Preprint. Genome Res. **12(1)**, January 2002, 26–36.

Genome-Scale Evolution: Reconstructing Gene Orders in the Ancestral Species

Guillaume Bourque* Pavel A. Pevzner †

Abstract

Recent progress in genome-scale sequencing and comparative mapping raises new challenges in studies of genome rearrangements. Although the pairwise genome rearrangement problem is well-studied, algorithms for reconstructing rearrangement scenarios for multiple species are in great need. The previous approaches to multiple genome rearrangement problem were largely based on the breakpoint distance rather than on a more biologically accurate rearrangement (reversal) distance. Another shortcoming of the existing software tools is their inability to analyze rearrangements (inversions, translocations, fusions, and fissions) of multichromosomal genomes. This paper proposes a new multiple genome rearrangement algorithm that is based on the rearrangement (rather than breakpoint) distance and that is applicable to both unichromosomal and multichromosomal genomes. We further apply this algorithm for genome-scale phylogenetic tree reconstruction and deriving ancestral gene orders. In particular, our analysis suggests a new improved rearrangement scenario for a very difficult *Campanulaceae* cpDNA dataset and a putative rearrangement scenario for human, mouse and cat genomes.

1 Introduction

The traditional phylogenetic tree reconstruction is based on the analysis of individual genes (Graur and Li, 2000 [12]). In contrast, genome rearrangement studies are based on genome-wide analysis of gene orders rather than individual genes (Palmer and Herbon, 1988 [29], Palmer, 1992 [28], Sankoff et al., 1992 [33], Olmstead et al., 1994 [27], Bafna and Pevzner, 1995 [1], Hannenhalli et al., 1995 [13], Blanchette et al., 1999 [5], Cosner et al., 2000 [10]). The study of genome rearrangements started more than 60 years ago (Dobzhansky and Sturtevant, 1938 [11]), but interest on the subject has flourished in recent years due to progress in large-scale sequencing and comparative mapping (O'Brien et al., 1999 [26], Murphy et al., 2000 [23], Lander et al., 2001 [20], Venter et al., 2001 [35]).

In the context of genome rearrangements, genomes are typically viewed as *signed permutations* where each integer corresponds to a unique gene/marker and the sign corresponds to its orientation (strand). For unichromosomal genomes, the most common rearrangements are *inversions* that are often referred to as *reversals* in bioinformatics. A *reversal* $\rho(i,j)$, applied to a permutation $\pi = \pi_1 \dots \pi_{i-1}\ \pi_i \dots \pi_j\ \pi_{j+1} \dots \pi_n$, reverses the segment $\pi_i \dots \pi_j$ and produces the permutation $\pi \cdot \rho(i,j) = \pi_1 \dots \pi_{i-1}\ -\pi_j\ -\pi_{j-1} \dots -\pi_i\ \pi_{j+1} \dots \pi_n$. For example, the effect of the reversal $\rho(4,8)$ on the identity permutation is the following:

*Department of Mathematics, University of Southern California. E-mail: gbourque@usc.edu

†Department of Computer Science and Engineering, University of California, San Diego. E-mail: ppevzner@cs.ucsd.edu

$$\begin{array}{c} 1\ 2\ 3\ \ 4\ \ 5\ \ 6\ \ 7\ \ 8\ 9\ 10 \\ \downarrow\ \ \rho(4,8)\ \ \downarrow \\ 1\ 2\ 3\ \text{-}8\ \text{-}7\ \text{-}6\ \text{-}5\ \text{-}4\ 9\ 10 \end{array}$$

Given two permutations π and γ, the *reversal distance*, $d(\pi, \gamma)$, is defined as the minimum number of reversals required to convert one permutation into the other. The study of reversal distance was pioneered David Sankoff (Sankoff, 1992 [31], Kececioglu and Sankoff, 1994 [19]) and increasingly efficient polynomial-time algorithms have been developed to compute the reversal distance (Hannenhalli and Pevzner, 1995 [14], Berman and Hannenhalli, 1996 [3], Kaplan et al., 1997 [18], Moret et al., 2000 [22], Bergeron, 2001 [2]).

The *Multiple Genome Rearrangement Problem* is to find a phylogenetic tree describing the most "plausible" rearrangement scenario for multiple species (Hannenhalli et al., 1995 [13], Sankoff et al., 1996 [34]). Formally, given a set of m signed permutations (existing genomes) of order n, find a tree T with the m permutations as leaf nodes and assign permutations (ancestral genomes) to internal nodes such that $D(T)$ is minimized, where

$$D(T) = \sum_{(\pi,\gamma)\in T} d(\pi, \gamma)$$

is the sum of the reversal distances over all edges of the tree. The special case of three genomes ($m = 3$) is called the *Median Problem* (Figure 6) .

Although the reversal distance for a pair of genomes can be computed in polynomial time (Hannenhalli and Pevzner, 1999 [17]), its use in studies of multiple genome rearrangements was somewhat limited since it was not clear how to combine pairwise rearrangement scenarios into a multiple rearrangement scenario. In particular, Caprara, 1999 [8] demonstrated that even the simplest version of the Multiple Genome Rearrangement Problem, the Median Problem, is NP-hard. As a result this line of research was later abandoned in favor of the breakpoint analysis approach and the existing tools use the so-called *breakpoint distance* (Watterson et al., 1982 [38], Nadeau and Taylor, 1984 [25]) to derive the rearrangement scenarios. However, the breakpoint analysis has some limitations in the analysis of *pairwise* genome rearrangements (Pevzner, 2000 [30]). One of the reasons why the breakpoint distance dominated the analysis of *multiple* rearrangements in the last few years is that it was not clear how to compute the plausible reversal-based evolutionary scenarios. This paper uses the reversal distance for computing multiple rearrangement scenarios and discusses some advantages of this approach over the breakpoint distance approach.

Our algorithm explores the specifics of the reversal distance and is based on the observation that the reversal distance is a good approximation of the true distance for many biologically relevant cases. Let γ be a genome that evolved from a genome π by k reversals (i.e., the true distance between π and γ is k). We say that π and γ form a *valid pair* if $d(\pi, \gamma) = k$; otherwise we say that $d(\pi, \gamma)$ *underestimates* the true distance (Wang and Warnow, 2001 [36]). Typically two genomes form a valid pair if the number of rearrangements between them is relatively small, exactly the case in a number of genome rearrangement studies. Figure 1 illustrates that for a genome with $n = 100$ markers the reversal distance approximates the true distance very well as long the number of reversals remains below $0.4n$. In many biologically relevant cases (e.g., rearrangements of the X chromosome in mammalian species) the number of rearrangement events is well below $0.4n$. As a result, the reversal distance often corresponds to valid (or "almost valid") pairs of genomes. Therefore the genome-based evolutionary trees are often *additive* or "almost additive" (Buneman, 1971 [7]). This property allows one to design new genome rearrangement algorithms that explore the specifics of additive trees.

Let π and γ be the leaves (existing genomes) in the evolutionary tree T and let $\pi = \sigma_1, \sigma_2, \ldots, \sigma_{k-1}, \sigma_k = \gamma$ be a path between π and γ in T passing through the ancestral genomes $\sigma_2, \ldots, \sigma_{k-1}$. Define

$$d_T(\pi, \gamma) = \sum_{1 \le i \le k-1} d(\sigma_i, \sigma_{i+1})$$

For a valid pair, $d_T(\pi, \gamma) = d(\pi, \gamma)$. We define the *deficit* of π and γ as $\mathrm{def}(\pi, \gamma) = d_T(\pi, \gamma) - d(\pi, \gamma)$. The deficit of the tree T is defined as $\mathrm{def}(T) = \sum \mathrm{def}(\pi, \gamma)$ where the sum is taken over all pairs of leaves. The closer the tree is to being additive, the smaller the deficit of the tree will be. Many genome-based trees are "almost additive", for example the herpes virus tree from Hannenhalli et al., 1995 [13] has deficit 1, while the mtDNA tree from Sankoff et al., 1996 [34] has deficit 3. Our algorithms are implicitly based on this observation and we demonstrate below that they provide an accurate reconstruction of the ancestral genomes for trees with small deficit. They use pairwise genomic distance software as a subroutine (implemented by Glenn Tesler and available via the GRIMM web server at Pavel Pevzner's laboratory web site `http://www-cse.ucsd.edu/groups/bioinformatics/software.html`). The multiple genome rearrangement software described in this paper will be available from the same web server in the near future.

The paper is organized as follows. In Section 2, we review previous work on the Multiple Genome Rearrangement Problem. In Section 3, we describe a new algorithm to solve the Multiple Genome Rearrangement Problem for unichromosomal genomes. Finally, in Section 4, we study rearrangements of multichromosomal genomes.

2 Previous Work

Studies of the Multiple Genome Rearrangement Problem started from the special case of the Median Problem. That is, given the gene order of three unichromosomal genomes G_1, G_2 and G_3, find the ancestral genome A which minimizes the total reversal distance $d(A, G_1) + d(A, G_2) + d(A, G_3)$. The *breakpoint analysis* (Blanchette et al., 1997 [4] and Sankoff and Blanchette, 1997 [32]) attempts to solve the median problem by minimizing the *breakpoint distance* instead of the reversal distance. A pair of elements in permutations π and γ form a *breakpoint* if they are consecutive in one permutation but non-consecutive in the other. The *breakpoint distance* between two permutations is simply the number of breakpoints. Blanchette et al., 1997 [4] and Sankoff and Blanchette, 1997 [32] generalized this notion for the case of more than two genomes.

The drawback of breakpoint analysis is that the breakpoint distance, in contrast to the reversal distance, does not correspond to a minimum number of rearrangement events. As a result, the *breakpoint median*, recovered by breakpoint analysis, rarely corresponds to the *ancestral median*, the genome that minimizes the overall number of rearrangements in the evolutionary scenario. Our simulations demonstrate that in many cases the ancestral median correctly reconstructs the ancestral genome. Another problem with breakpoint analysis is that it is not clear how to adapt it to multichromosomal genomes.

To illustrate the drawbacks of the breakpoint analysis, consider the following simple example. Suppose that the genomes G_1, G_2, and G_3, evolved from the ancestral genome $A = 1\ 2\ 3\ 4\ 5\ 6$ by one reversal each such that

$$\begin{aligned} G_1 &= 1\ 2\ \text{-}4\ \text{-}3\ 5\ 6 \\ G_2 &= 1\ \text{-}4\ \text{-}3\ \text{-}2\ 5\ 6 \\ G_3 &= 1\ 2\ 3\ 4\ \text{-}5\ 6\,. \end{aligned}$$

Searching for the breakpoint median will produce 4 optimal solutions: A, but also G_1, G_2, and G_3. If the median is A, then we have 2 breakpoints on each edge of the tree for a total of 6. But if the median is any of the 3 genomes, we also get a total of $6 = 0+3+3$ breakpoints. Hence, in this simple case, the breakpoint median fails to unambiguously identify the ancestor. Conversely, the only solution for the ancestral median is A since it is the only permutation generating a tree with a total score of 3 reversals.

This paper studies the ancestral median problem since it appears to be more biologically accurate than the breakpoint median. Initial attempts to recover the ancestral median were made by Hannenhalli et al., 1995 [13] and Sankoff et al., 1996 [34] who came up with approaches that may work well for 3 close genomes. However, it is not clear how to generalize their approaches for more than three genomes. In particular, Hannenhalli et al., 1995 [13] were successful in reconstructing a genome rearrangement scenario for 3 herpesviruses but failed to resolve a very complicated dataset of 13 *Campanulaceae* cpDNAs with 105 markers (the comparative maps were constructed by Mary Beth Cosner and colleagues in early 90s). The variety of rearrangements in these flowering plants far exceeds that reported in any group of land plants thus making this dataset a challenging problem for any genome rearrangement study.

The first relatively large dataset of rearranged genomes was studied by Blanchette et al., 1999 [5] who used `BPAnalysis`, the original implementation of breakpoint analysis [4], to analyze 11 metazoan mtDNA with 35 markers. As for the *Campanulaceae* problem, it remained unsolved for almost ten years until Cosner et al., 2000 [9], [10] improved on `BPAnalysis` and constructed a rearrangement scenario with 67 reversals. Recently, Moret et al., 2001 [22] developed the software `GRAPPA` which further improved on `BPAnalysis`. Finally, in a recent breakthrough, Moret et al., 2001 [21] described a million-fold speedup over `GRAPPA` and re-evaluated the *Campanulaceae* rearrangement scenario. Their analysis returned 216 trees with reversal distance 67 as compared to only 4 such trees in the previous analysis. Our `MGR` algorithm described below improves on this recent result by generating a rearrangement scenario with only 65 reversals that was overlooked by Moret et al., 2001 [21].

3 Multiple Genome Rearrangement Problem

3.1 Algorithm

We first explain the idea of our algorithm for the case of 3 genomes. Our algorithm evaluates all possible reversals for each of the 3 genomes, identifying *good reversals*. Intuitively, a reversal is good if it brings a genome closer to the ancestral genome. Of course, the ancestral genome is unknown and therefore it is unclear how to find good reversals. However, we argue (and confirm by simulations) that the reversals which bring G_1 closer to *both* G_2 and G_3 are likely to be good reversals. If this is correct, then we don't need the ancestral genome to find good reversals. We then carry on good reversals in the genomes G_1, G_2, and G_3 in the hope that they will bring us closer to the ancestor, and iterate until the genomes G_1, G_2 and G_3 are transformed into an identical genome (converge to the ancestor A). Ideally, at this point, we have reached the most likely ancestral median. Of course, this approach works well only for "almost additive" trees with small deficit and we argue that it is the case for many biologically interesting samples.

A *good reversal* ρ in genome G_1 is a reversal that reduces the reversal distance between G_1 and G_2 *and* the reversal distance between G_1 and G_3. Define $\Delta(\rho)$ as the overall reduction in the reversal distances:

$$\Delta(\rho) = (d(G_1, G_2) + d(G_1, G_3)) - (d(G_1 \cdot \rho, G_2) + d(G_1 \cdot \rho, G_3))$$

The reversal ρ is good if $\Delta(\rho) = 2$. Good reversals in genomes G_2 and G_3 are defined similarly.

The idea of our algorithm is embarrassingly simple: look for good reversals in G_1, G_2, and G_3 and perform them (if there are any) until each of the genomes is turned into the same ancestral genome. In many cases (in particular for additive trees) good reversals are sufficient to bring all three genomes to the ancestor. We call an instance of the median problem that can be resolved using only good reversals a *perfect triple*. If we run out of good reversals before the three genomes converge to the ancestor, we relax our definition of good reversals.

This approach leaves room for some flexibility. Often there is a variety of good reversals and it is not clear which one to choose. For example, the list of good reversals often contains non-overlapping reversals $\rho(i_1, j_1)$ and $\rho(i_2, j_2)$ with $i_1 \leq j_1 < i_2 \leq j_2$ and the order in which these reversals are performed is often irrelevant. Our objective is to choose good reversals in such a way that we don't run out of good reversals until all three genomes converge to the ancestor. One way to address this problem is to test all pairs/triples/... of reversals in order to avoid reversals that would cause us to run out of good reversals in a few steps. We also use a heuristic to choose the *best reversal* from the list of good reversals. The heuristic is based on an observation that good reversals, if carried out in the correct order, should not affect most of the other good reversals that are available. Hence, for each good reversal ρ, we compute n_ρ, the number of good reversals that will be available if ρ is carried out. The heuristic picks the good reversal ρ with the maximal n_ρ as the best reversal, the reversal to be carried out. We implemented these procedures in a program called `MGR-MEDIAN` and it turned out to work well in practice.

In some cases, no good reversal will be available, i.e. $\Delta(\rho) < 2$ for all reversals ρ in each of the three genomes. In those situations, the *best reversal* will be the result of a depth k search minimizing the total pairwise reversal distances. Suppose we have a sequence of k reversals $\rho_1, \rho_2 \ldots \rho_k$ to be applied to G_1. Define

$$\begin{aligned}\Delta(\rho_1, \rho_2, \ldots, \rho_k) &= d(G_1, G_2) + d(G_1, G_3) - \\ &\quad (d(G_1 \cdot \rho_1 \cdots \rho_k, G_2) + d(G_1 \cdot \rho_1 \cdots \rho_k, G_3))\,.\end{aligned}$$

Let

$$\Delta = \max_{\rho_1, \ldots, \rho_k} \Delta(\rho_1, \rho_2, \ldots, \rho_k)$$

and $\hat{\rho}_1, \ldots, \hat{\rho}_k$ be a sequence of reversals achieving this maximum. Δ then corresponds to the maximal reduction in the reversal distance after the depth k search. The *best reversal* in G_1 will be the first reversal of the sequence, i.e. $\hat{\rho}_1$ (the best reversals in G_2 and G_3 are defined similarly). When no good reversal is available, the reversal that will be carried out by `MGR-MEDIAN` will be the result of this search.

The depth k search should be taken with caution when one of the genomes is already within distance less than k from the ancestor. In this case we consider k reversals $\rho_1, \rho_2, \ldots, \rho_k$, where the first x reversals are applied to G_1, the next y reversals are applied to G_2, and the remaining $k - x - y$ reversals are applied to G_3, and maximize the function

$$\begin{aligned}\Delta(\rho_1, \rho_2, \ldots, \rho_k) &= d(G_1, G_2) + d(G_1, G_3) + d(G_2, G_3) \\ &\quad -d(G_1 \cdot \rho_1 \cdots \rho_x, G_2 \cdot \rho_{x+1} \cdots \rho_{x+y}) \\ &\quad -d(G_1 \cdot \rho_1 \cdots \rho_x, G_3 \cdot \rho_{x+y+1} \cdots \rho_k) \\ &\quad -d(G_2 \cdot \rho_{x+1} \cdots \rho_{x+y}, G_3 \cdot \rho_{x+y+1} \cdots \rho_k)\,.\end{aligned}$$

Now consider the case of $m > 3$ genomes $G_1, G_2, \ldots, G_m$. We generalize the previous definition of *good reversal* in G_i to be any reversal that reduces the reversal distance from G_i to all other

genomes. We define $\Delta(\rho)$ once again as the reduction in the reversal distances:

$$\Delta(\rho) = \sum_{j \neq i} d(G_i, G_j) - \sum_{j \neq i} d(G_i \cdot \rho, G_j)$$

A *good reversal* ρ in genome G_i is now a reversal with $\Delta(\rho) = m - 1$. We iteratively carry on good reversals until any two of m genomes become identical. When we reach that point, we remove one of the two genomes and start the procedure again with $m-1$ genomes. We keep removing genomes until we are back to solving the median problem.

In many cases, especially when m is large, we will run out of good reversals before converging to the ancestral genome. In such cases, we have developed a heuristic to complete the recovery of the phylogeny. The heuristic relies on the `MGR-MEDIAN` program for 3 genomes described in the previous section. Starting from the 3 closest genomes (in terms of the reversal distance), it iteratively adds one more genome to reconstruct the full phylogeny. Whenever possible, we choose the genome that is the "closest" to the partially reconstructed tree and such that it also forms a *perfect triple* with the two endpoints of one of the edges in the tree. We seek the closest genomes first because, as we will see in Section 3.2, the closer the genomes, the more accurate the ancestor produced by `MGR-MEDIAN`.

Assume that genomes G_1, G_2, ..., G_l are already included in the tree T that corresponds to the partially reconstructed phylogeny. The problem of adding genome G_{l+1} to T corresponds to identifying the edge of T that should be split by the edge leading to G_{l+1}. We call that edge the *split edge*. The heuristic once again uses a simple greedy approach to find the split edge. For each edge (u, v) in T, compute $M = M(u, v, G_{l+1})$ the median of u, v, and G_{l+1}. The cost of adding G_{l+1} to this edge, $C(u, v)$, is the total reversal score of the median less the reversal distance between u and v (the score of the edge being removed). Formally,

$$C(u, v) = d(u, M) + d(v, M) + d(G_{l+1}, M) - d(u, v)$$

The split edge on which to add G_{l+1} is then the one with the smallest cost. Putting all these steps together, we get the algorithm `MGR`.

The described algorithms rely on our ability to find *good reversals*. Instead of computing $\Delta(\rho)$ for all possible reversals while looking for good reversals, we have implemented a speedup making the algorithms more computationally efficient. The speedup makes use of the concept of *conserved adjacency*. We call an ordered pair of markers, (x, y), a *conserved adjacency* if (x, y) or its inverse $(-y, -x)$ is present in all genomes as consecutive elements. When looking for *good reversals*, we only consider reversals that do not break any *conserved adjacency*. The justification behind this shortcut comes from a result of Hannenhalli and Pevzner, 1996 [16] and a theorem recently proved by Glenn Tesler (personal communication).

3.2 Tests

3.2.1 Simulated data

We compared our MGR algorithm to the two implementation of breakpoint analysis: `BPAnalysis` [4] and `GRAPPA` [22]. Our initial tests showed that these two programs were producing nearly identical results, and so we decided only to include results from `GRAPPA` since it was a more efficient implementation. When testing the algorithm, we are interested not only in the phylogeny that we recover but also in the correct labeling of the internal (ancestral) nodes.

We used the following simulated data for benchmarking. Starting from the identity permutation A with n genes/markers, we performed k reversals to get genome G_1, k to get G_2 and k to get

G_3. We used the resulting 3 permutations as the input to MGR-MEDIAN and GRAPPA and checked whether they reconstructed the ancestral identity permutation.

Figures 2a,b show the difference between the total reversal distance $D(T)$ of the tree recovered by the algorithm and the actual number of reversals (equal to $3k$). Figures 2c,d show the reversal distance between the ancestral permutation recovered by the algorithm and the actual ancestor, the identity permutation. The tests are conducted for various ratios $r = \#\text{reversals}/\#\text{markers}$.

Both GRAPPA and MGR-MEDIAN produce very similar solutions for $r < 0.20$. But as the ratio r increases, GRAPPA starts making errors. In contrast, MGR-MEDIAN persists in finding correct solutions and in some cases find solutions that even have fewer reversals, than the actual ancestor. The issue here is that as the ratio r increases, the assumption that the ancestor corresponds to the most parsimonious scenario sometimes fails. In Figures 2c,d we see that as the ratio r increases, both algorithms start having difficulty recovering the actual ancestor, with the solution produced by GRAPPA further away on average than the ancestor produced by MGR-MEDIAN.

Figure 3 presents the results of similar experiments with nonequidistant genomes starting from the identity permutation A and performing k, k and $2k$ random reversals to obtain G_1, G_2 and G_3. Once again, GRAPPA starts failing to recover the optimal solution at $r > 0.20$, while MGR-MEDIAN keeps finding the true ancestor.

We tested the performance of MGR for four and more genomes using a similar setup. First, we considered a small tree with 4 genomes as leaves and 2 internal (ancestral) nodes. For simplicity, we picked one of the ancestral nodes to be the identity permutation. We then randomly simulated k reversals on each branch of the tree. We used the resulting 4 leaves of the tree as the input for MGR and GRAPPA and calculated the difference between the total reversal distance of the tree recovered with the actual number of performed reversals equal to $5k$ (Figure 4a,b). We also calculated how close the solutions recovered would get of the true ancestral permutation (the identity permutation). Since in each solution there are two internal nodes, we picked the one that is closer to the identity and recorded the reversal distance between it and the identity permutation (Figure 4c,d).

Finally, to see the effect of adding more genomes, we constructed larger complete unrooted binary trees and simulated k random reversals on each branch. To obtain a sample input with m genomes that we would feed into MGR and GRAPPA, we simulated the smallest complete binary tree such that the number of leaves was larger than m and randomly removed the extra leaves. The results in Figure 5 show the difference between the total reversal distance of the tree recovered and the total reversal distance of the simulated tree. Note that it is difficult to use the ratio $r = \#\text{reversals}/\#\text{markers}$ here as it changes depending on the size of the tree. For example, when $k = 1$, if m is 4 then $r = 5/30 \approx 0.167$, but if m is 8 then $r = 13/30 \approx 0.433$ and if m is 16 then $r = 27/30 = 0.9$. Unfortunately, running GRAPPA on more than 10 genomes genomes turned out to be impossible on our workstations, as the tree space was too large. The only way to get around this problem would have been to suggest a tree topology to GRAPPA (which is exactly what we are trying to recover in the first place). However, even if we did suggest the actual tree topology to GRAPPA, we would still get an average score difference of 7.3 for $n = 30$ and of 19.1 for $n = 100$.

3.2.2 Herpesvirus data

Hannenhalli et al., 1995 [13] used herpesvirus gene orders as a test case for one of the first studies on the Multiple Genome Rearrangement Problem. They developed a rather elaborate method to solve a relatively simple instance of the median problem for Herpes simplex virus (HSV), Epstein-Barr virus (EBV) and Cytomegalovirus (CMV) (Figure 6a). As the authors themselves pointed out, the method used would not be applicable to more complex problems and new algorithms would be required. The optimal solutions recovered involved 7 reversals. The ratio #reversals/#markers in

this example is: $r = 7/25 = 0.28$. Our simulations indicate that MGR-MEDIAN typically reconstructs the correct scenario with such ratios while GRAPPA typically fails for $r > 0.2$.

We tested both MGR-MEDIAN and GRAPPA on these 3 herpesviruses to see whether they would recover the ancestral genome suggested by [13]. MGR-MEDIAN found this genome and reconstructed the rearrangement scenario with 7 reversals (Figure 6b) even though it did not correspond to a perfect triple. In contrast, GRAPPA returned a suboptimal solution with 8 reversals. Actually, the ancestor suggested by GRAPPA was the genome HSV itself, indicating the problem with the breakpoint distance described in Section 2.

3.2.3 Human, fruit fly, and sea urchin mtDNA data

Sankoff et al., 1996 [34] analyzed human, sea urchin, and fruit fly mtDNA to derive the ancestral gene order. Using MGR-MEDIAN, we found the ancestral gene order A with a total reversal distance of 39 (Figure 7). The solution is different from the ones found in [34] but the total reversal distance is the same. The ratio #reversals/#markers for this data set is $r = 39/33 \approx 1.18$, an indication of a difficult problem. Running GRAPPA on these genomes, we obtained a solution that has a total reversal distance of 43.

3.2.4 Metazoan mtDNA data

Blanchette et al., 1999 [5] used BPAnalysis in the rearrangement study of 11 metazoan mtDNAs. The genomes come from 6 major metazoan groupings: nematodes (NEM), annelids (ANN), mollusks (MOL), arthropods (ART), echinoderms (ECH), and chordates (CHO). They were originally selected in [5] to provide the analysis with exemplar of the most diverse members of each group. The two "optimal" phylogenies recovered in [5] had 199 breakpoints.

We studied the same dataset with MGR and GRAPPA and used the curated gene order data of the 11 genomes from the MGA Source Guide compiled by Jeffrey L. Boore http://www.jgi.doe.gov/programs/comparative/MGA_Source_Guide.html. After removing two genes that were not shared by all mtDNAs we were left with a common set of 36 genes. MGR recovered a phylogeny with 150 reversals (Figure 8). The tree space for 11 genomes is very large and searching it exhaustively with GRAPPA is very time consuming. After 48 hours on a workstation, GRAPPA had recovered 3 "optimal" trees with 175 reversals and 200 breakpoints. Even suggesting the topology found by MGR to GRAPPA would only produce a fourth tree with 175 reversals.

The tree recovered by MGR is closely related to one of the optimal trees in Blanchette et al., 1999 [5]. The weak association of *Katharina tunicata* with the mollusks was already discussed in [5]. Apart from this and from the weak grouping of the two arthropods, the induced phylogeny also agrees with the metazoan phylogeny proposed by Boore and Brown, 2000 [6] (the nemathodes and the echinoderms were not discussed in this paper). We remark that Blanchette et al., 1999 [5] obtained their tree in a semi-automated regime by making a selection between the potential phylogenies and disregarding the ones breaking any one of the 6 metazoan groupings. Although such assumptions about the data were not included in MGR, it did not prevent it from the discovery of a very similar tree in a fully automated fashion (Figure 8). Rooted differently, we see that the nemathodes are a late-branching sister taxon of the annelids which is the same as in [5]. The deuterostomes (chordate and echinoderm) association was successfully identified both in [5] and in the tree from MGR but not in any of the first 3 trees produced by GRAPPA (excluding the one we suggested as a constraint).

3.2.5 *Campanulaceae* cpDNA data

We analyzed the *Campanulaceae* chloroplast dataset with 13 cpDNAs and 105 markers. It is one of the most challenging genome rearrangement datasets studied by Cosner et al, 2000 [9], [10] and Moret et al., 2001 [21]. The tree space for 13 genomes is too large and cannot be searched exhaustively by `GRAPPA`. To analyze the tree space in this case [9], [10], [21] described various techniques to obtain constraint trees to suggest to the program. `GRAPPA` then searched the space of refinements of these constraint trees trying to minimize the total number of reversals. Moret et al., 2001 [21] recovered 216 trees with a total of 67 reversals. `GRAPPA` was not able to decide which of those trees corresponds to the most likely reconstruction of the rearrangement scenario.

Running `MGR` on the same data set did not require the preprocessing of a constraint tree and recovered a tree with only 65 reversals, shown in Figure 9. The topology of the tree recovered actually corresponds to the topology of one of the trees recovered by `GRAPPA`, but the labeling of the internal nodes differs. Since our tree minimizes the number of reversals we argue that `MGR` provided a better reconstruction of the rearrangement scenario than `GRAPPA`.

4 Reconstructing Ancestral Gene Orders for Multichromosomal Genomes

4.1 Algorithm

Consider three multichromosomal genomes G_1, G_2 and G_3. The median problem is to find the ancestral genome A which minimizes the total *genomic distance* $d(A, G_1) + d(A, G_2) + d(A, G_3)$. The genomic distance in this case is defined in terms of *reversals, translocations, fusions, and fissions*, the most most common rearrangement events in multichromosomal genomes (Pevzner, 2000 [30]).

Let $\pi = \pi_1 \dots \pi_n$ be a chromosome in a multichromosomal genome and $1 \leq i \leq j \leq n$. A reversal $\rho(\pi, i, j)$ rearranges the genes *inside* π and transforms it into $\pi_1 \dots \pi_{i-1} -\pi_j -\pi_{j-1} \dots -\pi_i \pi_{j+1} \dots \pi_n$. Let $\pi = \pi_1 \dots \pi_n$ and $\sigma = \sigma_1 \dots \sigma_m$ be two different chromosomes and $1 \leq i \leq n+1$ and $1 \leq j \leq m+1$. A *translocation* $\rho(\pi, \sigma, i, j)$ exchanges genes *between* chromosomes π and σ and transforms them into chromosomes $\pi_1 \dots \pi_{i-1} \sigma_j \dots \sigma_m$ and $\sigma_1 \dots \sigma_{j-1} \pi_i \dots \pi_n$. A fusion concatenates the chromosomes π and σ, resulting in a chromosome $\pi_1 \dots \pi_n \sigma_1 \dots \sigma_m$. A fission "breaks" a chromosome π into two chromosomes $\pi_1 \dots \pi_{i-1}$ and $\pi_i \dots \pi_n$.

Given two genomes Π and Γ, the *genomic distance*, $d(\Pi, \Gamma)$, is defined as the minimum number of reversals, translocations, fusions and fissions required to convert one genome into the other. The genomic distance was first studied by Hannenhalli and Pevzner, 1995 [15] who developed a polynomial-time algorithm to compute a rearrangement scenario between man and mouse involving 131 rearrangements.

`MGR-MC` algorithm is a generalization of the `MGR-MEDIAN` algorithm for unichromosomal genomes. First, we evaluate all possible rearrangements (reversals, translocations, fusions, and fissions) for each of the 3 genomes, identifying *good rearrangements*. As in Section 3.1, a rearrangement is good if it brings a genome closer to the ancestral genome. We will argue once again that the rearrangements which bring G_1 closer to *both* G_2 and G_3 are likely to be good. We iteratively carry on these good rearrangements until the genomes G_1, G_2 and G_3 are transformed into an identical genome hoping that we have reached the most likely *ancestral median*.

Since we are dealing with multichromosomal genomes and with four different types of rearrangements, we need to be aware of an ambiguous situation that can occur when solving the median problem. Consider the following simple example:

$$G_1 = \ 1\ 2\ 3\ 4\ 5 \qquad G_2 = \ 1\ 2\ \text{-}5\ \text{-}4\ \text{-}3 \qquad G_3 = \ \begin{matrix} 1\ 2 \\ 3\ 4\ 5 \end{matrix}$$

In this example, the parsimony principle does not allow one to unambiguously reconstruct the evolutionary scenario. If the ancestor coincides with G_1 then a reversal occurred on the way to G_2 and a fission occurred on the way to G_3. But we could also have a similar scenario starting from G_2 as the ancestor or even starting from G_3 if we assume that 2 fusions occurred. In this example, $d(G_1, G_2) = d(G_1, G_3) = d(G_2, G_3) = 1$. We did not have this kind of ambiguity for unichromosomal genomes because it was impossible to find 3 genomes that would all be within 1 reversal of the each other. These ambiguities motivate a more careful selection within the good rearrangements. We use the observation that in most genomes of interest (e.g. mammalian genomes) reversals and translocations are more common than fusions and fissions. When looking for the *best rearrangement* to be carried out within the good rearrangements, we always select reversals/translocations before fusions/fissions. If we run out of good reversals before reaching a solution, the best rearrangement will be the result of a depth k search minimizing the total pairwise rearrangement distances. Putting all these steps together we get the `MGR-MC` algorithm for multichromosomal genomes that is easy to generalize for more than 3 multichromosomal genomes.

4.2 Tests

4.2.1 Simulated data

We used the following simulations to test the performance of `MGR-MC`. Starting from the identity permutation A of size n, we first randomly selected b chromosome breaks to simulate a multichromosomal ancestor. Next, to transform A into G_i ($1 \leq i \leq 3$), we performed k rearrangements where each rearrangement was randomly assigned to be a reversal/translocation with probability p and a fusion/fission with probability $1 - p$. We used the resulting 3 genomes as the input for `MGR-MC`. We are interested in the difference between the score (total number of rearrangements) of the solution recovered by the algorithm and the actual score of the simulated tree (equal to $3k$). We are also interested in the rearrangement distance between the ancestral genome recovered by the algorithm and the actual ancestor. The tests are conducted for various ratios $r = \#\text{rearrangements}/\#\text{markers}$.

Figure 10 illustrates that `MGR-MC` has no difficulty recovering ancestral genomes with a score which is at least as good as the one of the actual ancestor. Actually, in all the tests conducted, not once was the solution produced by `MGR-MC` worst than the true ancestor. The solutions produced for small ratios $r = \#\text{rearrangements}/\#\text{markers}$ tend to be very close the actual ancestor. But, as the ratio r increases, we see the same effect as for unichromosomal genomes: `MGR-MC` starts finding solutions with genomic distance which is smaller than the true number of rearrangement events. As a result, the average distance between the solution recovered and the true ancestor increases. The comparison of Figures 10a,b and 10c,d illustrates that accuracy of reconstructions deteriorates with the increase in the rate of fusions/fissions.

4.2.2 Gene order of the human-cat-mouse common ancestor

The modern comparative mapping studies generated a wealth of data on differences in genomic organization for many mammalian species. However, most existing comparative maps are *pairwise* maps representing genome organization of two species rather than *multiple* maps representing the genomic organization for more than two species. Since the number of established *universal* markers (O'Brien et al., 1999 [26]) that work across many genomes is relatively small, it is often not clear how to integrate pairwise comparative maps into multiple maps. The first sufficiently detailed

triple comparative maps appeared recently as the results of rat (Watanabe et al, 1999 [37]) and cat (Murphy et al., 2000 [24]) comparative mapping projects. We collaborated with Bill Murphy to integrate the pairwise human-mouse, human-cat, and mouse-cat comparative maps into a triple human-mouse-cat map and we use this map for deriving the ancestral genome organization.

The previous attempts to derive rearrangement history of multi-chromosomal genomes concentrated on human and mouse genomes (Nadeau and Taylor, 1984 [25], Hannenhalli and Pevzner, 1995 [15]). The cat data used in this paper comes from Murphy et al., 2000 [24] and consists of 193 markers shared by all three species. The number of markers is still too small to derive a detailed rearrangement scenario but it allows one to get some insights into a large-scale organization of the ancestor. Ultimately, this organization may be refined with `MGR-MC` as soon as more markers shared by all three species become available.

Comparative maps usually correspond to unsigned permutations, i.e., no information on the direction (signs) of the genes/markers is available. Since mammalian comparative maps contain many *singletons* (Pevzner, 2000 [30]) the existing algorithms for analyzing unsigned permutations become too time-consuming in this case. As a result we have to assign an orientation to the markers, since the current implementation of `MGR-MC` only supports signed permutations/genomes. Ultimately, this should not be a problem as more data becomes available. We used *strips* in unsigned permutations (Hannenhalli and Pevzner, 1996 [16]) to infer the signed permutations from the original unsigned permutations. Using human genome as a reference, we first identified all the strips both in cat and in mouse genomes. We then assigned an orientation to the markers based on these strips. Any marker for which we could not assign an orientation using this method either in cat or in mouse genome was removed and we were left with a common set of 114 markers. This process obviously inserts a bias towards blocks of preserved markers, while removing information about more local disruptions, e.g., single marker reversals. The resulting ancestral gene order generated by `MGR-MC` is shown in Figure 11. Although most of the elements of the ancestral organization in Figure 11 are consistent with the existing biological conjectures, the organization of ancestral chromosomes 4 and 17 is surprising and even counterintuitive. According to our scenario the chromosomes 4 and 17 in the ancestor were combined into the chromosome 5 in human and the chromosome A1 in cat. We do not argue that it is a correct reconstruction of the ancestral chromosomes 4 and 17 (more markers are needed to support this conjecture) but remark instead that `MGR-MC` provided us with solid combinatorial reasons why such scenario makes sense. Such reasons are not straightforward and hard to explain without the multichromosomal genome rearrangement software that never was available to evolutionary biologists in the past. Therefore, the non-trivial combinatorial arguments used by `MGR-MC` in the construction of Figure 11 may escape the attention of biologists that studied this problem in the past. The detailed biological analysis of our human-mouse-cat ancestral reconstruction will be discussed in another paper (joint project with Bill Murphy).

5 Acknowledgements

We are grateful to Bernard Moret and Glenn Tesler for providing their genome rearrangement programs. We are also indebted to Bill Murphy for providing us with the catdiscussions. Many thanks to Glenn Tesler for multiple discussions and comments that significantly improved the manuscript. We also thank Tandy Warnow for some useful references. This research is supported by grants from the Natural Sciences and Engineering Research Council of Canada.

References

[1] V. Bafna and P. Pevzner. Sorting by reversals: genome rearrangements in plant organelles and evolutionary history of X chromosome. *Molecular Biology and Evolution*, 12:239–246, 1995.

[2] A. Bergeron. A very elementary presentation of the Hannenhalli-Pevzner theory. In *Proceedings 12th Annual Symposium on Combinatorial Pattern Matching*, volume 2089 of *Lecture Notes in Computer Science*, pages 106–117, Jerusalem, Israel, 2001.

[3] P. Berman and S. Hannenhalli. Fast sorting by reversal. In *Combinatorial Pattern Matching. 7th Annual Symposium*, volume 1075 of *Lecture Notes in Computer Science*, pages 168–185, New York, 1996. Springer.

[4] M. Blanchette, G. Bourque, and D. Sankoff. Breakpoint phylogenies. In S. Miyano and T. Takagi, editors, *Genome Informatics Workshop (GIW 1997)*, pages 25–34, Tokyo, Japan, 1997. Univ. Academy Press.

[5] M. Blanchette, T. Kunisawa, and D. Sankoff. Gene order breakpoint evidence in animal mitochondrial phylogeny. *Journal of Molecular Evolution*, 49:193–203, 1999.

[6] J. Boore and W. Brown. Mitochondrial genomes of galathealinum, helobdella, and platynereis: Sequence and gene arrangement comparisons show that pogonophora is not a phylum and annelida and arthropoda are not sister taxa. *Molecular Biology and Evolution*, 7:87–106, 2000.

[7] P. Buneman. The recovery of trees from measures of dissimilarity. In *Mathematics in the Archaeological and Historical Sciences*, pages 387–395. Edinburgh University Press, 1971.

[8] A. Caprara. Formulations and complexity of multiple sorting by reversals. In S. Istrail, P. Pevzner, and M. Waterman, editors, *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB-99)*, pages 84–93, Lyon, France, April 1999. ACM Press.

[9] M. Cosner, R. Jansen, B. Moret, L. Raubeson, L. Wang, T. Warnow, and S. Wyman. An empirical comparison of phylogenetic methods on chloroplast gene order data in campanulaceae. In *Comparitive Genomics (DCAF-2000)*, pages 99–121, Montreal, 2000. Kluwer Acad. Pubs.

[10] M. Cosner, R. Jansen, B. Moret, L. Raubeson, L. Wang, T. Warnow, and S. Wyman. A new fast heuristic for computing the breakpoint phylogeny and experimental phylogenetic analyses of real and synthetic data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB-2000)*, pages 104–115, San Diego, 2000.

[11] T. Dobzhansky and A. Sturtevant. Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics*, 23:28–64, 1938.

[12] D. Graur and W.-H. Li. *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer Associates, Inc., 2000.

[13] S. Hannenhalli, C. Chappey, E. Koonin, and P. Pevzner. Genome sequence comparison and scenarios for gene rearrangements: A test case. *Genomics*, 30:299–311, 1995.

[14] S. Hannenhalli and P. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *Proceedings of the 27th Annual ACM-SIAM Symposium on the Theory of Computing*, pages 178–189, 1995.

[15] S. Hannenhalli and P. Pevzner. Transforming men into mice: polynomial algorithm for genomic distance problem. In *Proceedings of the 36th IEEE Symposium on Foundations of Computer Science*, pages 581–592, 1995.

[16] S. Hannenhalli and P. Pevzner. To cut... or not to cut (applications of comparative physical maps in molecular evolution). In *Proceedings of the 7th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 304–313, 1996.

[17] S. Hannenhalli and P. Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *Journal of ACM*, 46:1–27, 1999.

[18] H. Kaplan, R. Shamir, and R. Tarjan. Faster and simpler algorithm for sorting signed permutations by reversals. In *Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 344–351, New York, 1997. ACM.

[19] J. Kececioglu and D. Sankoff. Efficient bounds for oriented chromosome inversion distance. In M. Crochemore and D. Gusfield, editors, *Combinatorial Pattern Matching. 5th Annual Symposium*, volume 807 of *Lecture Notes in Computer Science*, pages 307–325, New York, 1994. Springer.

[20] Lander and et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[21] B. Moret, L. Wang, T. Warnow, and S. Wyman. New approaches for reconstructing phylogenies from gene order data. In *Proceedings of the 9th International Conference on Intelligent Systems for Molecular Biology (ISMB-2001)*, pages 165–173, 2001.

[22] B. Moret, S. Wyman, D. Bader, T. Warnow, and M. Yan. A new implementation and detailed study of breakpoint analysis. In *6th Pacific Symposium on Biocomputing (PSB 2001)*, pages 583–594, 2001.

[23] W. Murphy, R. Stanyon, and S. O'Brien. Evolution of mammalian genome organization inferred from comparative gene mapping. *Genome Biol.*, 2:1–8, 2001.

[24] W. Murphy, S. Sun, Z.-Q. Chen, N. Yuhki, D. Hirschmann, M. Menotti-Raymond, and S. O'Brien. A radiation hybrid map of the cat genome: Implications for comparative mapping. *Genome Research*, 10:691–702, 2000.

[25] J. Nadeau and B. Taylor. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences USA*, 81:814–818, 1984.

[26] S. O'Brien and et al. The promise of comparative genomics in mammals. *Science*, 286:458–481, 1999.

[27] R. Olmstead and J. Palmer. Chloroplast DNA systematics: a review of methods and data analysis. *American Journal of Botany*, 81:1205–1224, 1994.

[28] J. Palmer. Chloroplast and mitochondrial genome evolution in land plants. In R. Herrmann, editor, *Cell Organelles*, pages 99–133. Springer Verlag, 1992.

[29] J. Palmer and L. Herbon. Plant mitochondrial dna evolves rapidly in structure, but slowly in sequence. *Journal of Molecular Evolution*, 27:87–97, 1988.

[30] P. Pevzner. *Computational molecular biology: an algorithmic approach*, chapter 10. The MIT Press, 2000.

[31] D. Sankoff. Edit distance for genome comparison based on non-local operations. In A. Apostolico, M. Crochemore, Z. Galil, and U. Manber, editors, *Proceedings of the 3rd Annual Symposium on Combinatorial Pattern Matching*, pages 121–135, Tucson, AZ, 1992. Springer-Verlag, Berlin.

[32] D. Sankoff and M. Blanchette. The median problem for breakpoints in comparative genomics. In *Computing and Combinatorics, Proceeedings of COCOON '97*, Lecture Notes in Computer Science, pages 251–263, New York, 1997. Springer Verlag.

[33] D. Sankoff, G. Leduc, N. Antoine, B. Paquin, B. Lang, and R. Cedergren. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proceedings of the National Academy of Sciences USA*, 89:6575–6579, 1992.

[34] D. Sankoff, G. Sundaram, and J. Kececioglu. Steiner points in the space of genome rearrangements. *International Journal on Foundations of Computer Science*, 7:1–9, 1996.

[35] J. Venter and et al. The sequence of the human genome. *Science*, 291:1304–1352, 2001.

[36] L.-S. Wang and T. Warnow. Estimating true evolutionary distances between genomes. In *Proceedings 33rd Symposium on Theory of Computing (STOC'01)*, pages 637–646. ACM Press, 2001.

[37] T. Watanabe, M. Bihoreau, L. McCarthy, S. Kiguwa, H. Hishigaki, A. Tsuji, J. B. J, Y. Yamasaki, A. Mizoguchi-Miyakita, K. O. K, T. Ono, S. Okuno, N. Kanemoto, E. Takahashi, K. Tomita, H. Hayashi, M. Adachi, and C. W. et al. A radiation hybrid map of the rat genome containing 5,255 markers. *Nature Genetics*, 22:27–36, 1999.

[38] G. Watterson, W. Ewens, T. Hall, and A. Morgan. The chromosome inversion problem. *Journal of Theoretical Biology*, 99:1–7, 1982.

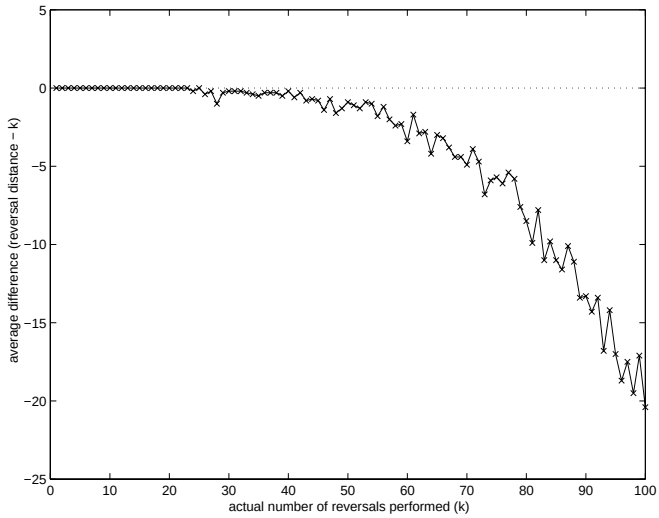

Figure 1: Reversal distance, $d(\pi, \gamma)$, versus actual number of reversals performed to transform π into γ, where γ is a genome/permutation that evolved from the identity permutation $\pi = 1, 2, \ldots, 100$ by k random reversals. The simulations were repeated 10 times for every k. We compute the average difference between the reversal distance and the actual number of reversals performed k.

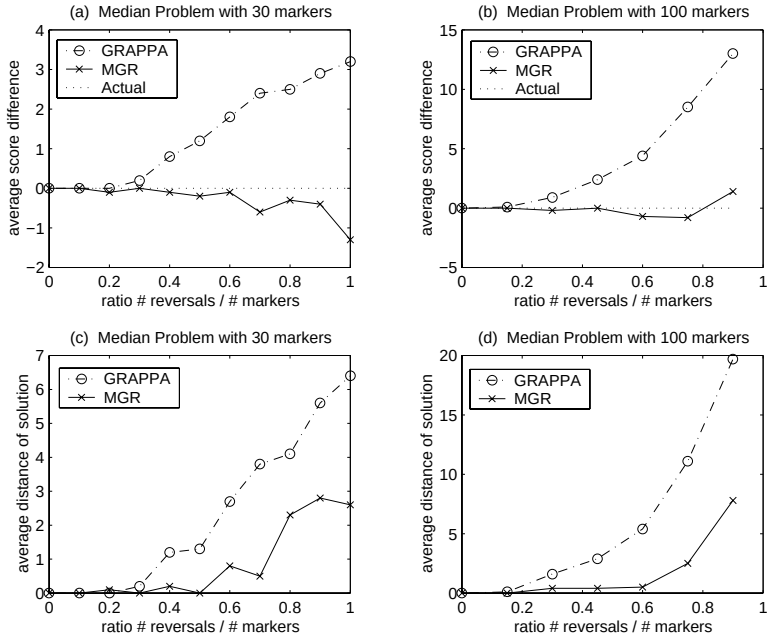

Figure 2: Comparison of `MGR-MEDIAN` and `GRAPPA` (3 genomes equidistant from the ancestor). The genomes G_1, G_2, G_3 are obtained by k reversals each from the ancestral identity permutation 1 2 ... n ($n = 30$ and $n = 100$). The simulations were repeated 10 times for every ratio #reversals/#markers $= 3k/n$. (a) and (b) The average difference between the number of reversals on the tree recovered by the algorithm and the number of reversals on the actual tree (equal to $3k$). (c) and (d) The average reversal distance between the solution recovered and the actual ancestor.

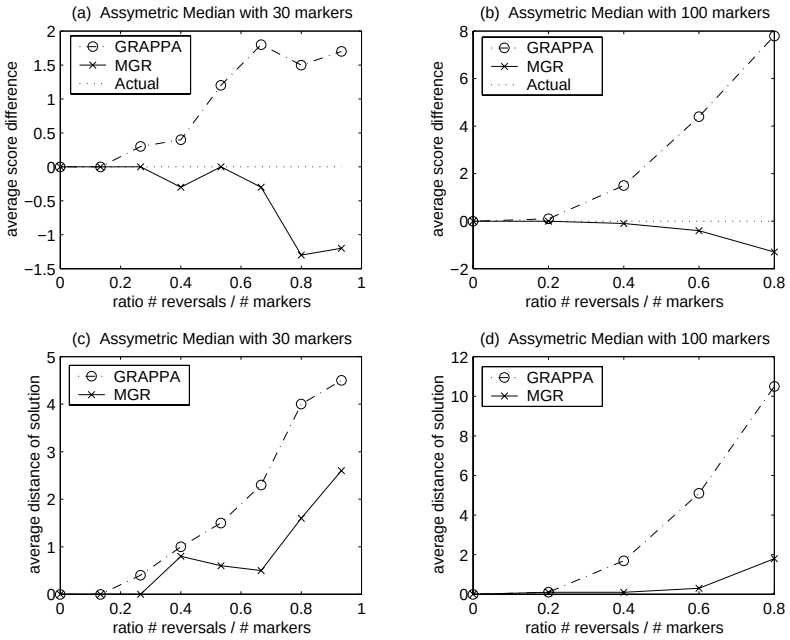

Figure 3: Comparison of `MGR-MEDIAN` and `GRAPPA` (3 genomes nonequidistant from the ancestor). The genomes G_1, G_2, G_3 are obtained by k, k and $2k$ reversals respectively each from the ancestral identity permutation 1 2 ... n ($n = 30$ and $n = 100$). The simulations were repeated 10 times for every ratio #reversals/#markers $= 4k/n$. (a) and (b) The average difference between the number of reversals on the tree recovered by the algorithm and the number of reversals on the actual tree (equal to $4k$). (c) and (d) The average reversal distance between the solution recovered and the actual ancestor.

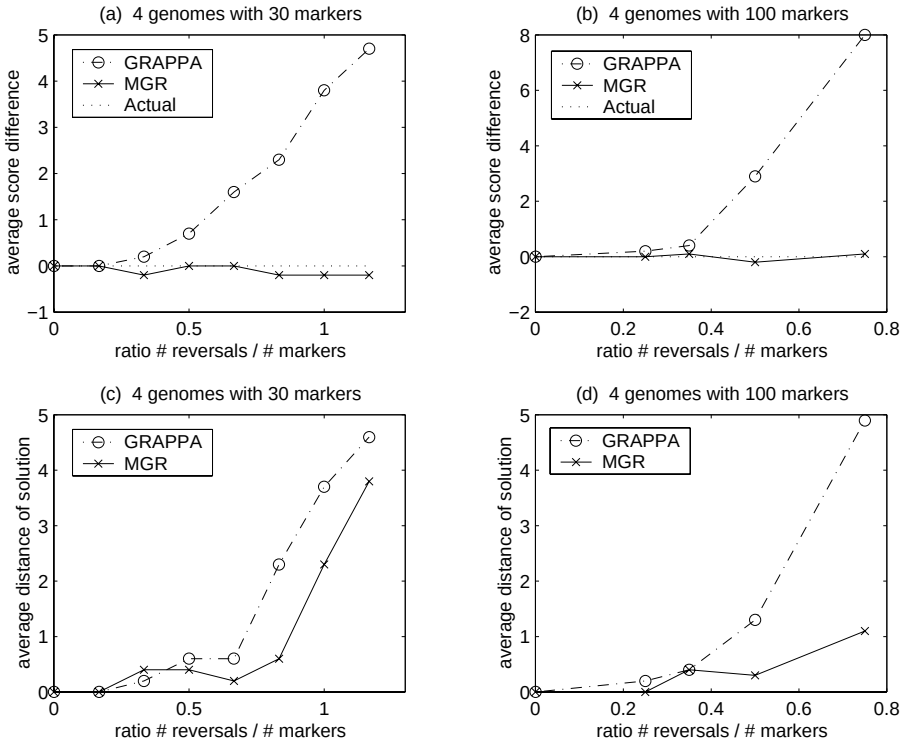

Figure 4: Comparison of `MGR` and `GRAPPA` (4 genomes). We start from an unrooted tree with 4 leaves and select one of the two internal nodes to be the identity permutation $1\ 2\ \ldots n$ ($n = 30$ and $n = 100$). We then perform k reversals on each branch of the tree to obtain the genomes G_1, G_2, G_3, G_4 as the 4 leaves of the tree. The simulations were repeated 10 times for every ratio #reversals/#markers $= 5k/n$. (a) and (b) The average difference between the number of reversals on the tree recovered by the algorithm and the number of reversals on the actual tree (equal to $5k$). (c) and (d) The average reversal distance between the *best* (i.e., closest) internal node in the solution recovered and the identity permutation.

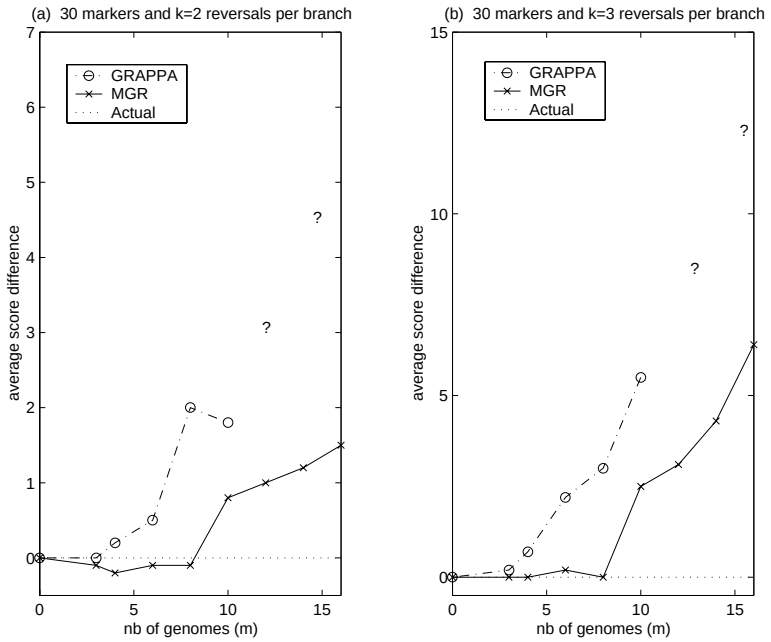

Figure 5: Comparison of `MGR` and `GRAPPA` (m genomes each with 30 markers). The genomes $G_1, G_2, \ldots, G_m$ correspond to a subset of leaves from a complete unrooted binary tree on which we have performed k reversals on each branch. The simulations were repeated 10 times for every m. (a) and (b) The average difference between the number of reversals on the tree recovered by the algorithm and the number of reversals on the actual tree when $k = 2$ and $k = 3$.

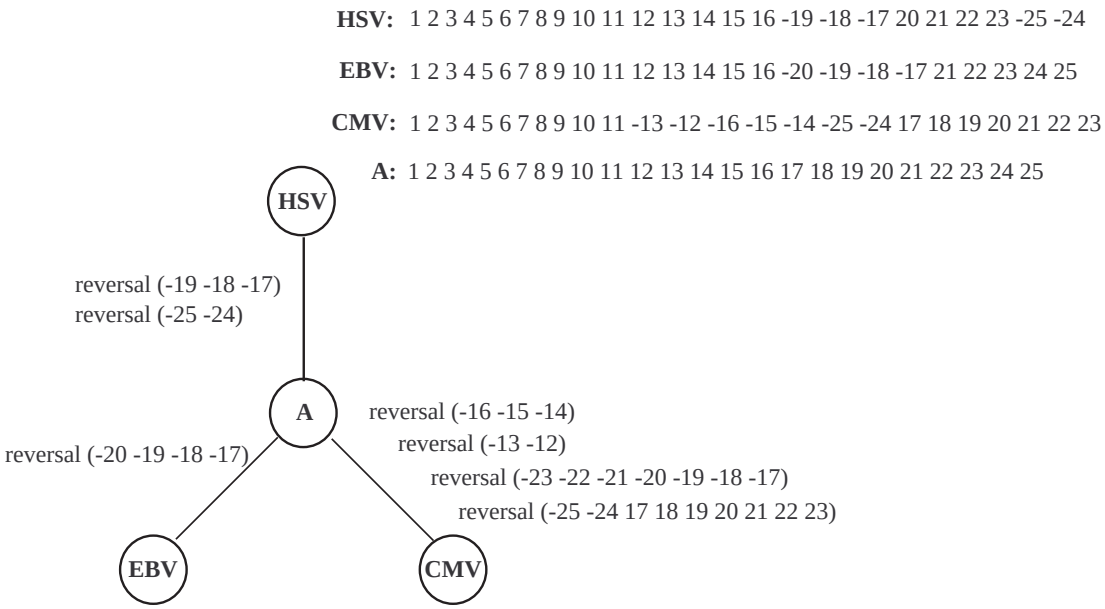

Figure 6: Herpes simplex virus (HSV), Epstein-Barr virus (EBV) and Cytomegalovirus (CMV) gene orders (Hannenhalli et al., 1995 [13]) as well as the ancestral gene order (A) and optimal evolutionary scenario recovered by `MGR-MEDIAN`.

Human:	26 13 17 12 -24 15 18 32 -2 -16 -3 -33 4 -28 7 5 1 10 19 25 22 11 29 14 20 -21 -8 6 30 -23 9 27 31
Sea urchin:	26 4 25 22 5 1 -28 19 11 29 20 -21 6 9 27 8 30 23 -24 16 14 -2 32 3 -31 15 -7 33 10 13 17 12 18
Fruit fly:	-26 -31 -27 12 -24 15 18 32 -3 -33 4 13 5 7 1 10 19 2 25 16 29 8 -9 -20 -11 -22 30 -23 21 6 28 -17 -14
A:	26 13 17 12 -24 15 18 32 -28 7 -6 21 -20 -29 -11 -22 -25 -16 8 -3 -33 4 14 -2 -19 -10 -1 -5 30 -23 9 27 31

Figure 7: Human, sea urchin and fruit fly mitochondrial gene order taken from Sankoff et al., 1996 [34]. A is the ancestral gene order suggested by `MGR-MEDIAN`.

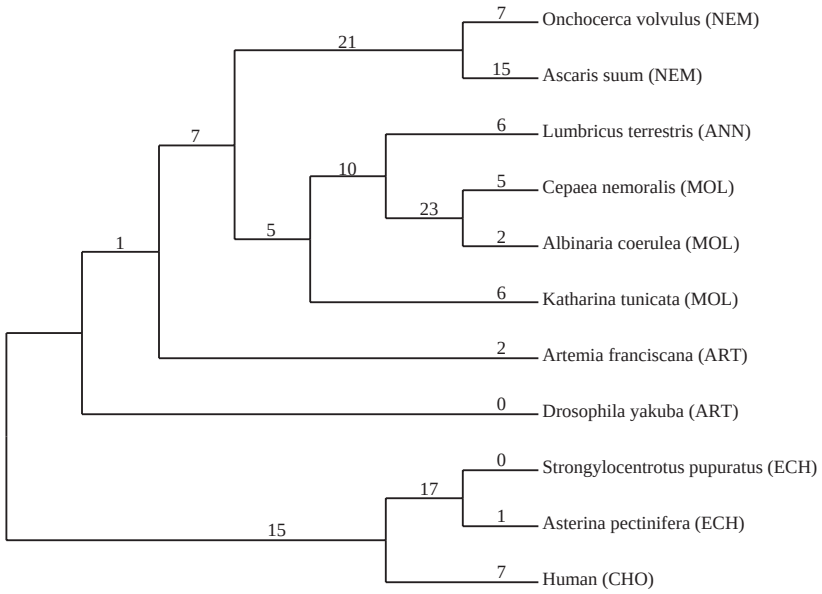

Figure 8: Phylogeny of 11 metazoan genomes reconstructed by `MGR`. The gene order data is taken from the MGA Source Guide which is compiled by Jeffrey L. Boore. The genomes come from 6 major metazoan groupings: nematodes (NEM), annelids (ANN), mollusks (MOL), arthropods (ART), echinoderms (ECH), and chordates (CHO). Numbers on the edges show the number of reversals.

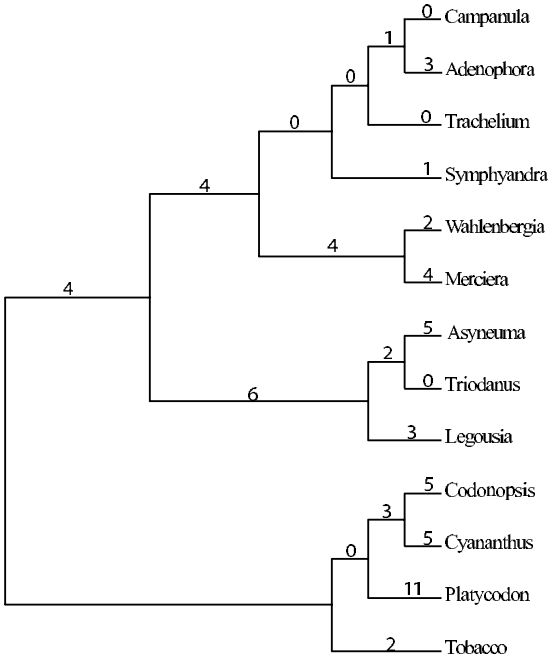

Figure 9: Phylogeny of the *Campanulaceae* cpDNA dataset as reconstructed by `MGR`. Numbers on the edges show the number of reversals.

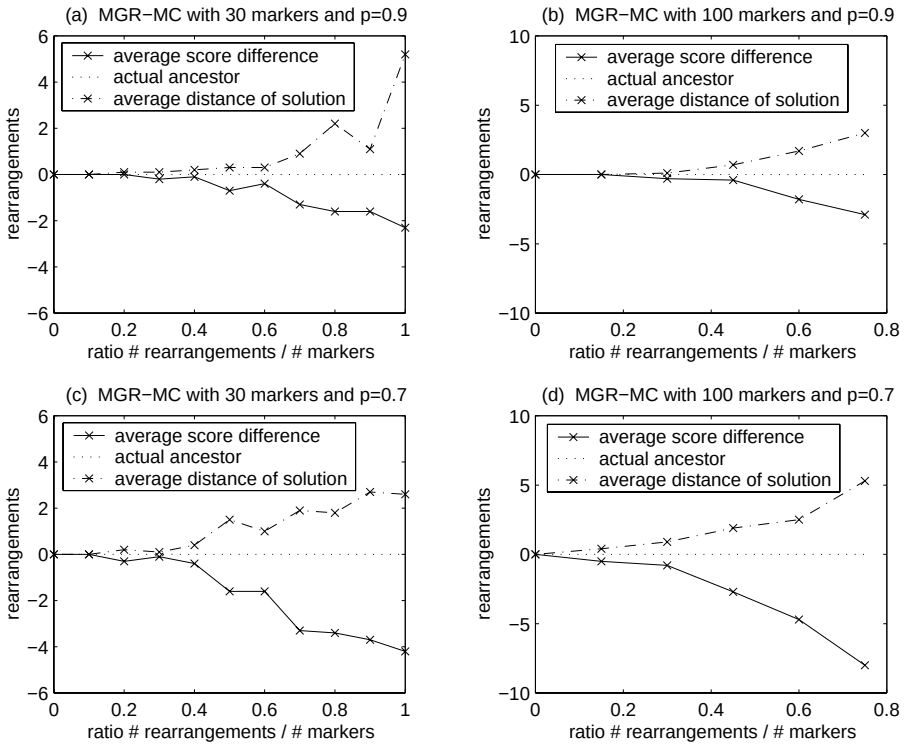

Figure 10: Performance of `MGR-MC` (3 multichromosomal genomes equidistant from the ancestor). The ancestral genomes are obtained from the identity permutation $1\ 2\ \ldots n$ ($n = 30$ and $n = 100$) by inserting b chromosomes breaks ($b = 2$ when $n = 30$ and $b = 9$ when $n = 100$). The genomes G_1, G_2, G_3 are obtained by k rearrangements each from the ancestral genomes. Each rearrangement is a reversal/translocation with probability p and a fusion/fission with probability $1 - p$. The simulations were repeated 10 times for every ratio #rearrangements/#markers $= 3k/n$. We compute the *average score difference* which is the difference between the number of rearrangements on the tree recovered by the algorithm and the actual number of rearrangements (equal to $3k$). We also compute the *average distance of solution* between the solution recovered and the actual ancestor.

Figure 11: Ancestral median for human, mouse, and cat genomes found by `MGR-MC`. We used the gene order of 114 markers spread over the chromosomes in all three species. The numbers above the chromosomes correspond to these 114 markers and the numbering is such that the human genome corresponds to the identity permutation broken in 20 pieces. The names below the chromosomes correspond to the name of the markers. We attribute a *color* to each human chromosome. The *color* of any marker (in any genome) indicates on which human chromosome is the homolog of this marker. Each marker *segment* is traversed by a diagonal line. These diagonal lines are such that the human chromosomes are traversed from top left to bottom right and are design to provide visual help to identify where rearrangements occurred. For example, for chromosome X, the gene order of the ancestor coincides with the cat gene order and only differs by one segment consisting of genes 108 and 109 (break in the diagonal line) from the human gene order. The mouse X chromosome is broken into 7 segments as compared to the ancestor (shown by 7 broken segments of the diagonal line).