# DUPLICATION, REARRANGEMENT AND RECONCILIATON

David Sankoff

Nadia El-Mabrouk

A method to account for gene order data from $N$ genomes according to a given species tree, with no restriction on the number of approximate copies of a gene (or of members of a gene family) in a genome. Gene orders, together with gene trees produced by sequence comparison, are submitted to an analysis that integrates the concepts of phylogenetic reconciliation, exemplar strings and breakpoint medians.

## 1 Introduction

This paper seeks to integrate, at the conceptual and methodological levels, three approaches to genomic evolution: genome rearrangement theory and particularly its extension to include multigene families, breakpoint-based phylogeny and genome reconstruction, and the theory of gene tree/species tree reconciliation. Starting from a given species tree $\mathcal{T}$, as well as $N$ genomes (gene orders) and $F$ gene trees summarizing the results of independent phylogenetic analyses within each of the $F$ multigene families represented in these genomes, we will show how to reconstruct gene orders at the ancestral nodes of $\mathcal{T}$. The strategy is to minimize the extent of genome rearrangement from each ancestral node of $\mathcal{T}$ to its offspring necessitated by the ancestral gene orders found.

### 1.1 Genome rearrangements

The theory of genome rearrangements, exemplified by the polynomial algorithms for signed reversal distance and for translocation distance (Hannenhalli and Pevzner, 1995a,b), and the NP-hardness result for unsigned reversal distance (Caprara, 1997), takes for input two different orders of the same set of genes. These rearrangement distances measure the number of elementary operations necessary to transform one linear order on the genes into another, where the operations model genome-level evolutionary processes such as inversion (reversal) of a chromosomal segment, transposition of a segment from one site on a chromosome to another, or translocation (exchange) of terminal segments between two chromosomes.

## 1.2 Exemplar extraction

Implicit in the rearrangements literature is that both genomes contain an identical set of genes and the one-to-one homologies (orthologies) between all pairs of corresponding genes in the two genomes have previously been established. While this hypothesis of *unique genes* may be appropriate for some small genomes, e.g. viruses and mitochondria, it is clearly unwarranted for divergent species where several copies of the same gene, or several homologous (paralogous) genes–a *multigene family*, may be scattered across a genome.

In a recent publication (Sankoff, 1999), we formulated a generalized version of the genomic rearrangement problem, where each gene may be present in a number of copies in the same genome. The central idea, based on a model of gene copy movement, is the deletion of all but one member of each gene family–its *exemplar*–in each of the two genomes being compared, so as to minimize some rearrangement distance between the two reduced genomes thus derived.

## 1.3 The median

The solution of the *median problem* is of key importance in inferring the ancestral states in a phylogenetic tree. Given a distance $d$, three genomes $A, B$ and $C$, as well as a set $\mathcal{S}$ of genomes, the median is a genome $M \in \mathcal{S}$ such that the sum $d(A,X) + d(B,X) + d(C,X)$ is minimal for $X = M$. Algorithms for finding the median can be used to reconstruct ancestors in a given phylogeny through the process of *steinerization*. Unfortunately, the median problem is NP-hard, even in the case of unique genes, for all known rearrangement distances $d$ including signed inversion distance (Caprara, 1999). Even heuristic approaches to this problem work well only for very small instances (cf Hannenhalli et al. (1995); Sankoff et al. (1996)).

On the other hand, for the breakpoint distance $d$, where $d(Y,Z)$ is the number of pairs of genes that are adjacent in genome $Y$ but not in $Z$, the median problem can be solved in a relatively simple manner for three genomes $A, B$ and $C$ having the same gene content. Indeed, in this case, the problem can be reduced to the Traveling Salesman Problem (TSP) (Sankoff and Blanchette, 1997). Although, theoretically, the latter problem is also NP-hard, there are a number of algorithms and software packages applicable in particuliar contexts (Reinelt, 1991). These allow us to find the median of three genomes of size $n = 100$ in a matter of minutes. Recently, we have developed a heuristic for this problem in the case where the genomes do not have the same set of genes (Sankoff et al., 2000). This algorithm requires calculation time of a hour or more for a similarly sized problem.

## 1.4 Reconciliation

Exemplar analysis was designed for genomic rearrangements between two genomes containing multigene families, and not for phylogenetic trees. Conversely, the median problem is an important component of phylogenetic analysis based on rearrangements, but does not apply to multigene families. A third area of research on multigene families and phylogeny is the *reconciliation* of a gene tree with a given species tree (Chen et al., 2000; Eulenstein et al., 1997; Guigó et al., 1996; Ma et al., 1998; Page, 1994; Page and Charleston, 1997). Given a gene tree constructed from an alignment of nucleotide or amino acid sequences of the same gene in several species, and a species tree, reconciliation explains the non-congruence between these two trees by duplication events that have affected some lineages in the tree but not others. This analysis, contrary to the two preceeding, does not take into account the order of genes in the genomes.

### 1.5 Integration

We thus have three concepts as well as associated algorithmic methods, each pertinent to different aspects of the reconstruction of ancestral genomes:

- exemplar analysis, to handle multigene families in the comparison of two genomes,
- breakpoint median, providing ancestral gene order in the unique gene context, i.e., without multigene families, and
- reconciliation, for interpreting the gene tree of a multigene family in terms of a given species tree.

In this paper, we will show how to integrate these three ideas as a strategy for reconstructing ancestral gene orders, without restriction on copies of genes.

## 2 The data and the problem

Our starting point is a (binary rooted) phylogenetic tree $\mathcal{T}$ that recapitulates the evolutionary relationships among the $N$ species considered. Our problem is phrased in terms of genomic data at secondary and tertiary levels, i.e., genomic sequences already transformed by various analyses:

1. The detection of genes.
2. The determination of relationships of homology
   (a) to define multigene families, and
   (b) to produce gene orders based on sets of genes that are comparable between genomes.
3. Phylogenetic analysis within each multigene family to produce gene trees.

Analyses 1 and 2 allow us to rewrite the genomes as strings of gene symbols. These strings represent the order of the genes in the genomes, where a single symbol may represent the same gene in different genomes (orthologs) or copies, or related genes, within a single genome (paralogs). Analysis 3 allows us to determine the evolutionary relationships among all the genes, orthologs and paralogs, in the same multigene family.

In the framework of this study, the role of the reconciliation analysis is to allocate, to each ancestor in the species tree, the right number of copies of each gene, and to indicate, for each copy of a gene in a genome, its parent copy in the immediate ancestor genome. Figure 1 depicts this kind of information for two examples of multigene families. Although they are labeled by the same symbol in the figure, each copy of a gene in an ancestral genome is distinguished from the others by the positions of its offspring in its immediate descendant genomes and, eventually, in one or in several of the $N$ data genomes.

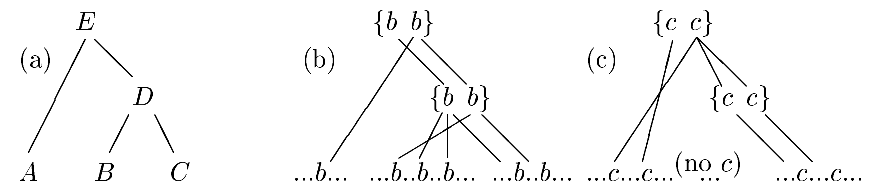


Figure 1: (a) Given species tree. Given genomes: $A, B$ and $C$. Ancestral genomes: $D$ and $E$ (root). (b) and (c) Two gene families, with an indication of the relationships of each copy. This is the kind of information produced by the reconciliation of gene trees to a species tree.

As mentioned above, we use exemplar analysis to calculate the distance between two genomes $X$ and $Y$, by retaining only one member–the exemplar–of each family of genes in a genome. The idea of this analysis is to identify those gene copies–the set of *true* exemplars–that have moved the least, each with respect to the others, during the divergence of $X$ and $Y$ through rearrangement. In this sense, these copies furnish the best indication, or the best reflection, of the original positions of the genes in the common ancestor of $X$ and $Y$; whence the designation "exemplar".

The same concept will be used to formalize the problem of ancestral gene order reconstruction. Suppose that a genome $G$ contains $r \geq 1$ genes homologs (copies), all descendants of a gene $x$ in $A$, the immediate ancestor of $G$. One of the $r$ copies is designated as being the exemplar of $x$ in $G$. If several copies of $x$ are present in $A$, then each will have its exemplar in $G$, as long as it has at least one descendant in $G$.

Given

- a phylogenetic tree $\mathcal{T}$ on $N$ species;
- their $N$ genomes: strings of symbols belonging to an alphabet of size $F$;
- $F$ gene trees, each relating all occurrences of one symbol in the $N$ genomes;
- a distance $d$ between two gene orders containing only unique genes,

the problem is to find, in each ancestral genome (internal node) of $\mathcal{T}$,

- its set of genes, as well as

- their relationships with respect to genes in the immediate ancestor,
- the order of these genes in the genome, and
- one gene in each family, designated as the exemplar,

such that the sum of the branch lengths of the tree $\mathcal{T}$ is minimal. The length of the branch connecting a genome $G$ to its immediate ancestor $A$ is $d(G', A)$, where $G'$ is the genome built from $G$ by deleting all but the exemplar from each family.

Note that the exemplars in a genome are defined with respect to the genome of the immediate ancestor. There are therefore no exemplars in the root $R$ (except for a modified version of the problem taking into account an outgroup). In fact, in the sum of the branch lengths of $\mathcal{T}$, distances $d(R, S')$ and $d(R, T')$ between $R$ and its two offpring $S$ and $T$ are replaced by $d(S', T')$. On the other hand, exemplars can and have to be determined in the $N$ data genomes.

# 3 Tools

Before presenting our solution to this problem, we describe in more detail the concepts of exemplar, median and reconciliation.

## 3.1 Exemplar analysis

Consider an alphabet of $F$ symbols. Let $G$ and $H$ be two signed ($+$ or $-$) strings of symbols, of lengths $n_G$ and $n_H$, respectively. For each symbol $a$, let $k_X(a)$ be the number of occurrences ($+$ or $-$) of $a$ in the genome $X$. Without loss of generality $k_G(a) > 0$ and $k_H(a) > 0$, and therefore $F \leq n_G$ and $F \leq n_H$. The set of the occurrences of the symbol $a$ in the two genomes constitutes a family of genes, or multigene family.

¿From the two genomes $G$ and $H$, we construct two reduced strings $g$ and $h$ by deleting all but one occurrence from each family of genes in each genome. These are the exemplar strings. Note that $h$ is a permutation of the symbols in $g$.

Consider two exemplar strings $g = g_1 \ldots g_F$ and $h = h_1 \ldots h_F$. We say that $g_i$ *precedes* $g_{i+1}$ in $g$. If the gene $a$ precedes $b$ in $g$, but does not precede $b$, nor does $-b$ precede $-a$, in h, then $a$ and $b$ determine a *breakpoint* in $g$. In addition there are supplementary breakpoints if $g_1 \neq h_1$ or $g_F \neq h_F$. The breakpoint distance between $g$ and $h$ is the number from breakpoints in $g$ (which is also the number of breakpoints in $h$). The *exemplar distance* between $G$ and $H$ is the minimum, over all choices of exemplar strings $g$ and $h$, of the breakpoint distance between $g$ and h.

**Example.** Let $G = -b - a\ b\ a - c\ d\ c,\ H = a - a\ c\ a - c\ b\ d$. Based on exemplar strings $-b - a - c\ d$ and $c\ a\ b\ d$, the exemplar distance is equal to 2.

We use a branch-and-bound algorithm (Sankoff, 1999) to calculate the exemplar distance. This treats one gene family at a time,inserting the new exemplars in the two strings already partially constructed. At each step, the exemplars are formed from the gene pair that least increases the distances between the two partial strings. As soon as the distance between the partial strings exceeds that the best complete exemplar strings already found during previous steps, these partial strings as well as all other pairs of strings that include them, are excluded, i.e., are not considered by the algorithm. This is justified by a monotonicity property of breakpoint distance: adding further gene families cannot decrease the exemplar distance–the increment of exemplar distance with the addition of new families is bounded below by zero.

What families should be examined first? Our strategy is to begin with families of size 2, for example $k_G(a) = 1$ and $k_H(a) = 1$, then those of size $3, 4$ , $\cdots$. This approach ensures that the depth-first search wastes as little time as possible evaluating short partial strings and attains meaningful temporary solutions (upper bounds) as rapidly possible. The search tree can then be pruned earlier and more often, postponing and avoiding the combinatorial explosion due to repeated examination of families with high of values $k_G \times k_H$.

## 3.2 The median of three genomes

Given

- a distance $d$ defined on gene orders of genomes,
- three genomes $A, B, C$ where each gene appears at most once in a genome, and a subset $E$ included in the union of genes of $A, B$ and $C$,

the median is the genome $M$ that contains the genes in $E$ and that minimizes $d(A, X) + d(B, X) + d(C, X)$. In the notation of Section 1.3, $\mathcal{S}$ = the set of all permutations on $E$.

### 3.2.1 Breakpoint medians

Let $d$ be the breakpoint distance. We present the reduction of the median problem to the TSP (Sankoff and Blanchette, 1997) for the case where $A, B, C$ and $M$ are unsigned genomes containing the same genes; the signed case is slightly more complicated.

Define $v(x, y)$ for an unordered pair of genes $(x, y)$ by

$$v(x, y) = |\{X \in \{A, B, C\} : x \text{ and } y \text{ adjacent in } X\}|.$$

Let

$$\Psi(S) = d(A, S) + d(B, S) + d(C, S).$$

The problem of determining the genome $S$ that minimizes $\Psi(S)$ is the same as determining a minimal length tour in terms of the distance matrix $\delta$, where $\delta_{x,y} = 3 - v(x, y)$.

The TSP is NP-hard; so is the median breakpoint problem (Pe'er and Shamir, 1998). Nevertheless, the relatively efficient algorithms available for the TSP can be applied to the median problem, as long as all the genomes are made up of the same set of genes.

To compare two genomes whose gene sets differ, we make use of the concept of *induced breakpoints* (cf Sankoff and Blanchette (1997)). First we delete all genes present in only one of the two genomes. Then we count the breakpoints in the genomes thus reduced (now of identical composition). The calculation of induced breakpoints is a subtler way to evaluate the parallelism of two gene orders than ordinary breakpoints. Indeed, if a breakpoint exists between $a_i$ and $a_j$ in the reduced genome, where $j \neq i+1$, this breakpoint cannot be detected in the original genome. (See Figure 2.)

<table>
<tr><td>genome $A$</td><td>1 4 5 6 7 3 8</td></tr>
<tr><td>genome $A$, reduced</td><td>1| 4 5 7 |3 8 |</td></tr>
<tr><td>genome $B$, reduced</td><td>1 | 3 8| 4 5 7 |</td></tr>
<tr><td>genome $B$</td><td>2 1 3 10 8 9 4 5 7</td></tr>
</table>

Figure 2: Induced breakpoints (vertical lines) in genomes with different gene sets.

In a phylogenetic analysis, the number of induced breakpoints can be a seriously biased estimate of the degree of evolutionary divergence if the genomes differ greatly in the total number of genes that they contain. To eliminate this problem, we normalize the number of induced breakpoints by the number of genes in common in the two genomes (Sankoff et al., 2000).

#### 3.2.2 A heuristic for the median problem

The reduction of the median breakpoint problem (normalized or not) to the TSP is no longer directly valid in the case of genomes containing different genes; TSP algorithms are no longer applicable. For this case we use a heuristic, which is computationally costly, but which proves to be rather accurate when verified with the help of good lower bounds Sankoff et al. (2000).

The median $M$ is constructed one gene at a time. Beginning with those genes of $E$ that are in all three genomes $A, B$ and $C$, we insert each gene at the point in the partially-constructed median where it least increases the sum of distances to $A, B$ and $C$. The process is repeated with genes that are present only in two of $A, B$ and $C$. Of course, here only two distances are involved. Finally, we add genes in $E$ that are in only one of the three genomes.

At each step of this algorithm, there may be several genes that could have been inserted with equivalent costs. The solution is improved by repeating the algorithm several times, altering these choices each time, and retaining the best final result.

### 3.3 Reconciliation of a gene tree with a tree species

An important new direction in the study of the phylogeny of a set of $N$ species is to consider all occurrences (orthologs and paralogs) of a gene $a$ in all the species, to produce a *gene tree* $T_a$ using standard phylogenetic procedures, and to reconcile this tree with a given species tree $\mathcal{T}$ (cf references in Section 1.4). The working hypothesis is that not only the species tree, but also the gene trees, are correct. Potential non-congruences between trees are explained by the duplication and loss of genes in particular lineages.

The first step in a reconciliation algorithm, and the one which concerns us, is to situate the duplications in the gene tree, and to locate them with respect to the speciation events in the species tree. This is done in terms of a projection $P$ from the set of nodes of $T_a$ onto the set of nodes and leaves of $\mathcal{T}$. For a node $\nu$ of $T_a$, let $R = R(\nu)$ be the subset of the $N$ genomes containing at least one copy of $a$ which is a descendant of $\nu$ in $T_a$. Now let $\mu$ be the most recent common ancestor in $\mathcal{T}$ of all the genomes in $R$. Then $P(\nu) = \mu$. A key property is: A node $\nu$ of $T_a$ corresponds to a duplication if and only if $P(\nu) = P(\nu_1)$ or $P(n) = P(\nu_2)$, where $\nu_1$ and $\nu_2$ are the offspring of $\nu$.

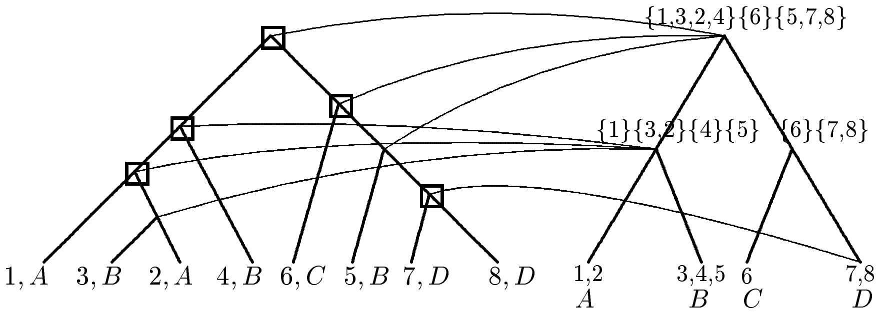


Figure 3: Projection $P$ from a gene tree $T_a$ (left) to the species tree $\mathcal{T}$ (right). Numbers correspond to the different copies of gene $a$; letters refer to the species (i.e., the genome). Duplication nodes in $T_a$ are drawn as squares. At each ancestral node, $P$ induces a number of groupings of copies of $a$ as indicated by the sets enclosed in braces. Each set refers to a distinct copy of $a$ which is inferred to have been present in that ancestral genome.

As illustrated in Figure 3, the projection $P$ helps us determine the duplication events in the gene tree, and to relate each of these events to an event in the species tree. More precisely, the fact that a duplication node $\nu$ of $T_a$ is associated to a node $\mu$ in $\mathcal{T}$ means that the duplication event ($\nu$) occurred before the speciation event ($\mu$).

Reconciliation has not yet been used in the analysis of whole genomes. Rather, each gene family has been treated independently, without worrying about gene order. The sole purpose has been to detect duplications. We use reconcilation to obtain somewhat different information: the number of copies of a gene in each ancestral genome of the species tree, as well as the historical relationships among these copies (as a preliminary to the reconstruction of gene order). How this is done is illustrated in Figure 3.

1. Associate to each ancestral node $\mu$ in $\mathcal{T}$ a *grouping* $\{a_1, a_2, \cdots a_i\}$ containing all copies of gene $a$ in all genomes in the subtree with root $\mu$. The initial grouping at the root of $\mathcal{T}$ contains all copies of the gene;

2. Starting with the root of the tree, at each ancestral node $\mu$,

   (a) Reproduce any groupings within the immediate ancestor of $\mu$, restricted to the subset of copies at $\mu$;

   (b) Further subdivide groupings thus obtained according to the duplications associated with node $\mu$ by the projection $P$.

Each grouping represents a single distinct copy of the gene which is inferred to have existed at the ancestral node.

# 4 Solution

The steinerization procedure for finding the ancestral gene orders in the species tree focuses on each ancestral node as the median of its three neighbours: its immediate ancestor, and its two descendants. A preliminary initialisation, random or other, is undertaken for each ancestral gene order. These orders are improved one by one by using, at each step, the most recent versions of the neighbours. After some iterations of all ancestors of the tree, we can hope that each ancestor will converge to a minimizing configuration.

## 4.1 Separation of the median and exemplar analyses

The calculation of breakpoint distances is not possible for genomes containing several copies of a same gene, simply because the question of adjacency becomes ambiguous. If gene $a$ is adjacent to gene $b$ in genome $G$, and in genome $H$ there are two copies of $a$, one adjacent to $b$ and one not, there is no clear way of deciding whether a breakpoint should be counted or not. This means that algorithms for finding the median cannot be applied to such genomes.

How can we transform the situation into one where each gene family is represented at most once in a genome, so that the median algorithm could be applicable? Consider two genomes $G$ and $H$, where $H$ is the immediate ancestor of $G$ in a species tree. For gene $a$, consider the groupings of genes in $G$ and $H$ produced by the reconciliation analysis. Because of duplications, each grouping of form $\sigma = \{a_1, \cdots, a_r\}$ in $H$ may give rise to $p \geq 2$ groupings in $G$. We consider, for the moment, only the gene subfamily containing one copy represented by $\sigma$ in $H$ and the corresponding $p$ copies in $G$. Then choosing an exemplar appropriately among the $p$ copies, repeating this procedure for any other groupings $\tau$ of gene $a$ in $H$, and so on for all the other gene families, is an evolutionarily principled way of removing the ambigities from the calculation of the breakpoints between $G$ and $H$. Note that by considering one subfamily at a time, every copy $\sigma, \tau, \cdots$ of every gene in $H$ is included in the calculation; the exclusion of non-exemplars only affects $G$.

Consider now a subtree consisting of genomes $A, B$, their common immediate ancestor $X$, and its ancestor $C$, as represented in Figure 4. The median calculation requires that all four genomes contain at most one member of each gene family. As in the preceding paragraph, we re-interpret this in terms of subfamilies, applying exemplar analysis to pick, for each copy of each gene in $X$, a single corresponding copy in $A$ (assuming there is at least one to pick) and a single copy in $B$ (if there is at least one). We do the same for the case where a gene copy in $C$ has several descendants in $X$, picking just one to be the exemplar.

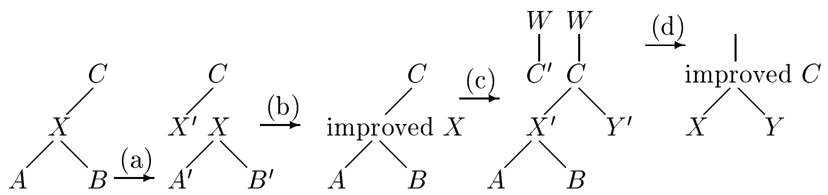


Figure 4: Alternating application of exemplar and median analyses. (a) and (c): Exemplar extraction. (b) and (d): calculation of the median.

Note that at the end of this choice procedure if an exemplar in $X$ with respect to some copy in $C$ also has exemplars in $A$ and $B$, all four copies are considered to be the same gene for the purposes of the subsequent median analysis. Note as well that a copy in $A$ may have correspondences *via* the exemplar relationship in $X$ and $B$ only, $X$ and $C$ only, in $X$ only or in none of $B, C$ or $X$. An analogous statement may be made for $B$. Even $C$ may contain a copy which has no descendant in $X$ and thus none in $A$ or $B$. But if it does have an exemplar in $X$, it will also correspond to a copy in either $A$ or $B$ or both. These possibilities are illustrated in Figure 5.

We are now in a position to use the median algorithm to calculate an improved gene order for genome $X$.

### 4.2 Iterating the overlapping median analyses

In Figure 4, it can be seen that the once the median analysis has reordered the genes in $X$, the previous choice of exemplars in genomes $A, B$ and $X$ may no longer be optimal. We need not change the designation of $A'$ and $B'$ as exemplar strings of $A$ and $B$, but in improving $C$, we may have to change the exemplars in $X$. Such a change will not affect the quantities $d(A', X)$ and $d(B', X)$. By accepting only changes that improve $d(X', C)$ we can ensure that the quality of the solution increases.

The origin of the phylogenetic signal is in the $N$ given genomes (the leaves

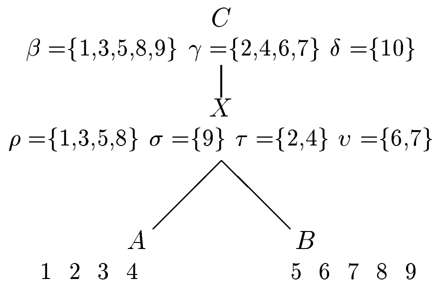


Figure 5: Subtree consisting of genomes $A, B$, their common immediate ancestor $X$, and its ancestor $C$. The groupings in braces represent a possible output from the reconciliation analysis. Each grouping represents a gene copy whose descendants in $A$ and $B$ are just the copies listed between the braces. Possible exemplar choices 1 and 4 in $A$ to reflect $\rho$ and $\tau$, respectively, 5, 7 and 9 in $B$ to reflect $\rho, \upsilon$, and $\sigma$, respectively, and $\rho$ and $\tau$ in $X$ to reflect $\beta$ and $\gamma$, respectively, lead to five sets of corresponding copies: (i) 1 in $A$, 5 in $B$, $\rho$ in $X$, $\beta$ in $C$; (ii) 4 in $A$, $\tau$ in $X$ and $\gamma$ in $C$; (iii) 7 in $B$, $\upsilon$ in $X$; (iv) 9 in $B$, $\sigma$ in $X$; (v) $\delta$ in $C$ (the other descendant of $C$, not shown, must have a descendant of $\delta$). Copies $\rho, \sigma, \tau$ and $\upsilon$, must be ordered by the median analysis together with copies of all the other genes in $X$, based on the orders in $A, B$, and $C$.

of the species tree). Applying our analyses, first to the ancestors that have two descendants among the $N$, and then to the ancestors whose descendants have already been improved, the signal is transmitted towards the root during a single pass of the ancestral nodes of a tree. But for the signal from each of the leaf nodes to "interact" with that of each other, necessary if a global solution is hoped for, it also has to be transmitted back from the root. The speed of this transmission will be only one generation per pass (e.g., in Figure 4, from $W$ to $C$ in one pass, then from $C$ to $X$ in the next pass). Therefore, for a balanced tree, the time required to propagate the signal between two distant genomes, descendants of different offspring of the root, is of the order of $\log N$. This is a rough indication of the number of passes necessary before the convergence.

# 5 Discussion

The details of this paper have been phrased in terms of breakpoints, since this is the only measure of genomic difference for which a feasible extension to phylogeny exists. The main contribution of this work, however, is the discovery of the natural and unexpected meshing of the three concepts of reconciliation, median, and exemplar, all developed for other purposes. Thus, the analysis could all be rephrased in terms of other genomic distances, such as signed reversals distance (Hannenhalli and Pevzner, 1995a), for which exemplar extraction works efficiently, but for which median analysis is impractical for the moment. And in any approach requiring the optimization of ancestors, it will always be necessary to settle the question of content (what genes, how many copies) and the problem of multigene families. This is why reconciliation (or a similar method), as well as exemplars (or an analogous analysis), seem to be essential to the study of phylogeny for this kind of data.

Where is such data to be found? It is true that numerous genomes have been sequenced. Primary analyses, however–the detection of all genes and the construction of multigene families–are far from complete, except for some small genomes such as those in organelles. For the moment, our method remains a tool awaiting the development of appropriate data before being applied.

## Acknowledgments

Research supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Genome Analysis and Technology program. DS is a Fellow, and N E-M a Scholar, in the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

## References

Caprara, A. (1997). Sorting by reversals is difficult. In *Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 97)*, pages 75–83, New York. ACM.

Caprara, A. (1999). Formulations and hardness of multiple sorting by reversals. In Istrail, S., Pevzner, P., and Waterman, M., editors, *Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB 99)*, pages 84–93, New York. ACM.

Chen, K., Durand, D., and Farach-Colton, M. (2000). Notung: Dating gene duplications using gene family trees. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB 2000)*, New York. ACM.

Eulenstein, O., Mirkin, B., and Vingron, M. (1997). Comparison of annotating duplications, tree mapping, and copying as methods to compare gene trees with species trees. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 37:71–93.

Guigó, R., Muchnik, I., and Smith, T. (1996). Reconstruction of ancient molecular phylogeny. *Molecular Phylogenetics and Evolution*, 6:189–213.

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

Hannenhalli, S. and Pevzner, P. (1995a). 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.

Hannenhalli, S. and Pevzner, P. (1995b). Transforming men into mice. In *Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science*, pages 581–592.

Ma, B., Li, M., and Zhang, L. (1998). On reconstructing species trees from gene trees in term of duplications and losses. In Istrail, S., Pevzner, P., and Waterman, M., editors, *Proceedings of the Second Annual International Conference on Computational Biology (RECOMB 98)*, pages 182–191, New York. ACM.

Page, R. (1994). Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, 43:58–77.

Page, R. and Charleston, M. (1997). Reconciled trees and incongruent gene and species trees. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 37:57–70.

Pe'er, I. and Shamir, R. (1998). The median problems for breakpoints are np-complete. Electronic Colloquium on Computational Complexity Technical Report 98-071, http://www.eccc.uni-trier.de/eccc.

Reinelt, G. (1991). *The traveling salesman - computational solutions for TSP applications*. Springer, New York.

Sankoff, D. (1999). Genome rearrangements with gene families. *Bioinformatics*, 15:909–917.

Sankoff, D. and Blanchette, M. (1997). The median problem for breakpoints in comparative genomics. In Jiang, T. and Lee, D., editors, *Computing and Combinatorics, Proceeedings of COCOON '97*, number 1276 in Lecture Notes in Computer Science, pages 251–263, Berlin. Springer.

Sankoff, D., Bryant, D., Deneault, M., Lang, B., and Burger, G. (2000). Early eukaryote evolution based on mitochondrial gene order breakpoints. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB 2000)*, New York. ACM.

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

Centre de recherches mathématiques, Université de Montréal, CP 6128 succursale Centre-Ville, Montréal, Québec H3C 3J7.
*E-mail:* sankoff@ere.umontreal.ca

Département d'informatique et de recherche opérationnelle, Université de Montréal, CP 6128 succursale Centre-Ville, Montréal, Québec H3C 3J7.
*E-mail:* mabrouk@iro.umontreal.ca