literature to the best of our knowledge.

Based on this negative result, in Section 4, we adopt a maximum-parsimony framework instead and study median problems for our two distance measures. In the traditional undirected version of the median problem, we are given $k$ genomes and asked to find the genome which minimizes the sum of the distances to the other $k$. The median problem for the Kendall-Tau distance has previously been studied in the context of social choice theory and rank aggregation, where it is called the *Kemeny optimal ranking problem*, and shown to be NP-hard for $k \geq 4$ [18]. A recent paper gives a 1.57 approximation algorithm for the minimization version of the problem [19] and a PTAS was previously known for the maximization version [20]. For $k = 3$, the case of most interest in the context of phylogeny, since it is often used as a subroutine in phylogenetic reconstruction [1], NP-hardness has not been established, but a trivial $\frac{4}{3}$ approximation can be achieved by simply picking one of the three permutations as the median. We also note that although we concentrate in this paper on the case where all genomes have the same gene content, it is also important to study the unequal gene content case. Some of the results in the rank aggregation literature generalizing the Kendall-Tau distance to partial lists may be applicable to our problem, and our work also suggests further generalizations of of the rank aggregation problem. For example, a gene duplication in our scenario would correspond to an element with two different ranks in the same list in the context of rank aggregation.

For the case $\alpha = 1$, since the distance measure is asymmetric, we define a directed version of the median problem in which we are given two child genomes and asked to find the parent which minimizes the sum of the distances to the two children.

DEFINITION 1.2. *Given two permutations $\pi_1$ and $\pi_2$, find the permutation $\hat{\pi}$ such that $d_1(\hat{\pi}, \pi_1) + d_1(\hat{\pi}, \pi_2)$ is minimized.*

Our second main result is an approximation algorithm for this problem:

THEOREM 1.2. *There exists a polynomial-time approx-*

*imation algorithm for the directed median problem, under the* whole-genome duplication and random loss *distance measure, with* additive *error* $O(\log \log n)$.

We conclude in Section 5 with some experiments on real biological data.

## 2 Computing the Distance

In the rest of this paper, each round of tandem duplication and gene loss will be referred to as a *duplication-loss step* or, when the context is clear, simply a *step*. When we wish to specify the size of the duplication precisely, we shall refer to it as a *k-step*, and we shall specify a step by a $k$-bit vector in which the 1 bits correspond to those genes which are lost in the second copy of the genome and the 0 bits to those which are lost in the first copy (alternatively, 1 bits are the genes "selected" in the first copy and 0 bits are those "selected" in the second copy).

**2.1** $\alpha = 1$**.** For the case $\alpha = 1$, it is easy to see that it is always optimal to duplicate the entire genome in each round, so that it reduces to the *whole-genome duplication-random loss model*. In this section, we prove Theorem 1.1, which states that the distance from the identity permutation to a given permutation, $\pi$, is exactly $\lceil \log_2 \rho(\pi) \rceil$, and give an algorithm that reconstructs the duplication-loss steps on a particular shortest path between the two permutations.

Our algorithm is based on the insight that a duplication-loss step is equivalent to one step of the classical radix sort algorithm. We initially label each element of the $i^{th}$ maximal increasing substring in $\pi$ with the binary representation of $i$. We then define the $k^{th}$ duplication-loss step to be the $n$-bit vector which has 0 everywhere, except in those positions which have labels containing a 0 in their $k^{th}$ least significant bit. This series of steps is sufficient to transform the identity permutation into $\pi$. Figure 2.1 shows the steps in an example execution of the algorithm.

We now prove Theorem 1.1.

*Proof.* We prove that $d_1(\pi) \leq \lceil \log_2 \rho(\pi) \rceil$ by induction on the number of steps. We claim that after the first $k$ steps, the maximal increasing substrings are sorted with respect to the $k$ least significant bits of their labels. This is obvious for the base case $k = 1$, and the inductive step follows from the fact that sorting in step $k$ does not interfere with the sorted order produced by the previous steps (i.e. radix sort is *stable*). The elements within a particular maximal increasing substring were in increasing order in the identity permutation to start, and the sorting process does not change their order at any point, again using the stability property. This

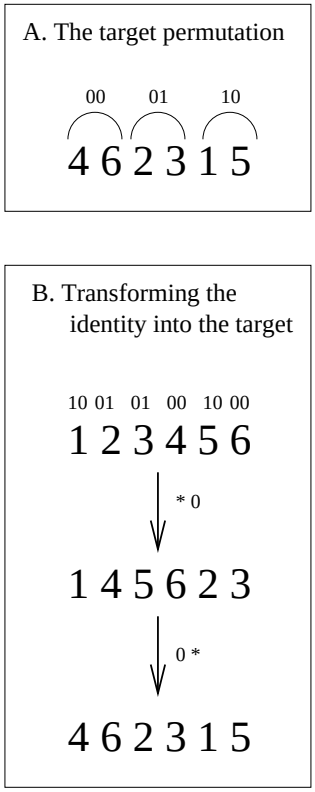


Figure 2: Example of an execution of the radix sort algorithm. A) We label each maximal increasing substring with its index in binary. B) We transfer these labels to the identity and use them to sort the identity permuation into the target permutation. The first step selects the elements whose least significant bit is 0, and the second step selects the elements whose second least significant bit is 0. The execution of the algorithm take $\lceil \log_2 3 \rceil = 2$ steps.

shows that after $\lceil \log_2 \rho(\pi) \rceil$ steps, the elements are sorted with respect to their entire labels.

To show that $d_1(\pi) \geq \lceil \log_2 \rho(\pi) \rceil$, the main insight is that a duplication-loss step can create at most two maximal increasing substrings from each maximal increasing substring of the previous permutation. Since the identity permutation has one maximal increasing substring and $\pi$ has $\rho(\pi)$ maximal increasing substrings, the proof follows.

An immediate corollary of Theorem 1.1 is that the diameter of the space of permutations is $\lceil \log_2 n \rceil$ and the worst case is achieved for an inversion of the entire genome. This proves mathematically what was previously believed by the mitochondrial genome community, namely that a pure duplication-loss model cannot accomodate inversions in a natural way.

**2.2** $\alpha \geq 2$**.** For the case $\alpha \geq 2$, we show that the distance is exactly equal to the Kendall-Tau distance and can thus be computed in $O(n \log n)$ time [18].

First observe that when we are restricted to 2-steps (duplications of size 2), the problem is exactly equivalent to bubble-sort, since the only operation possible is to swap two adjacent elements.

To complete the proof, we need the following additional lemma:

LEMMA 2.1. *If $\alpha \geq 2$, given a sequence of steps of arbitrary size, it is always possible to construct a sequence using only 2-steps of cost no greater than the original sequence.*

*Proof.* It is possible to directly simulate a k-step with two (k-1)-steps and thus to show the result by induction, but we find it clearer to give an alternative proof as follows. The idea is to simulate a k-step with multiple 2-steps. Suppose we start with the identity permutation and we perform a k-step on it, producing the permutation $\pi$. We want to compute the maximum possible Kendall-Tau distance between the identity and $\pi$. Observe that $\pi$ has exactly two maximal increasing substrings. Let $A$ be the set of elements in the first and $B$ the set of elements in the second respectively. To compute the Kendall-Tau distance, we consider all ordered pairs $(i, j)$ where $1 < i < j < k$.

No pair $(i, j)$ where $i \in A$ contributes to the Kendall-Tau distance, since by definition, no element larger than $i$ is to the left of $i$. Now for each pair $(i, j)$ where $i \in B$, the only pairs that can contribute to the Kendall-Tau distance are those where $j \in A$. Therefore the Kendall-Tau distance is at most $|A| \cdot |B|$ and this is maximized by taking $|A| = |B| = \frac{k}{2}$ if $k$ is even and $|A| = \frac{k+1}{2}, |B| = \frac{k-1}{2}$ if $k$ is odd. Therefore the maximum Kendall-Tau distance is $\frac{k^2}{4}$ if $k$ is even and $\frac{(k+1)(k-1)}{4}$ if $k$ is odd.

Now, it is optimal to use only 2-steps if for even $k$, $\alpha^k \geq \frac{k^2}{4} \cdot \alpha^2 \Rightarrow \alpha \geq (\frac{k^2}{4})^{\frac{1}{k-2}}$ and for odd $k$, $\alpha^k \geq \frac{(k+1)(k-1)}{4} \cdot \alpha^2 \Rightarrow \alpha \geq (\frac{(k+1)(k-1)}{4})^{\frac{1}{k-2}}$.

Plotting these functions for $k \geq 2$, we see that the maximum value of the function attained is 2.

## 3 Distance-based Phylogenetic Reconstruction

Given our distance computation, the natural next step is to extend the model to reconstruct phylogenies. For the case $\alpha > 2$, it is easy to see that the Kendall-Tau distance is a metric (in particular, it is symmetric) and so any distance-based phylogeny method (e.g. neighbor joining) can be applied.

The case $\alpha = 1$ is more complicated since the distance measure is asymmetric: for example, we can transform 1234 to 2413 in one step but not the reverse. On the other hand, asymmetry in and of itself is not necessarily an insurmountable problem, if it is the case that the distance measure is only "weakly" asymmetric, in the sense that asymmetric steps are rare events. In this section, we give negative results showing that this in fact, does not hold, and that our distance measure is in some sense, "strongly asymmetric".

**3.1 Probability of an Asymmetric Step** First, we consider the possibility that asymmetric steps are rare events. Recall that each whole-genome duplication-loss step is specified by an $n$-bit vector, in which the 1 bits correspond to genes in which the first copy is selected and the 0 bits to those in which the second copy is selected. How many of these $2^n$ steps are asymmetric? We give an exact combinatorial characterization of the steps that are asymmetric, and show that the number of such steps grows exponentially in the size of the duplicated region (i.e. in $n$). To do this, we will in fact prove a stronger lemma, which characterizes exactly how asymmetric a step is, in terms of the structure of the step:

LEMMA 3.1. *Let $\pi_2$ be derived from $\pi_1$ by one duplication-loss step, $s$. Consider the longest subsequence of $s$ of the form $\{01\}*$. (Note that here we consider* subsequences*, which are not necessarily consecutive positions in the string, while in Theorem 1.1 we consider* substrings*, which must be consecutive.) Let $r(s)$ be the number of repetitions of $\{01\}$. Then $d_1(\pi_2, \pi_1) = \lceil \log_2[r(s) + 1] \rceil$.*

*Proof.* If $\pi_2$ is derived from $\pi_1$ by step $s$, then the number of maximal increasing substrings of $\pi_1$ with respect to $\pi_2$ is precisely $r(s) + 1$. The lemma follows

by application of Theorem 1.1.

From this lemma, we see that a duplication-loss step is asymmetric if and only if it contains a 0101-subsequence. This result allows to compute the probability of a symmetric step:

COROLLARY 3.1. *Let $\Sigma_n$ be the sum of the first n natural numbers. The number of steps that are symmetric is $\Sigma_n + (\Sigma_{n-3} + \Sigma_{n-2} + \ldots + \Sigma_1)$.*

*Proof.* Suppose a step does not contain a 0101-subsequence. Call a maximal contiguous substring of 1's a *block*. Clearly the step cannot contain more than two blocks, so there two cases to consider: either there is one block, or there are two blocks but the first one starts at the first position in the step. The number of steps with one block is $1 + 2 + \ldots + n = \Sigma_n$. The number of symmetric steps with two blocks is $\Sigma_{n-3} + \Sigma_{n-2} + \ldots + \Sigma_1$.

This result shows that the number of symmetric moves is only $O(n^3)$ out of a possible $2^n$ moves, so if in our evolutionary process the gene losses occur uniformly at random, the process is asymmetric with exponentially high probability.

## 4 Maximum Parsimony Phylogenetic Reconstruction

Motivated by the negative results of the previous section, we approach the problem of phylogenetic reconstruction for the whole-genome duplication case within a maximum parsimony framework. We start with the simplest case of the maximum parsimony problem, the directed median problem: given two genomes, find the parent genome which minimizes the sum of the distances to these two children. [1]

Undirected medians are frequently used as a subroutine in phylogenetic tree reconstruction (e.g. [5]). They have been shown to be NP-hard for the breakpoint and inversion distances [21, 22], and *multiplicative* constant factor approximations are known for the breakpoint median problem [23]. [2] As noted in the introduction, the median problem for three genomes under the Kendall-Tau distance is not known to be NP-complete, and the best known algorithm achieves an approximation factor of $\frac{4}{3}$.

We first note that the trivial *leaf lifting algorithm* which labels the parent node with one of the two children achieves the worst possible approximation factor. Consider the following example:

$$\begin{aligned}\pi_1 &= 1, 2, 3, \ldots, n \\ \pi_2 &= 1, \frac{n}{2}+1, 3, \frac{n}{2}+3, \ldots, 2, \frac{n}{2}+2, \ldots, \frac{n}{2}, n\end{aligned}$$

Applying Theorem 1.1, we see that both children are at distance $\log_2 n - 1$ from each other (note that $\pi_2$ is its own inverse). However, there exists a parent node

$$\hat{\pi} = 1, \frac{n}{2}+1, 2, \frac{n}{2}+2, \ldots, \frac{n}{2}, n$$

which is at distance 1 from each child. Since the diameter of the space is $\log_2 n$, this is essentially the worst possible distortion. This is in stark contrast to the undirected case, where choosing one of the three permutations as the median is a $\frac{4}{3}$-approximation for any metric.

In this section, we give an $O(n^3 \log n)$ time algorithm which approximates the optimum up to an additive error of $O(\log \log n)$. The algorithm returns a simple graph data structure which represents not just a single optimal solution, but *all* optimal solutions.

### 4.1 A Graph-Theoretic Formulation

The analysis of the previous sections shows that for two permutations $\hat{\pi}$ and $\pi_i$, $d_1(\hat{\pi}, \pi_i) = \lceil \log \rho_{\hat{\pi}}(\pi_i) \rceil$, where $\rho_{\hat{\pi}}(\pi_i)$ is the number of maximal increasing subsequences of $\pi_i$ with respect to $\hat{\pi}$. Alternatively, we can rephrase this in terms of the number of *descents* of $\pi_i$ with respect to $\hat{\pi}$ - this number is just $\rho_{\hat{\pi}}(\pi_i) - 1$. Now consider the directed graph $G_{\pi_i}$ on the vertex set $[n]$ and with edges $e_i = (\pi_i(i), \pi_i(i+1))$ for $i \in [n-1]$. This is a chain with $n-1$ edges. A descent in $\pi_i$ with respect to $\hat{\pi}$ corresponds to an edge $e = (i, j)$ in $G_{\pi_i}$ such that $i$ occurs after $j$ in $\hat{\pi}$. We say that $e$ violates $\hat{\pi}$. We conclude that given two permutations $\pi_i$ and $\hat{\pi}$, $d_1(\hat{\pi}, \pi_i) = \lceil \log(k+1) \rceil$ where $k$ is the number of edges in $G_{\pi_i}$ which violate $\hat{\pi}$.

In the median problem, we are looking for the permutation $\hat{\pi}$ which minimizes the quantity $\lceil \log(k_1 + 1) \rceil + \lceil \log(k_2 + 1) \rceil$, where $k_i$ is the number of edges of $G_{\pi_i}$ which violate $\hat{\pi}$. Now consider the graph $G_{\pi_1, \pi_2}$ on vertex set [n] and with edge set the union of the edge sets of $G_{\pi_1}$ and $G_{\pi_2}$. Color the edges of $G_{\pi_1, \pi_2}$ that come from $G_{\pi_1}$ blue and the edges that come from $G_{\pi_2}$ red. To solve the median problem, we need to find the

[1] Note that we assume that gene loss events occur instantaneously with respect to speciation events, so duplicated genes cannot be carried across speciation boundaries, and if both genomes are duplicated, we charge 2 for this and not 1.

[2] The inversion distance is defined on signed permutations (e.g. $+1 - 2 + 4 - 3$). An inversion takes a contiguous segment of the genome and reverses both the elements and their signs (e.g. an inversion of the last three elements produces $+1 - 2 + 4 - 3 \rightarrow +1 + 3 - 4 + 2$). The inversion distance between two signed permutations is the minimum number of inversions needed to convert one into the other. The breakpoint distance is defined on unsigned permutations and is defined as the number of adjacencies in one genome that are not present in the other.

pair $(b, r)$ which minimizes $\lceil \log(b+1) \rceil + \lceil \log(r+1) \rceil$, such that it is possible to remove $b$ blue edges and $r$ red edges from $G_{\pi_1,\pi_2}$ such that the remaining graph is a directed acyclic graph (DAG). We can then choose $\hat{\pi}$ to be any permutation that is consistent with the remaining graph, in other words any linear ordering of this remaining graph. This problem is a variation of the Feedback Arc Set problem.

**4.2 A Polynomial Time Approximation Algorithm** In this section, we give a reduction of the median problem to the Feedback Arc Set problem which gives a $O(n^3 \log n)$ time algorithm with a guarenteed $O(\log \log n)$ additive error.

We are trying to optimize the quantity $\lceil \log(b+1) \rceil + \lceil \log(r+1) \rceil$, which is within less than two units of $\log(b+1) + \log(r+1)$. As previously stated, it is trivial to decide what is the minimum number $b$ of blue edges which need to be removed if $r = 0$. So we may assume that $b \geq 1$ and similarly $r \geq 1$. Therefore the quantity $(r+1)(b+1) < 4rb$, and $\log r + \log b + 4 > \lceil \log(b+1) \rceil + \lceil \log(r+1) \rceil$. Thus, it suffices to approximate $\log r + \log b$ to within $\log \log n + \log \log \log n + O(1)$ and we can replace our objective function by $\log r + \log b$, or equivalently $rb$.

Now let $r_{opt}$ and $b_{opt}$ give the minimum for this new objective function. Suppose for now that we know these quantities in advance. Let $\lambda = \lfloor \log(r_{opt}/b_{opt}) \rfloor$ and $\Lambda = 2^\lambda$. Then $b_{opt}\Lambda < r_{opt} \leq 2b_{opt}\Lambda$.

Now in the graph $G_{\pi_1,\pi_2}$ assign weight $\Lambda$ to the blue edges and weight 1 to the red edges. By the algorithm of Even et al. [24], we can compute an $O(\log n \log \log n)$ approximation to a minimum feedback arc set in this new graph in time $O(n^3)$. Suppose that the result of our computation has $r$ red edges and $b$ blue edges. Then $r + b\Lambda < A(r_{opt} + b_{opt}\Lambda)$, where $A = O(\log n \log \log n)$. But $b_{opt}\Lambda < r_{opt} \leq 2b_{opt}\Lambda$, therefore

$$\begin{aligned} rb\Lambda &< \frac{(r+b\Lambda)^2}{4} \\ &< \frac{A^2(r_{opt}+b_{opt}\Lambda)^2}{4} \\ &< \frac{9}{8}A^2(r_{opt}b_{opt}\Lambda) \end{aligned}$$

This implies

$$\log(rb) < log(r_{opt}b_{opt}) + \log \log n + \log \log \log n + O(1).$$

However, we do not know the value of $\Lambda$. Notice that $-L \leq \Lambda \leq L$, with $L = \lceil \log n \rceil$. We can therefore iterate over all possible $2L$ values of $\Lambda$ and simply pick the best overall result. Since each iteration takes $O(n^3)$ time, the total running time will be $O(n^3 \log n)$.

## 5 Experiments

For our experiments, instead of our approximation algorithm, we implemented a brute-force exact algorithm that simply searches exhaustively over all solutions $(b, r)$. Although it requires time exponential in the total distance, we have been able to run it at modest running times since the distances are usually quite small in practical data sets.

For our experiments, we chose one small example which illustrates that although our primary focus is on the mathematical properties of our model, it is not so abstracted so as to be devoid of biological insight. The evolutionary history of the three sequenced Brachiopod mitochondrial genomes is an extremely vexing problem in which the genomes are so scrambled that no reasonable evolutionary scenario is known [25]. We have carefully selected this particular example because it does not contain any gene inversions, which are not currently accomodated by our model, because the topology of the tree on the three species is known independently (though not, of course, the ancestral genomes or branch lengths), because the homology relationships between the genes are unambiguous and because the gene contents are equal. Some authors have speculated that multiple rounds of tandem duplication and loss might be able to explain such scrambled gene orders. However, our experiments show that this is unlikely to be true for the Brachiopods. This is possible because Theorem 1.1 gives an exact condition for the number of whole genome duplications needed and so in particular, it provides a *lower bound* on the number of duplications of *any* length required.

Using our implementation, we establish that the most parsimonious evolutionary scenario for the Brachiopod clade requires at least 8 duplications, which is close to the diameter of the space since mitochondrial genomes have only 37 genes. The implication is that other rearrangement processes, most likely transpositions, are required to explain the gene order of the Brachiopod clade.

## 6 Acknowledgements

We thank Jeff Boore for providing us with the data used in Section 5 and Andrej Bogdanov for helpful discussions. This work was supported by NSF grant EF 03-31494. Radu Mihaescu was supported by the Fannie and John Hertz Foundation graduate fellowship.

## References

[1] B. Moret, J. Tang, and T. Warnow. Reconstructing phylogenies from gene-content and gene-order data.

In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*. Oxford Univ. Press, 2004.

[2] D. Sankoff and J. Nadeau, editors. *Comparative Genomics*. Kluwer Academic Publishers, 2000.

[3] N. El-Mabrouk. Genome rearrangements by reverals and insertions/deletions of contiguous segments. In *CPM*, 2000.

[4] M. Marron, K.M. Swenson, and B.M.E. Moret. Genomic distances under deletions and insertions. In *COCOON*, 2003.

[5] J. Tang, B. Moret, L. Cui, and C. dePamphilis. Phylogenetic reconstruction from arbitrary gene-order data. In *BIBE*, pages 592–599, 2004.

[6] K.M. Swenson, M. Marron, J.V. Earnest-DeYoung, and B.M.E. Moret. Approximating the true evolutionary distance between two genomes. Tech. Report TR-CS-2004-15, Univ. of New Mexico, 2004.

[7] J. L. Boore. The duplication/random loss model for gene rearrangement exemplified by mitochondrial genomes of deuterostome animals. In D. Sankoff and J. Nadeau, editors, *Comparative Genomics*. Kluwer, 2000.

[8] R. L. Mueller and J. L. Boore. Molecular mechanisms of extensive mitochondrial gene rearrangement in plethodontid salamanders. *Mol. Bio. Evol.*, 22(10):2104–2112, 2005.

[9] C. Moritz, T. E. Dowling, and W. M. Brown. Evolution of animal mitochondrial DNA: relvance for population biology and systematics. *Annu. Rev. Ecol. Syst.*, 18:269–292, 1987.

[10] S. Bensch and A. Härlid. Mitochondrial genomic rearrangements in songbirds. *Mol. Biol. Evol.*, 17:107–113, 2000.

[11] D. V. Lavrov, J. L. Boore, and W. M. Brown. Complete mtDNA sequences of two millipedes suggest a new model for mitochondrial gene rearrangements: duplication and nonrandom loss. *Mol. Biol. Evol.*, 19(2):163–169, 2002.

[12] J. G. Inoue, M. Miya, K. Tsukamoto, and M. Nishida. Evolution of the deep-sea gulper eel mitochondrial genomes: large-scale gene rearrangements originated within the eels. *Mol. Biol. Evol.*, 20:1917–1924, November 2003.

[13] I. Miklos and J. Hein. Genome rearrangement in mitochondria and its computational biology. In J. Lagergren, editor, *Comparative Genomics: RECOMB 2004 Int. Workshop*, volume 3388. Springer-Verlag, 2005.

[14] S. Ohno. *Evolution by Gene Duplication*. Springer, New York, 1970.

[15] M. Kellis, B. Birren, and E. Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428:617–624, 2004.

[16] O. Jaillon, J. Aury, F. Brunet, J. Petit, N. Stange-Thomann, et al. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, 431:946–957, October 2004.

[17] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, 2004.

[18] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW*, 2001.

[19] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: Ranking and clustering. In *STOC*, 2005.

[20] S. Arora, A. Frieze, and H. Kaplan. A new rounding procedure for the assignment problem with applications to dense graph arrangement problems. In *FOCS*, 1996.

[21] I. Pe'er and R. Shamir. The median problem for breakpoints are NP-complete. Technical Report TR98-071, ECCC, 1998.

[22] A. Caprara. Formulations and hardness of multiple sorting by reversals. In *RECOMB*, 1999.

[23] I. Pe'er and R. Shamir. Approximation algorithms for the permutations median problem in the breakpoint model. In D. Sankoff and J. Nadeau, editors, *Comparative Genomics*, pages 225–241. Kluwer Academic Press, 2000.

[24] G. Even, J. Naor, B. Schieber, and M. Sudan. Approximating minimum feedback sets and multi-cuts in directed graphs. In *Proc. 4th Int. Conf. on Integer Prog. and Combinatiorial Optimization*, 1995.

[25] K. G. Helfenbein, W. M. Brown, and J. L. Boore. The complete mitochondrial genome of the articulate brachiopod *Terebratalia transversa*. *Mol. Biol. Evol.*, 18(9):1734–1744, September 2001.