

Models and Methods in Comparative Genomics

Guillaume Bourque¹ and Louxin Zhang²

¹ Genome Institute of Singapore, 60 Biopolis Street
#02-01 Genome, Singapore 138672
bourque@gis.a-star.edu.sg

² Department of Mathematics, National University of Singapore
2 Science Drive 2, Singapore 117543
matzlx@nus.edu.sg

Abstract. Comparative genomics is the analysis and comparison of genomes from different species. In recent years, in conjunction with the growing number of available sequenced genomes, this field has undergone a rapid expansion. In the current survey, we review models and methods associated with four important research topics in this area that have interesting computational and statistical components: genome rearrangements, gene duplication, phylogenetic networks and positional gene clustering. The survey aims at balancing between presenting classical results and promising new developments.

1 Introduction

Advances in sequencing and comparative mapping have enable a new period in biology referred to as the *Genomics Era*. The tremendous efforts invested into this domain have provided the community with the complete sequence of whole genomes for a wide range of organisms ranging from atypical bacteria living in harsh conditions to large eukaryotic genomes such as Human [94] and Mouse [159].

Although the availability of these genomic sequences has facilitated great leaps in our understanding of many biological processes, they have also highlighted the complexity of fully deciphering genomes. Even questions as simple as determining the exact number of genes in the human genome have turned out to be quite difficult [146]. In this context, it is of no surprise that more elaborate problems, such as fully understanding how genes are being regulated (e.g. which genes are expressed, when are they expressed, etc.), have remain very challenging. This last question in particular is important because a lot of the animal diversity is thought to be harbored in gene regulation [99]. Other similarly exciting questions will hopefully incite the development of computational and statistical tools that will help further our comprehension of the forces that shape modern genomes.

Decoding the sequenced genomes is analogous to decoding a hard disk with no information about the file structure or even the type of information that is encoded. The difference is that we expect the challenge of decoding the genome

to be greater given the complexity of the final product. In this setting, comparative genomics comes in as a powerful tool contrast and recognize some of the features that play a crucial role in the different genomes. By the identification of similarities and differences, the hope is that we will gain a first handle on some of these important problems [124]; this is the field of Comparative Genomics.

The chapter focuses on mathematical and algorithmic aspects of four general research topics in comparative genomics. In Section 2, we first introduce models and methods used in the analysis of genome rearrangements. In this section we present some of the similarity measures use to study gene order conservation across genomes. This section also includes some of the details of the Hannenhalli-Pevzner algorithm for computing the inversion distance between two genomes, probably one of the strongest algorithmic result in computational molecular biology. We also present two recently introduced alternative model for genome evolution: the block-interchange and the double-cut-and-join operation. Finally, we summarize the recent progress in genome rearrangement with gene family and with partial order genomes. In Section 3, we summarize the mathematical models for dating both large scale genomic duplications and tandem duplications. In Section 4, we presents three different network models for studying horizontal gene transfer, recombination and other reticulations. We also summarize different methods for reconstructing these networks from gene trees and sequences. In Section 5, we point out two basic statistical models for analytically testing positional gene clusters.

2 Genome rearrangements

The study of genome rearrangements is the analysis of mutations affecting the global architecture of genomes as oppose to local mutations affecting individual regions. This type of analysis dates back to the early 1920s with pioneering studies on the evolution of the *Drosophila* genome (e.g. [114]). To study genome rearrangements, it is in general sufficient to view genomes as permutations on a set of markers common to the group of genomes. Different types of marker can be used for this purpose but the key is that these markers must be unambiguously identifiable across genomes. In most of the current section (except 2.4), we restrict the comparison to a common set of markers such that each marker is found exactly once in each genome. See also Section 3 for other results where this restriction is alleviated to allow unequal marker content.

An obvious set of markers than can be use for this purpose is the set of genes observed in the group of genomes. Based on sequence similarity, a set of homologous genes can be identified across genomes and by labeling these genes from 1 to n , one can obtain permutations that encapsulate the relative order of these markers in the genomes. If available, the relative orientation can also be associated to the set of markers (e.g. for genes-based markers, the direction of transcription can be used) and leads to *signed permutations*. Otherwise *unsigned permutations* will be used. For an example with two small mitochondria DNA (mtDNA) genomes and their signed permutations see Figure 1. By convention,

obtain a distance measure, we actually use breakpoints which are pairs of markers adjacent in one genome but not the other. This easily calculated measure was first explicitly presented in the context of genome rearrangements by Watterson et al. 1982 [162]. Formally, we get:

Definition 1. *The breakpoint distance, $b(\pi)$, is the number of pairs (π_i, π_{i+1}) , $0 \leq i \leq n$ such that $(\pi_{i+1} - \pi_i) \neq 1$ where $\pi_0 = 0$ and $\pi_{n+1} = n + 1$.*

In the example shown in Figure 1, there are 9 such pairs:

$$(3, 5), (5, -10), (-10, 11), (11, 4), (4, 9), (9, 7), (8, 12), (12, 6), (6, 13),$$

and so $b(\pi) = 9$. See also Figure 2.

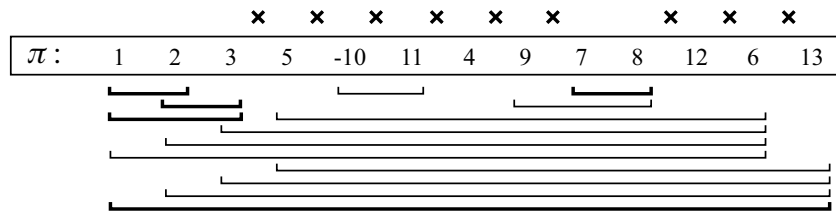


Fig. 2. Breakpoints, common and conserved intervals for π , the permutation associated with Earthworm (see Figure 1). The 9 breakpoints are indicated using crosses above the permutation. The 14 common intervals are shown below the permutation. The 5 conserved intervals correspond to a subset of the common intervals and are shown in black.

Common and conserved intervals Recently, two new criteria were introduced as an extension of the breakpoint distance to measure the similarity between sets of genomes: common intervals [156, 77] and conserved intervals [10]. We introduce these definitions in the context of two general permutations of size n , π and γ . When one of these permutation is the identity we have $\gamma = I$ as above.

Definition 2. *A common interval is a set of two or more integers that is an interval in both π and γ .*

Uno and Yagiura [156] presented three algorithms for finding all common intervals between two permutations π and γ : two simple $O(n^2)$ time algorithms and one more complex $O(n + K)$ time algorithm where $K \leq \binom{n}{2}$ is the number of common intervals between π and γ . Heber and Stoye [77] extended on this result by developing an algorithm to find the common intervals in a family of m permutations in optimal $O(nm + K)$ time where K is the number of common intervals in the m permutations defined similarly to Definition 2.

To continue with the example of the two mtDNA shown in Figure 1, we see that these two permutations harbor 14 common intervals displayed only in π in Figure 2. For two permutations of size n , the maximum number of common intervals is $\binom{n}{2}$ and so, in this case, the maximum would have been $\binom{13}{2} = 78$ common intervals.

We now define a conserved interval between two arbitrary permutations as introduced by Bergeron and Stoye [10].

Definition 3. A conserved interval $[a, b]$ is an interval such that a precedes b , or $-b$ precedes $-a$ in both π and γ , and the set of elements, without signs, between a and b is the same in both π and γ .

Conserved intervals are common intervals with additional constraints on their endpoints. In Figure 2 we see that 5 of the 14 common intervals of the two mtDNA also qualify as conserved intervals.

Although the definition of conserved intervals may seem unnatural at first, it is intimately connected to the concept of *subpermutations* [73] in the Hannenhalli-Pevzner Theory (see Section 2.2). Moreover, it was shown that it can be used to efficiently sort permutations by reversals [11].

2.2 Edit distance between two genomes

In the early 90s, a series of paper revived the interest in the problem of computing the edit distance between a pair of genomes under different edit operations [137, 89, 14]. This problem had been posed by Watterson et al. [162] and even earlier in the genetics literature (e.g. [150]). Examples of edit operations that are frequently considered are displayed in Table 1. These operations can be considered separately or in different combinations and lead to different models of evolution of gene order. We now review some of the key results in this area and also present recent advances.

Reversal distance The reversal-only, or inversion-only, edit distance is probably the most studied edit distance in the context of gene order. Initially, we will focus on the problem of computing the reversal between two signed permutations.

Definition 4. Given a permutation π , a reversal $\rho_{i,j}$, $1 \leq i, j \leq n$, applied to π produces:

$$\rho_{i,j}(\pi) = \pi_1 \dots \pi_{i-1} - \pi_j \dots - \pi_i \pi_{j+1} \dots \pi_n$$

In this context, the *reversal distance*, $d_{rev}(\pi)$, is defined as the minimum number of reversals required to convert π into the identity permutation I . Since every reversal can reduce by at most two the number of breakpoints, a trivial first result is the following:

Lemma 1.

$$d_{rev}(\pi) \geq \frac{b(\pi)}{2}$$

| Mutation Type | Before | After |
|-------------------|---------------------------------------|---------------------------------------|
| Reversal | 1 2 3 4 <u>5 6 7</u> 8 9 10 | \Rightarrow 1 2 3 4 -7 -6 -5 8 9 10 |
| Translocation | <u>1 2 3 4 5 6</u> <u>7 8 9 10</u> | \Rightarrow 7 8 5 6 1 2 3 4 9 10 |
| Fusion | <u>1 2 3 4 5 6</u> <u>7 8 9 10</u> | \Rightarrow 1 2 3 4 5 6 7 8 9 10 |
| Fission | <u>1 2 3 4 5 6 7 8 9 10</u> | \Rightarrow 1 2 3 4 5 6 7 8 9 10 |
| Transposition | 1 <u>2 3 4 5</u> _ 6 7 8 9 10 | \Rightarrow 1 4 5 2 3 6 7 8 9 10 |
| Block Interchange | 1 <u>2 3</u> 4 5 <u>6 7 8</u> 9 10 | \Rightarrow 1 6 7 8 4 5 2 3 9 10 |

Table 1. Examples of chromosomal mutations, or edit operations, affecting gene order.

To obtain an exact formula to compute the reversal distance, we now present a summary of the terminology frequently referred to as the *Hannenhalli-Pevzner Theory* [14, 74]. First, we convert π , a signed permutation, into π' , an unsigned permutation, by mimicking every directed element i by two undirected elements i^t and i^h representing the tail and the head of i . Since π is a permutation of size n , π' will be a permutation of size $2n$. The permutation π' is then extended by adding $\pi'_0 = 0$ and $\pi'_{2n+1} = n + 1$. Next, we construct the breakpoint graph associated with π .

Definition 5. *The breakpoint graph of π , $G(\pi)$, is an edge-colored graph with $2n + 2$ vertices. Black edges are added between vertices π'_{2i} and π'_{2i+1} for $0 \leq i \leq n$. Grey edges are added between i^h and $(i + 1)^t$ for $0 < i < n$, between 0 and 1^t , and between n^h and $n + 1$.*

In the breakpoint graph, black edges correspond to the actual state of the permutation while grey edges correspond to the sorted permutation we seek. See Figure 3 for an example.

Bafna and Pevzner [14], and later Hannenhalli and Pevzner [74], showed that $G(\pi)$ contains all the necessary information for efficiently sorting the permutation π . The first step is to look at the maximal cycle decomposition of the breakpoint graph. Finding the maximal cycle decomposition of a graph in general can be a very difficult problem but, fortunately, because of the way the breakpoint graph was constructed for a signed permutation, each vertex has degree two and so the problem is trivial. Suppose $c(\pi)$ is the maximum number of edge-disjoint alternating cycles in $G(\pi)$. The cycles are *alternating* because, in the breakpoint graph of a signed permutation, each pair of consecutive edges always has different colors. We then get:

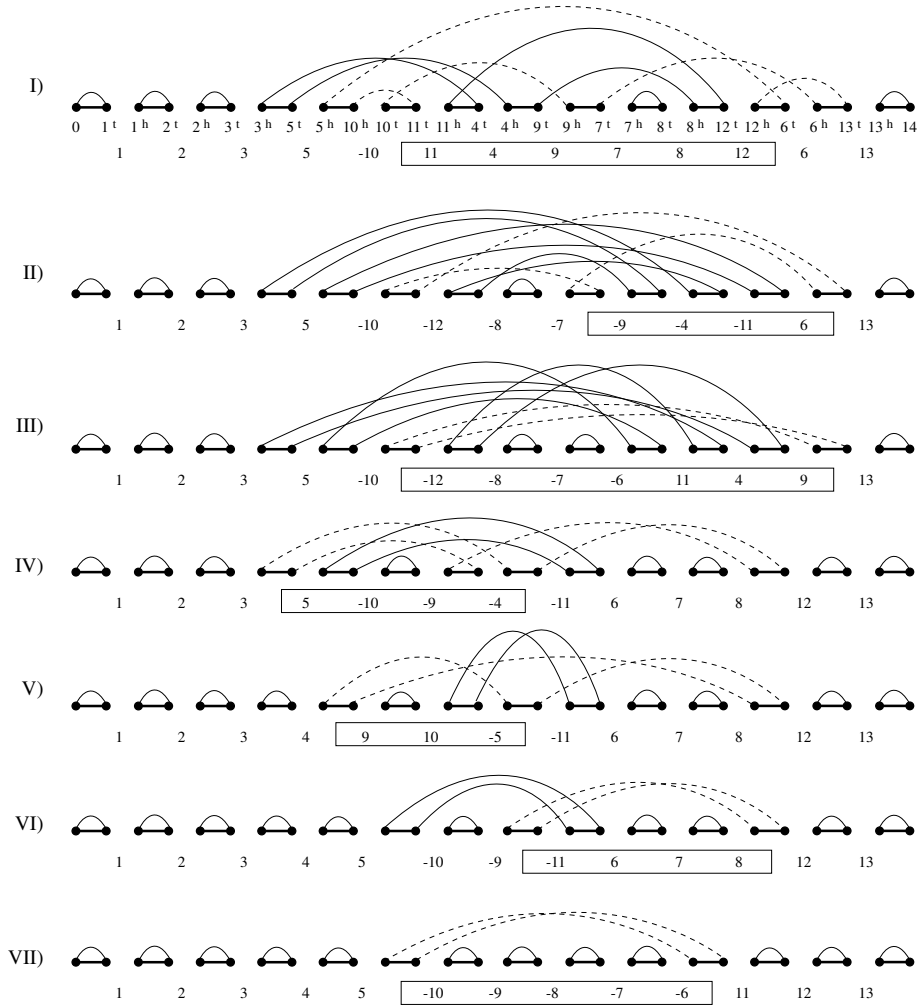


Fig. 3. I) Breakpoint graph associated with the two permutations from Figure 1. Black edges are shown using thick lines. All other lines (both solid and dashed) correspond to grey edges. I-VII) A sequence of sorting reversals. The fragment of the permutation to be inverted is shown using a box. Dashed lines are used to highlight the cycle that will be affected by the reversal. Black edges represent the “current” state of the permutation to be sorted while grey edges represent the “desired” state corresponding to a sorted permutation. Note that in the final stage (not shown) the black and grey edges are perfectly matched.

Lemma 2. ([14, 89])

$$d_{rev}(\pi) \geq n + 1 - c(\pi).$$

An edge in $G(\pi)$ is said to be *oriented* if it spans an odd number of vertices (when the vertices of $G(\pi)$ are arranged in the canonical order $\pi'_0, \dots, \pi'_{2n+1}$). A cycle is said to be *oriented* if it contains at least one oriented gray edge. Cycles which are not oriented are said to be *unoriented* unless they are of size 2 in which case they are said to be *trivial*. The term oriented comes from the fact that if we traverse an oriented cycle we will traverse at least one black edge from left to right and one black edge from right to left. In the breakpoint graph shown in Figure 3 (I), there are only two non-trivial cycles: one where the gray edges are displayed using solid lines and one where the gray edges are displayed using dashed lines. The cycle with solid lines is unoriented since it does not contain an oriented edge but the cycle with dashed lines is oriented because it contains an oriented edge (e.g. $(10^h, 11^t)$).

For each grey edge in $G(\pi)$ we will now create a vertex v_e in the *overlap graph*, $O(G(\pi))$. Whenever two grey edges e and e' overlap or cross in the canonical representation of $G(\pi)$, we will connect the corresponding vertices v_e and $v_{e'}$. A *component* will mean a connected component in $O(G(\pi))$. A component will be *oriented* if it contains a vertex v_e for which the corresponding grey edge e is oriented. As for cycles, a component which consists of a single vertex (grey edge) will be said to be *trivial*. In Figure 3 (I), there are 5 trivial components and one larger oriented component since at least one of its grey edge is oriented. The challenge in sorting permutations comes from unoriented components.

Unoriented components can be classified into two categories: hurdles and protected nonhurdle. A *protected nonhurdle* is an unoriented component that separates other unoriented components in $G(\pi)$ when vertices in $G(\pi)$ are placed in canonical order. A *hurdle* is any unoriented component which is not a protected nonhurdle. A hurdle is a *superhurdle* if deleting it would transform a protected nonhurdle into a hurdle, otherwise it is said to be a *simple hurdle*. Finally, π is said to be a *fortress* if there exists an odd number of hurdles and all are superhurdles in $O(G(\pi))$ [143]. We then get the main result of the HP theory:

Theorem 1. ([72])

$$d_{rev}(\pi) = n + 1 - c(\pi) + h(\pi) + f(\pi),$$

where $h(\pi)$ is the number of hurdles in π and $f(\pi)$ is 1 if π is a fortress and 0 otherwise.

For instance, using Figure 3 (I), we see that the reversal distance between Human mtDNA and Earthworm mtDNA is $d_{rev}(\pi) = 13 + 1 - 7 + 0 + 0 = 7$. In [72], Hannenhalli and Pevzner also showed how to recover an optimal sequence of sorting reversals using the breakpoint graph in $O(n^4)$ (see Figure 3 (II-VII) for an example).

Since these initial results, there has been a number of improvement on the performance of these algorithms. Berman and Hannenhalli [17] improved the bound for the sorting problem to $O(n^2\alpha(n))$ (where α is the inverse of Ackermann's function) and Kaplan et al. [85] reduced it further to $O(n^2)$. Later, Bader et al. [4] showed that without recovering an actual optimal sequence of steps,

the reversal distance can be computed in linear time ($O(n)$). Finally, Bergeron and Stoye [10] have described an alternative sorting algorithm that takes $O(n^2)$ but bypass much of the complexity of the earlier algorithms.

So far, the discussion was centered around the problem of sorting two signed permutations. In this context, it is interesting to highlight the following result by Caprara [31].

Theorem 2. ([31]) *The problem of sorting an unsigned permutation by the minimum number of reversals is NP-Hard.*

Transposition distance A transposition is an edit operation, in which a segment is cut out of the permutation, and pasted in a different location (for an example, see Table 1).

Definition 6. *Given a permutation π , a transposition is an operation $\theta_{i,j,k}$, $1 \leq i, j \leq n$, $k < i$ or $k > j$, that once applied to π produces:*

$$\theta_{i,j,k}(\pi) = \pi_1 \dots \pi_{i-1} \pi_{j+1} \dots \pi_{k-1} \pi_i \dots \pi_j \pi_{k+1} \pi_n$$

The problem of sorting by transposition was first studied by Bafna and Pevzner [13] who presented a 1.5 approximation algorithm which runs in $O(n^2)$ time. Using an alternative data structure, Walter et al. [158] developed a 2.25 approximation algorithm for the same problem. More recently, Elias and Hartman [50] improved on these bounds by presenting a 1.375 approximation algorithm that required an elaborate computer assisted proof. The complexity of sorting by transpositions remains an open problem.

Block interchange distance The notion of a block interchange operation in the context of genome rearrangements was introduced by Christie [38]. In a block-interchange, two non-intersecting substrings of any length are swapped in the permutation. This type of event can be viewed as a generalized transposition.

Definition 7. *Given a permutation π , a block-interchange is an operation $\beta_{i,j,k,l}$, where $1 \leq i < j \leq k < l \leq n$, that once applied to π produces:*

$$\beta_{i,j,k,l}(\pi) = \pi_1 \dots \pi_{i-1} \pi_k \dots \pi_l \pi_{j+1} \dots \pi_{k-1} \pi_i \dots \pi_j \pi_{l+1} \pi_n$$

Note that the special case of $j = k$ leads to an alternative and equivalent definition of a transposition.

Christie [38] showed that by considering the block interchange operation, one can efficiently sort unsigned permutations in $O(n^2)$. This algorithm can also serve as a 2-approximation algorithm for the problem of sorting by transpositions.

Theorem 3. ([38]) *The block-interchange distance for an unsigned permutation, $d_{BI}(\pi)$, is*

$$d_{BI}(\pi) = \frac{1}{2}[(n+1) - c(\pi)],$$

where $c(\pi)$ is the number of alternating cycles in the cycle graph of π (Note here that the cycle graph is defined slightly differently since the permutations are unsigned, see [38]).

Recently, the analysis of block-interchanges was revisited by Lin et al. [101]. By focusing on circular chromosomes (such as the mtDNA in Figure 1), that are also unsigned, and making use of permutations groups in algebra, they designed an algorithm for sorting by block-interchanges with time-complexity $O(\delta n)$, where δ is the minimum number of block-interchanges required for the transformation and can be calculated in $O(n)$ time in advance. The approach was also implemented in a tool called ROBIN [102].

Taking the permutations displayed in Figure 1 and treating them as unsigned, we can compute an optimal scenario with 3 block-interchange operations, see Table 2.

| | | | | | | | | | | | | | |
|--------------------|---|---|---|---|----|----|----|----|----|----|----|----|----|
| Earthworm | 1 | 2 | 3 | 5 | 10 | 11 | 4 | 9 | 7 | 8 | 12 | 6 | 13 |
| $\beta_{4,6,8,10}$ | 1 | 2 | 3 | 5 | 10 | 11 | 4 | 9 | 7 | 8 | 12 | 6 | 13 |
| $\beta_{4,4,7,12}$ | 1 | 2 | 3 | 9 | 7 | 8 | 4 | 5 | 10 | 11 | 12 | 6 | 13 |
| $\beta_{6,8,9,12}$ | 1 | 2 | 3 | 4 | 5 | 10 | 11 | 12 | 6 | 7 | 8 | 9 | 13 |
| Human | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |

Table 2. Sorting using bloc-interchanges the two permutations, viewed as unsigned, associated with the mtDNA displayed in Figure 1. The scenario requires only 3 steps.

Reversal, translocation, fusion and fission distance The results presented thus far have been centered around unichromosomal genomes, i.e. genomes that have a single chromosome. Table 1 shows examples of events that specifically affect genomes with multiple chromosomes, mainly: translocations, fusions and fissions.

Kececioglu and Ravi [88] began the investigation of translocation distances by giving a 2-approximation algorithm for multichromosomal genomes when the orientation of the genes are unknown (“unsigned permutation”). We also use the terminology permutation when dealing with multichromosomal genomes because, by using special markers to delimitate chromosome boundaries, it is still possible to represent such genomes using permutations. For signed permutation, Hannenhalli and Pevzner [73] derived an equation related to Theorem 1 to compute the rearrangement distance between two multichromosomal genomes when permissible operations are: reversals, translocations, fusions and fissions. We refer the reader to Pevzner [133] and Tesler [154] for the details of the calculation but we will briefly present how the formula can be obtained.

The main idea to compute the rearrangement distance between two multichromosomal genomes Π and Γ is to concatenate their chromosomes into two permutations π and γ . The purpose of these concatenated genomes is that every rearrangement in a multichromosomal genome Π can be mimicked by a reversal

in a permutation π . In an *optimal* concatenate, sorting π with respect to γ actually corresponds to sorting Π with respect to Γ . Tesler [154] also showed that when such an optimal concatenate does not exist, a *near-optimal* concatenate exists such that sorting this concatenate mimics sorting the multichromosomal genomes and uses a single extra reversal which corresponds to a reordering of the chromosomes. The algorithm was implemented into a program called GRIMM [155]. Ozery-Flato and Shamir [127] identified a case where the algorithm does not apply but also suggested a correction.

Double-cut-and-join distance Recently, in an attempt to reconcile the various edit distances, Yancopoulos et al. [165] presented a universal edit operation, the *double-cut-and-join* (DCJ), that could seamlessly model inversions, transpositions, translocations, fusions and fissions. The last two had already been identified as special cases of translocations [73]. This elementary operation is a local operation on four markers initially forming two adjacent pairs. It consists of cutting two adjacencies in the first genome and rejoining the resulting four unconnected markers to form two new pairs [165].

Under this model, any rejoining is *proper* as long as $b(\pi) - c(\pi)$ is reduced by 1. The major difference with the HP-Theory presented above is that some of the proper ways of reconnecting these two pairs cannot be associated with a reversal (or a reversal mimicking a translocation). Actually, some of these operations lead to the creation of a circular intermediate (CI). Reabsorbing the CI actually correspond to doing a block-interchange (see [165]) but since it required two steps, it will be associated with a weight of two in the final edit scenario.

Theorem 4. ([165]) *The double-cut-and-join distance for a permutation π , $d_{DCJ}(\pi)$, is*

$$d_{DCJ}(\pi) = b(\pi) - c(\pi)$$

2.3 Genome rearrangements with multiple genomes

Extending the two way measures and edit distance algorithms to multiple genomes has proven to be challenging. Formally, the problem is the following:

Definition 8. *Given a set of m genomes, the Multiple Genome Rearrangement problem is to find an unrooted tree T , where the m genomes are leaf nodes, and assign internal ancestral nodes such that $D(T)$ is minimized where:*

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

and $d(\pi, \gamma)$ can be any distance measure discussed in Sections 2.1-2.2.

The problem is also known as the problem of reconstructing the most parsimonious phylogenetic tree under the metric d (see Figure 4 for an example).

The simplest extension, the case with $m = 3$ signed permutations, also called the *Median problem*, was shown to be NP-Hard [32] for both the breakpoint

distance ($d = b$) and the reversal distance ($d = d_{rev}$). Nonetheless, we now present a few heuristic that have been developed for this problem under various distance metrics.

Breakpoint phylogenies Sankoff and Blanchette [139] studied the median problem for the breakpoint distance; they showed how the problem could be reduced to an instance of the Traveling Salesman Problem (TSP), a problem for which reasonably efficient algorithms are available. Using this result, Blanchette et al. [19] developed **BPA**nalysis, a method to recover the most parsimonious scenario for m genomes under the breakpoint distance. The approach was to look for an optimal assignment of internal nodes for a given topology by solving a series of median problem (this is also known as the small parsimony problem). The next step in the approach was to scan the space of all possible tree topologies to find the best tree (large parsimony problem). One of the downside of this approach is that the tree space quickly becomes prohibitive. This limitation was partially addressed by Moret et al. [112] who developed **GRAPPA** which, by computing tight bounds, was able to efficiently prune the tree space.

Conservation phylogenies The first method that use the concept of conserved intervals as the criterion for the phylogenetic reconstruction problem was presented by Bergeron et al. [12]. Even though the problem was restricted to finding an optimal assignment of internal nodes on a fixed phylogeny (small parsimony problem), this is an auspicious area of research.

Rearrangement phylogenies Siepel and Moret [113] also studied the median problem but under a different metric: the reversal distance. They presented a branch-and-bound algorithm to prune the search space using simple geometric properties of the problem. Concurrently, Bourque and Pevzner [24] implemented a method called **MGR** for both the median and the full phylogeny by making use of properties of additive or nearly additive trees. This approach, combined with **GRIMM** [155] was shown to be applicable to both unichromosomal genomes for the reversal distance [24] and for multichromosomal genomes for a rearrangement distance that combines reversals, translocations, fusions and fissions [24, 25]. In a recent analysis [116], this algorithm was applied to 7 mammalian genomes (Human, Mouse, Rat, Cat, Dog, Pig, Cow) and for which the recovered unrooted tree is shown in Figure 4 (see [116] for full rearrangement scenario including recovered ancestral genomes).

2.4 Genome rearrangement with gene families

As we have seen in the last several sections, each genome is viewed as a permutation in which each gene has exactly one copy in the traditional study of genome rearrangement. While this may be appropriate for small viruses and mitochondria genomes, it may not realistic when applied to eukaryotic genomes where

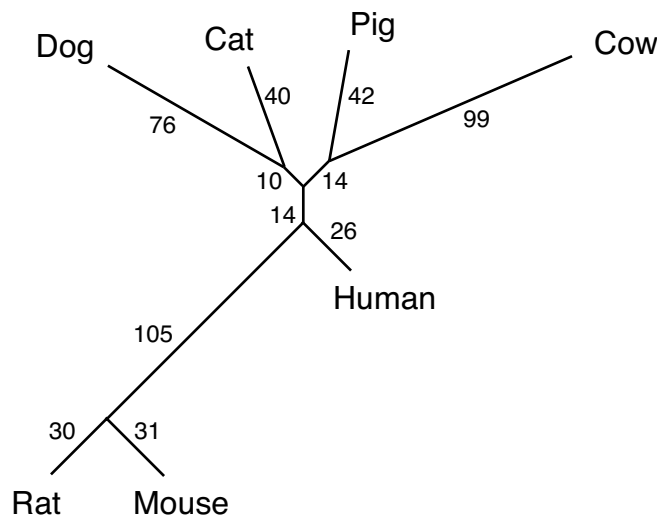


Fig. 4. Unrooted binary tree showing phylogenetic relationships between 7 mammalian genomes recovered by MGR. The number on each edge indicates the minimum number of rearrangements (reversals, translocations, fusions, fissions) required to convert between the two genomes connected by the edge. Extracted from Murphy et al. [116].

paralogous genes often exist. Hence, a more generalized version of the genome rearrangement problem was proposed by Sankoff [140] where multiple copies of the same gene can now be found in the same genome. His idea is to delete all but one member of a gene family in each genome, so as to minimize the total breakpoint (or other) distance between the reduced genomes. In this approach, the retained copies are called exemplars. One of the applications of the exemplar problem is in orthologous gene assignment [33, 151]

Even though it is almost trivial to calculate the breakpoint distance between two genomes with single-gene families, the exemplar problem for breakpoint distance is not only NP-hard [29], but also unlikely to have polynomial-time constant-ratio approximation algorithm unless $NP=P$ [122, 34]. Nevertheless, by observing the monotonicity of the exemplar problem, Sankoff proposed a branch-and-bound method to tackle this problem [140]. His idea is to work on each gene family separately, choosing the pair of exemplars that least increases the distance when inserted into the partial exemplar genomes already constructed.

Recently, Nguyen, Tay and Zhang [123] proposed a divide-and-conquer approach to calculating the exemplar breakpoint distance. Their idea is to partition the gene families into disjoint subsets such that two gene families in different subsets are ‘independent’, then to find the pair of exemplars of gene families in each independent subset at a time, and finally to merge all the exemplars together to obtain good exemplars of the given genomes. Tests with both simulated and real

datasets show that the combination of the divide-and-conquer and branch-and-bound approaches is much more efficient than the branch-and-bound approach.

Finally, an alternative way to look at the exemplar problem is to identify the gene copies that maximize the conserved or common intervals (see Section 2.1). Using this approach, Bourque et al. [26] showed that, under certain conditions, it is possible to improve on a method that would only utilize breakpoints.

2.5 Genome rearrangement with partially ordered genomes

Another restriction in the traditional study of genome rearrangement is the total order of genes in a genome inherent in the representation of the genome as a permutation. In practice, the total order of genes can only be determined after the sequenced genomes are completely annotated. Many genomes are currently only sequenced at a level that prevents a whole and accurate assembly and this problem will probably not be fixed in the near future because of the prohibitive sequencing costs. When the complete sequence of a genome is not available, one can rely on gene mapping data as input for rearrangement studies but even for these datasets, due to the relatively low resolution, several genes are often mapped to the same chromosomal position.

To deal with these ambiguities but also to work in general context of partial gene orders, it is possible to represent a chromosome, or genome, as directed acyclic graph (DAG) [168, 169]. The genome rearrangement problem with partially ordered genomes can be restated as the problem of inferring a sequence of mutational operations which transform a linearization of the DAG for one genome to a linearization of the DAG for the other genome that minimizes the number of operations required [168, 169, 141]. Obviously, such a general rearrangement problem is computationally challenging. Therefore, the development of efficient heuristic algorithms that could tackle some of the real datasets in the near future are highly desired.

3 Gene duplication and multigene families

In the human and other higher organisms, there are numerous gene families. The number of genes in each gene family ranges from several to hundreds. Some families contain genes with similar functions; others contain genes with very diverse functions. The large copy number of members in some gene families such as histone families is due to need for large amounts of gene product.

Gene duplication has been proposed as a major mechanism for generating multigene families, because duplicated genes provide raw genetic materials for the emergence of new functions through point mutation, natural selection and random drift [125]. Such gene duplication processes include polyploidization, tandem duplication, and retrotransposition. During polyploidization, whole genomes are duplicated. Tandem duplication is responsible for positional clustered gene families. It is probably caused by unequal crossing over during meiosis and mitosis in a germ cell lineage [144]. Although repetitive sequences derived

from reverse transcription are numerous in the human genome, there are not many retrogenes.

Since 1970s, phylogenetic analysis has been used for understanding relationships of gene family members, identifying gene duplication events and for orthologous gene assignment.

3.1 Gene trees and species trees

Bifurcating trees (called *phylogenies* or *phylogenetic trees*) have been used as models to represent the evolution of species, in which evolutionary lineages splits and evolve independently for each other, since Charles Darwin first pointed out that the simplest pattern that might lie in the heart of evolutionary history can be represented by a tree [41]. Indeed, Darwin called the evolution of species the Tree of Life.

For a set I of N taxa, their evolutionary history is represented by a rooted full binary tree T where there are N leaves each uniquely labeled by a taxon in I and $N - 1$ unlabeled internal nodes. Here the term “full” means that each internal node has exactly two children. Such a tree is called a *species tree*. In a species tree, we also consider an internal node as a subset (called a *cluster*) which includes as its members its subordinate species represented by the leaves below it. Thus, the evolutionary relation “ m is a descendant of n ” is expressed using set-theoretic notation as “ $m \subset n$.”

The model for gene evolution is a rooted full binary tree with leaves labeled by gene copies. Usually, a gene tree is constructed from a collection of genes each having several copies appearing in the studied species. For example, the gene family of hemoglobin genes in vertebrates contains α -hemoglobin and β -hemoglobin. A gene tree based on these two genes in human, chimpanzee and horse is shown in Figure 5. We use the species to label the genes appearing in it. Thus, the labels in a gene tree may not be unique since there are usually multiple genes under consideration in each species. Therefore, each internal node g in a gene tree corresponds to a multiset $\{x_1^{i_1}, x_2^{i_2}, \dots, x_m^{i_m}\}$, where i_j is the number of its subordinate leaves labeled with x_j . The *cluster* of g is simply the set

$$S_g = \{x_1, x_2, \dots, x_m\}.$$

Finally, we use $L(T)$ to denote the set of leaf labels in a species or gene tree T .

3.2 Gene duplications and losses

Detecting gene duplication and loss events is based on a node mapping from a gene tree to a species tree. Such a mapping was first considered by Goodman *et al.* [59] and later was popularized by Page in a series of papers [128–130, 132]. Given a gene tree G and a species tree S such that $L(G) \subseteq L(S)$. For any node $g \in G$, we define $M(g)$ to be the least common ancestor (lca) of g in S , i.e. the smallest node $s \in S$ such that $S_g \subseteq s$. Here we used term “smallest” to mean

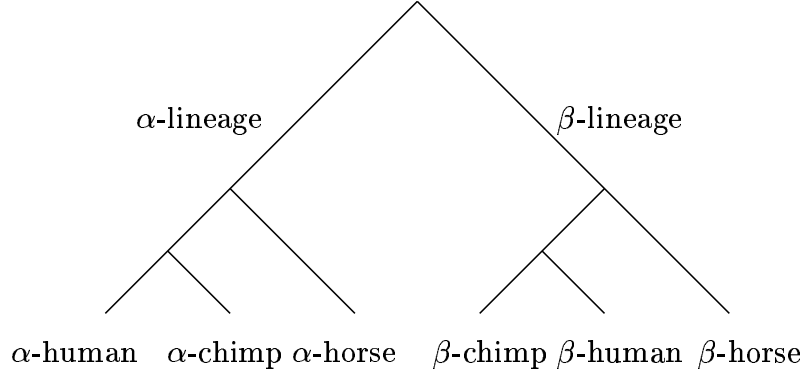


Fig. 5. A gene tree based on α -hemoglobin and β -hemoglobin.

“farthest from the root.” We call M the *LCA mapping* from G to S . Obviously, if $g' \subset g$, then $M(g') \subseteq M(g)$, and any leaf is mapped onto a leaf with the same label. For an internal node g , we use $c(g)$ (sometimes $a(g)$ and $b(g)$) to denote a child of g and $G(g)$ the subtree rooted at g .

If $M(c(g)) = M(g)$ for some child $c(g)$ of g , then we say a *duplication* happens at g . The total number $t_{dup}(G, S)$ of duplications happening in G under the LCA mapping M is proposed as a measure for the similarity between G and S [59, 128]. We call such a measure the *duplication cost*.

Let $a(g)$ and $b(g)$ denote the children of g . For $c(g) = a(g), b(g)$, if $M(c(g)) \neq M(g)$, we let P be the path from $M(g)$ to $M(c(g))$. We say that the gene gets lost on each lineage between a species X on the path P to its child $c(X)$ that is not on P [63]. Therefore, the *number of gene losses* l_g associated to g is

$$l_g = \begin{cases} 0 & \text{if } M(g) = M(a(g)) = M(b(g)); \\ d(a(g), g) + 1 & \text{if } M(a(g)) \subset M(g) \ \& \ M(g) = M(b(g)); \\ d(a(g), g) + d(b(g), g) & \text{if } M(a(g)) \subset M(g) \ \& \ M(b(g)) \subset M(g). \end{cases}$$

The *mutation cost* is defined as the sum of t_{dup} and the total number of losses $loss(G, S) = \sum_{g \in G} l_g$. This measure turns out to have a nice biological interpretation [109, 166, 51].

Since the LCA mapping from a gene tree to a species tree can be computed in linear time [166, 35, 170], the gene duplication and loss events can be identified effectively.

3.3 Reconciled tree

The *reconciled tree* concept gives another way to visualize and compare the relationship between gene and species trees [59]. Such a tree is constructed from a gene tree and a species tree and has two important properties. The first property is that the observed gene tree is a ‘subtree’ of the reconciled tree. The second

property is that the clusters of the reconciled tree are all clusters of the species tree. Formally, the reconciled tree is defined as follows.

Let T' and T'' be two rooted trees, we use $T' \Delta T''$ to denote the rooted tree T obtained by adding a node r as the root and connecting r to $r(T')$ and $r(T'')$ so that T' and T'' are two subtrees rooted at the children of r . Further, let t be an internal node in T' , then, $T'|_{t \rightarrow T''}$ denotes the tree formed by replacing the subtree rooted at t with T'' . Similarly, $T'|_{t \rightarrow T_1, t' \rightarrow T_2}$ can be defined for disjoint nodes t and t' .

For a gene tree G rooted at g and a species tree S rooted at s such that $L(G) \subseteq L(S)$, let M be the LCA mapping from G to S and let $s' = M(a(g))$ and $s'' = M(b(g))$. The *reconciled tree* $R = R(G, S)$ of G with respect to S is defined as:

$$R = \begin{cases} R(G(a(g)), S) \Delta R(G(b(g)), S) & \text{if } s' = s'' = s, \\ S|_{s' \rightarrow R(G(a(g)), S(s')), S(s'')} \Delta R(G(b(g)), S) & \text{if } s' \subseteq a(s), s'' = s, \\ S|_{s' \rightarrow R(G(a(g)), S(s')), s'' \rightarrow R(G(b(g)), S(s''))} & \text{if } s' \subseteq a(s), s'' \subseteq b(s), \\ S|_{a(s) \rightarrow R(G, S(a(s)))} & \text{if } M(g) \subseteq a(s). \end{cases} \quad (1)$$

Such a concept is illustrated in Figure 6. An efficient algorithm was presented in [128] for computing a reconciled tree given a gene and species tree. It is easy to see that the reconciled tree $R(G, S)$ satisfies the following three properties, of which the first two are mentioned above:

1. It contains G as a subtree, i.e. there is a subset L of leaves such that $R(G, S)|_L$ is isomorphic to G ;
2. All clusters are in S , where a cluster is defined as a subset of species below an internal node in S (see Subsection 3.1);
3. For any two children $a(g)$ and $b(g)$ of a node $g \in R(G, S)$, $a(g) \cap b(g) = \phi$ or $a(g) = b(g) = g$.

Actually, Page also defined the reconciled tree $R(G, S)$ as the smallest tree satisfying the above properties. However, these two definitions are not obviously equivalent. A rigorous proof of this equivalence is given in [21].

Obviously, duplication events are one-to-one correspondent to the internal nodes with two identical children in the reconciled tree. Moreover, in [61], Gorecki and Tiuryn proved an earlier conjecture that the number of gene losses is also equal to the number of the maximal subtrees that do not contains any nodes in the image of the gene tree in the reconciled tree.

3.4 From gene trees to species trees

Over the years, biomolecular sequence information has been applied effectively toward to reconstructing the species tree – the evolution history of species. Under the gene duplication model, the problem is formulated:

Definition 9. Species Tree Problem: *Give a set of gene trees G_i ($1 \leq i \leq n$), find a species tree T that has the minimum duplication cost $\sum_{1 \leq i \leq n} t_{dup}(G_i, T)$ or mutation cost $\sum_{1 \leq i \leq n} (t_{dup}(G_i, T) + t_{loss}(G_i, T))$.*

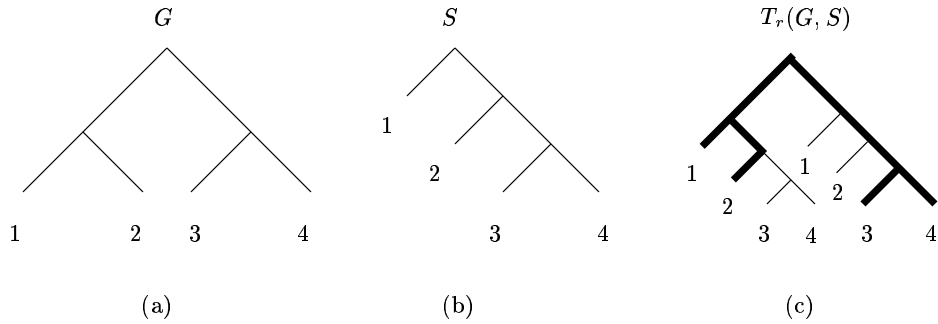


Fig. 6. (a) A gene tree G ; (b) a species tree S ; (c) the reconciled tree $T_r(G, S)$ of G with respect to S .

For either cost, this problem was proved to be NP-hard by Ma, Li and Zhang in [103] and the parametric complexity of this problem was studied by Fellows *et al.* in [149, 52]. Moreover, various heuristic algorithms have been proposed by Page [130], Arvestad *et al.* [3] and Durand, Halldorsson and Vernot [46].

3.5 Tandem gene duplication

Tandem duplication tree model In the study of tandem duplication history of human hemoglobin, Fitch first observed that tandem duplication histories are much more constraint than speciation histories and proposed to model them assuming that unequal crossover is the biological mechanism from which they originate [54], and the corresponding trees are now called *tandem duplication trees*.

Assume n sequences $\{1, 2, \dots, n\}$ were formed from a locus through a series of tandem duplications, where each duplication replaced a stretch of DNA sequences containing several repeats with two identical and adjacent copies of itself. If the stretch contains k repeats, the duplication is called a k -duplication.

A rooted *duplication tree* \mathcal{M} for tandemly repeated segments $\{1, 2, \dots, n\}$ is a rooted binary tree that contains blocks as shown in Figure 7. A node in \mathcal{M} represents a repeat. Obviously, the root represents the original copy at the locus and leaves the given segments.

A *block* in \mathcal{M} represents a duplication event. Each non-leaf node appears in a unique block; no node is an ancestor of another in a block. If the block corresponds to a k -duplication, it contains k nodes, say, u_1, u_2, \dots, u_k from left to right. Assume $lc(u_i)$ and $rc(u_i)$ are the left and right children of u_i , $1 \leq i \leq k$. Then, in the model \mathcal{M} ,

$$lc(u_1), lc(u_2), \dots, lc(u_k), rc(u_1), rc(u_2), \dots, rc(u_k)$$

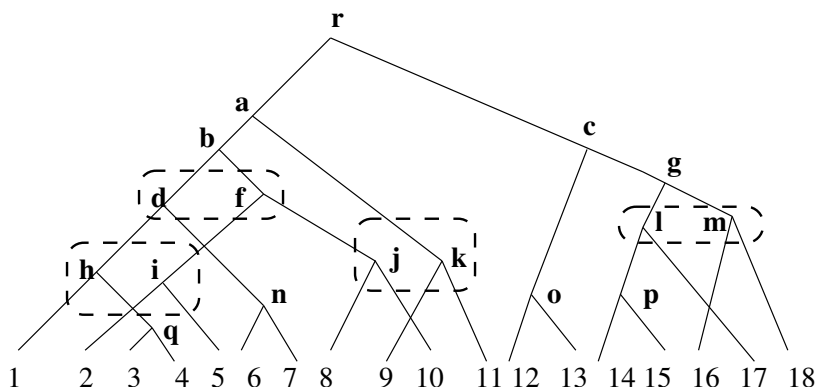


Fig. 7. A rooted duplication tree \mathcal{M} . Multi-duplication blocks are $[d, f]$, $[h, i]$, $[j, k]$, and $[l, m]$.

are placed from left to right. Hence, for any i and j , $1 \leq i < j \leq k$, the directed edges $(u_i, rc(u_i))$ and $(u_j, lc(u_j))$ cross each other. But no other edges cross in the model. For simplicity, we will only draw blocks corresponding to multi-duplication events that contain more than one internal nodes.

The leaves representing given segments are placed from left to right in the same order as the segments appear on the chromosome. Here we assume such an order is the increasing order.

Combinatorics of tandem duplication models Duplication trees containing only 1-duplications are called *ordered phylogenies*. They form a proper subclass of the duplication models. Since an ordered phylogeny with n leaves corresponds uniquely to a triangulation of a regular $(n + 1)$ -polygon, the number of ordered phylogenies with n leaves is just $\binom{2(n-1)}{n-1}/n$, the n -th Catalan number [167, 48].

Like phylogenies, both rooted and unrooted duplication trees are studied. An unrooted phylogeny is a duplication tree if it can be rooted (on some edge) into a duplication tree.

Theorem 5. ([58]) *The number of rooted duplication trees for n segments is twice the number of unrooted duplication trees for n segments.*

A simple non-counting proof of this theorem is given by Yang and Zhang in [163]. Moreover, using a recurrence relation in [58], they also obtained the following recurrence relation for computing the number of duplication trees.

Theorem 6. ([163]) *Let r_n denote the number of duplication trees for n segments. For any $n \geq 2$,*

$$r_n = \begin{cases} 1 & \text{if } n = 2, \\ \sum_{k=1}^{\lfloor (n+1)/3 \rfloor} (-1)^{k+1} \binom{n+1-2k}{k} r_{n-k} & \text{if } n \geq 3. \end{cases}$$

Reconstruction algorithms Tandem repeats are everywhere in the genome of higher organisms. A good method for reconstructing the parsimonious duplication tree is extremely useful for identifying orthologous genes in genome annotation. However, there has not been a good solution for it up to now.

Since the duplication model space is huge, the problem of constructing a parsimonious duplication tree given a set of duplicated sequences is believed to be NP-hard. Indeed, Jaitly *et al* proved that finding the parsimonious ordered tree is NP-hard [84]. Therefore, one approach to the problem is to search the phylogenetic trees that are duplication trees after the parsimony score is computed. Indeed, whether a phylogeny is a duplication tree or not can be determined efficiently. Tang, Waterman and Yooseph first gave a quadratic time algorithm for the problem [152]. Later, Zhang et al. and Elemento, Gascuel and Lefranc presented two different linear-time algorithms for the problem [167, 49]. Other heuristic reconstructing methods can be found in [152, 47].

Finally, different algorithms for reconstructing parsimonious ordered trees were developed in [9, 84, 152].

4 Phylogenetic networks

4.1 Tree of life or net of life?

As we mentioned in the last section, phylogenetic trees have a long history as models to represent the evolution of species. However, in the last decade, the large-scale availability of genomic sequences indicates that horizontal gene transfer (HGT), gene conversion, and recombination events have often occurred in genome evolution.

Horizontal gene transfer events occur when genetical material transfers across from a species to another distantly related species. They are common in the prokaryotes, especially bacterial genomes [95, 44, 119]. Additional evidence suggests that it might also occur in eukaryotes [79]. In many cases, horizontal gene transfers are very interesting in their own. Indeed, many reflect the most innovative adaptations in all of biology such as bacterial photosynthesis and nitrogen fixation. Horizontal transfers are not restricted to single genes. Genes, operons and a large segment of genomes are commonly exchanged among prokaryotes.

Recombination is another important mutational process that is common to most forms of life. A species is defined as a potentially interbreeding group of organisms that are capable of producing fertile offspring. Within a species, gene phylogenies are often inconsistent due to high rate gene flow and meiotic recombination. Meiotic recombination takes two equal length sequences and produces a third one of the same length by concatenating a prefix of one sequence and a suffix of another one. Other forms of recombination such as transformation, conjugation and transduction allow the sharing of genetic material between species as indicated by the recent completion of the sequences of different bacterial genomes. Efforts to identify patterns of recombination and the location of recombination are central to modern genetics.

In a nutshell, genomes have evolved not only vertically, but also horizontally. As a result, directed networks (i.e. trees with reticulation branches) are probably more appropriate mathematical model for the study of genome evolution. Here, we shall present the recent study of the algorithmic aspects of reconstructing phylogenetic networks. Readers are referred to [115, 134] for more information on phylogenetic networks.

4.2 Horizontal gene transfer detection and models

G+C content-based detection The G+C content of a genome is determined by mutation and selection pressures. Hence, the sequences from a genome share a common feature of compositional bases, codons and oligonucleotides [62, 86]. This makes it possible to identify horizontally transferred genes as those whose G+C content is atypical for a particular genome [95, 79, 171]. For example, *M. thermoautotrophicum* contains several regions that have about 10% lower G+C content than that of the whole genome on average [145]. ORF's in these regions exhibit a codon usage pattern atypical of *M. thermoautotrophicum*, suggesting that they code some genes acquired by HGT. HGT genes are usually G+C poor. However, this method should be used with caution since sequences can quickly adjust to the new genome pattern and a gene with different G+C content does not necessarily originates in distant organisms [42].

Phylogeny-based detection model Comparison of a gene tree and a species tree provides a reliable method for identifying horizontally transferred genes [1, 70, 71, 60, 104, 110]. It is based on the following simple idea: If A and B are siblings in the gene tree, then, either the parent gene AB must present in the last common ancestor of A and B in the species tree or a horizontal gene transfer has occurred from the lineage A to the lineage B or vice versa. Horizontal gene transfers are modeled as a species graph formed from a species tree by adding additional horizontal edges [70, 60]. The horizontal edges represent the hypothetical horizontal gene transfers.

Let $S = (V, E)$ be a rooted tree in which each internal node has at most two children. A time stamp for S is a function t from V to non-negative integers with the following property: for any $v \in V$, $t(p(v)) < t(v)$, where $p(v)$ is the parent of v in S .

A relation $H \subset V \times V$ is *horizontal* for S with respect to a time stamp t if the following conditions are true:

H1: $(v, v) \notin H$;

H2: for each $(v, w) \in H$, both v and w have only one child;

H3: for any different $(v, w), (v', w') \in H$, $\{v, w\} \cap \{v', w'\} = \emptyset$;

H4: for any $(v, w) \in H$, $t(v) = t(w)$;

H5: for any different $(v, w), (v', w') \in H$, $t(v) \neq t(v')$.

Each element of H is a *horizontal transfer*. Intuitively, (H3) prevents more than one horizontal transfer events from occurring on the same lineage and (H4) indicates that the ends of a horizontal transfer should exist at the same time.

A *species graph* $\mathcal{G} = (S, t_S, H)$ consists of a rooted tree S , a time stamp t_S and an horizontal relation H on S with respect to t_S . We use $S(H)$ to denote the directed graph obtained by adding elements in H as arcs on S . Then, the following condition is true:

For each directed path v_1, v_2, \dots, v_k in $S(H)$, $t_S(v_1), t_S(v_2), \dots, t_S(v_k)$ is a non-decreasing integer sequences.

Obviously, the above property implies that $S(H)$ is a directed acyclic graph.

Let G be a gene tree and T a species tree. A rooted binary tree S is an extension of T if T can be obtained from S by contracting all the degree-2 nodes. Each species graph (S, t_S, H) is a *model* of horizontal transfers of the gene occurring in the gene tree G on the species tree T if S is an extension of T and G can be embedded into $S(H)$ as illustrated in Figure 8. Notice that the time stamp and condition (H5) are used to prevent inconsistent transfer events as indicated in Figure 9.

The problem of inferring horizontal gene transfers is formulated as follows:

Definition 10. HGT Inference Problem: *given a species tree T and a gene tree G , find a horizontal gene transfer model (S, t_S, H) with the smallest gene transfer set H over all the models.*

Combining the horizontal gene transfer model and the concept of reconciliation tree described in the last section, one obtains a mathematical model for simultaneously detecting gene duplication, loss and horizontal transfer events [60, 71]. Like duplication inference problems, the HGT inferring problem and its generalization to simultaneously detecting gene duplication, loss and horizontal transfers are obviously NP-hard [40]. Hence, an important and practical problem is to develop efficient algorithm for these two problems.

4.3 The recombination model

A recombination network \mathcal{N} over a set S of 0-1 sequences of length L has four basic components:

- i. The topology structure of \mathcal{N} is a directed acyclic graph D . It contains a unique node (called the *root*) with no incoming edges, a set of internal nodes that have both incoming and outgoing edges, and a set of nodes (called the *leaves*) with no outgoing edges. Each internal node has one or two incoming

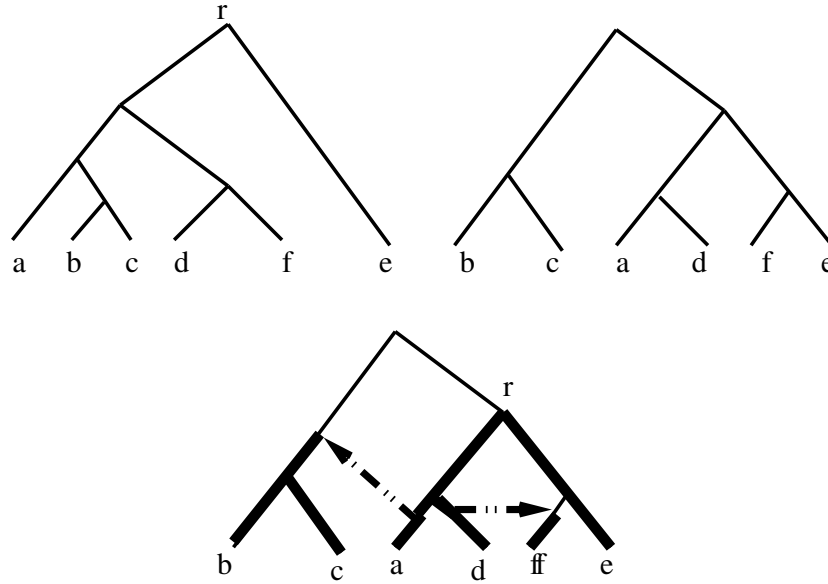


Fig. 8. A species graph. On the top, the left tree is a gene tree, the right one is a species tree. The species graph is shown at the bottom, in which the gene tree is embedded.

edges. A node with one incoming edge is called *tree node*; a node with two incoming edges is called *recombination node*. An edge is called a *tree edge* if it enters a tree node, and called *recombination edge* if it enters a recombination node.

- ii. There is a mapping w from the integer set $[1, L]$ to the set of tree edges of D . It assigns a site i in the sequences to a unique tree edge e . We write $i = w^{-1}(e)$. It is possible that there are more than L tree edges in D and hence some tree edges might not receive a site assignment.
- iii. There are exactly two recombination edges entering each recombination node. These two edges are labeled with p and s respectively. In addition, a site is associated with a recombination node.
- iv. There is also a mapping that labels each node of D with a 0-1 sequence. The labels satisfy the following conditions:
 - (a) For a tree-node v , let e be the edge entering into v . Then, the label of v is only different from its parent's label in site $w(e)$. This models a mutation in site i occurring on edge e .
 - (b) For a recombination node v , let e and e' be the edges coming into v from v' and v'' , labeling with p and s respectively, and let the integer

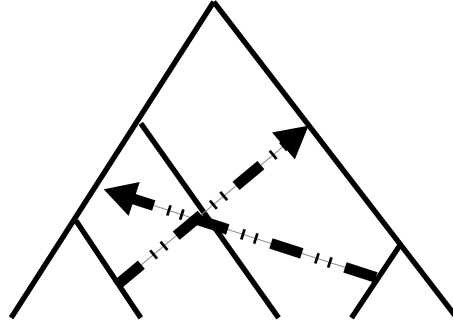


Fig. 9. Inconsistent transfer events that could not occur in genome evolution.

assigned to v in (iii) is i . The label of v is identical to v' 's label in the first $i - 1$ sites and to v'' 's label in the last $L - i + 1$ sites. This models a single-cross recombination occurring at site i .

- (c) All the leaves are uniquely labeled with the given sequences in S .

One phylogenetic network is shown in Figure 10. In this network, the ancestral sequence is 0000000, the leaves are labeled with sequences a, b, c, d, e , and there are three recombination nodes. The model presented here can be generalized in different ways. This simple version is the one that has been extensively studied currently. It is also the topology part of the stochastic process model called an ancestral recombination graph (ARG) in the population genetics study (see [120] for example).

In recombination model, the central problem is to infer a rooted or unrooted phylogenetic network with minimum number of recombination nodes on a set of given sequences. Although such a problem is NP-hard [161], it is polynomial-time solvable for two classes of phylogenetic networks: perfect phylogeny and galled phylogenetic networks [64–66, 68, 161].

Perfect phylogeny A *perfect phylogeny* is a recombination network without recombination nodes. This is a simple combinatorial characterization for the 0-1 sequences that can be derived from a perfect phylogeny. Given a set S of 0-1 sequences, two sites in the sequences are said to be *incompatible* in S if and only if S contains four rows where sites i and j contains all four possible ordered pairs: 00, 01, 10, 11. A site is *compatible* if it does not form any incompatible pair together with any other site. For a sequence s , two sites i and j in S are said to *conflict* relative to s if i and j are incompatible in $S \cup \{s\}$. A classical

edge connecting two nodes if the corresponding sites are incompatible. Similarly, given a sequence s , we can define the “conflict graph” $G_s(S)$ for S (relative to s).

A *connected component* of a graph is a maximal subgraph in which for any two nodes in it there is at least one path between them in the graph. If all the sites are compatible, then $G(S)$ has only trivial components that have only one node and no edges.

In a series of papers [68,67,66], Gusfield and his collaborators studied the structural properties of a connected component that corresponds to a gall in a phylogenetic network. By giving an efficient algorithm to determine how the sites in a connected component with the desired structural properties are arranged on a gall, they obtained the following theorem.

Theorem 8. *There is a polynomial-time algorithm that determines whether a set of 0-1 sequences can be derived from a galled phylogenetic network or not and constructs such a galled network.*

Full-decomposition optimality conjecture In a non-galled phylogenetic network, recombination cycles are not necessarily edge-disjoint. A maximal set of cycles C_1, C_2, \dots, C_k form a *blob* if they can be arranged in such a way $C_{i_1}, C_{i_2}, \dots, C_{i_s}$ that the successive cycles share edges. In [65], Gusfield and Bansal proved the following theorem.

Theorem 9. *Let S be a set of 0-1 sequences of equal length. Then*

- (i) *there is an unrooted recombination network \mathcal{N} deriving S in which each blob contains exactly all sites in a single non-trivial connected component in $G(M)$, and every compatible site is assigned to a tree edge;*
- (ii) *there is a recombination network \mathcal{N} that derives S , with ancestral sequence s , in which each blob contains exactly all sites in a non-trivial connected component of $G_s(S)$ and any non-conflicting site is assigned to a tree edge.*

Obviously, there are other recombination networks that derive S , where a blob may contain sites from different connected components in $G(S)$. Therefore, the following conjecture arises. It is considered as one of the most important open problems in the study of phylogenetic networks.

Conjecture 1. Full-decomposition Optimality Conjecture [65] For any 0-1 sequence set S , there is always a recombination network that satisfies the condition in the above theorem and has the minimum number of recombinations over all possible phylogenetic networks for S .

In [66], Gusfield showed that for the sequence set that can be derived from a galled network, there is an efficient algorithm for producing a galled network with the minimum number of recombinations, over all possible phylogenetic networks for the sequences. Recent progress toward this conjecture can be found in [69].

4.4 The hybrid model

In the hybrid model, a phylogenetic network for a set of taxa (species, DNA, protein sequences, or other objects) has all the properties of a recombination network except for edge site-labels. It is a directed acyclic graph in which there is a unique node (the *root*) without coming edges, a set of nodes (*leaves*) without outgoing leaves, and some other nodes with one outgoing edges and one or two incoming edges. The leaves are one-to-one labeled with given taxa. The nodes with one incoming edge are *tree nodes*; the nodes with two incoming edges are *reticulation nodes*. The edges entering tree nodes are tree edges; the edges entering reticulation nodes are network edges. As in the recombination model, we can define the reticulation cycle similarly. A reticulation network without overlapping reticulation cycles is galled.

Different variants of the hybrid model have been proposed. For example, a phylogenetic network is not necessarily binary. A phylogenetic network may have ‘time-weighted’ edges that satisfy time constraints so that two parents of a reticulation nodes should coexist in the same time [111].

Network reconstruction from gene trees In genomic evolution, one usually obtains a set of gene trees that have evolved from a common ancestor in a series of reticulate events, and would like to reconstruct the underlying phylogenetic network from these gene trees. By removing exactly one of the two edges entering every reticulation node in a network, we obtain a phylogenetic tree. Such a tree is said to be induced by the network. Formally, the computing problem arise from above procedure is

Definition 11. Parsimonious Reticulate Network from Gene Trees Problem: *given a set S of gene trees over a taxon set X , construct a phylogenetic network that induces all the trees in S and has the minimal number of reticulation nodes, over all the phylogenetic networks.*

In 1997, Maddison first proposed this problem and studied how to construct a phylogenetic network with one reticulation node for two gene trees [105]. Since this problem is NP-hard in general [161], different algorithms have been proposed recently [37, 82, 83, 111, 118, 147]. In particular, the problem is also polynomial-time solvable for galled phylogenetic networks. More specifically, Nakhleh et al. proved the following theorem

Theorem 10. ([118]) *Given two binary trees T_1 and T_2 , it is polynomial-time computable a galled phylogenetic network, if existence, that induces T_1 and T_2 and has the minimum number of reticulation nodes, over all the galled phylogenetic networks.*

The above theorem is recently generalized to multiple (not necessarily binary) trees by Huynh *et al.* Let T and t be two arbitrary trees. We say T refines t if t can be obtained by a series of edge contractions. A phylogenetic network \mathcal{N} refines t if \mathcal{N} induces a binary tree that refines t .

Theorem 11. ([83]) *Given a set of trees, it is polynomial-time computable a galled phylogenetic network, if existence, that refines the given trees and has the minimum number of reticulation nodes, over all the galled phylogenetic networks.*

When a set of phylogenetic trees cannot be combined into a galled phylogenetic network, one may be interested in knowing to what extent, these trees admit a solution. Hence, the following problem is interesting:

Definition 12. *Phylogenetic Network Compatibility Problem: given a class of phylogenetic network, and a set of trees over a taxon set X , find a largest subset X' of X such that the set of trees restricted on X' have a refined network in the given class.*

This problem is believed to be NP-hard. For the class of galled phylogenetic networks, the following result was obtained.

Theorem 12. ([83]) *Given a set of k trees each with maximum degree d over a taxon set X , it is $O(2^{3kd}n^{2k})$ -time computable a largest subset X' of X such that the restriction of the given trees on X' admit a refining galled phylogenetic network.*

Network reconstruction from sequences All phylogeny reconstruction methods can and will probably generalize to phylogenetic network reconstruction in future. Here, we shall examine how the parsimony method is generalized in detail.

Let S be a set of equal-length DNA or protein sequences. For two sequences $x, y \in S$, we use the Hamming distance $H(x, y)$ between them to measure their dissimilarity. It is defined as the number of mismatch positions between x and y . Let T be a rooted phylogeny over S . Then, each internal node is implicitly labeled with a sequence s_v of the same length as those in S ; each leaf is labeled uniquely with a sequence in S . The *parsimony score* $s(T, S)$ of T is defined as $\sum_{(u,v) \in E(T)} d(u, v)$, where $E(T)$ denotes the set of tree edges in T . The parsimony phylogeny for sequence set S is a rooted phylogeny that has the minimum parsimony score overall the phylogenies.

The *parsimony problem* is, given a set S of equal-length sequence set, to compute a parsimony phylogeny for S . It is known that this problem is NP-hard (see [57] for example). Therefore, one practical approach is through exhaustive search over the phylogeny space after the parsimony score of a phylogeny is computed, which is linear-time solvable (see [53] for example). In literature, computing the parsimony score of a phylogeny for a set of equal-length sequences is called the *small parsimony problem*.

Three parsimony problems arise from reconstructing a phylogenetic network from biomolecular sequences. Since each phylogenetic network \mathcal{N} induces a set of phylogenies $P(\mathcal{N})$, we have the following problem:

Definition 13. *Small Parsimony Phylogenetic Network Problem: given a phylogenetic network \mathcal{N} for a set S of equal-length sequences, find a labeling of*

the internal nodes of \mathcal{N} that has the minimum parsimony score $s(\mathcal{N}, S) = \min_{T \in \mathcal{P}(\mathcal{N})} s(T, S)$.

To take non-point-mutation events such as recombination into account, we assume the given sequences are partitioned into different blocks b_i ($1 \leq i \leq n$) such that only point-mutation events occurred in each block and non-point-mutation events combined sequences from different blocks, where each block is specified by start and end positions. If we use $S|_{b_i}$ to denote the resulting sequence set in the block b_i , then, the parsimony problem for phylogenetic networks is formulated as

Definition 14. Parsimony Phylogenetic Network Problem: *given a set S of equal-length sequences that are partitioned into block b_i ($1 \leq i \leq n$), find a phylogenetic network \mathcal{N} with the minimum score $s(\mathcal{N}, S, \{b_i\}) = \sum_{i=1}^n s(\mathcal{N}, S|_{b_i})$.*

In the study of genomic evolution, we are only given a set of genomic sequence from different species. Without knowing the true evolutionary history of these sequences, we do not have the true partition blocks on the sequences. Hence, the following problem is also interesting and practical:

Definition 15. Large Parsimony Phylogenetic Network Problem: *given a set S of equal-length sequences and an integer k , find a phylogenetic network \mathcal{N} and a block partition $\{b_i | 1 \leq i \leq k\}$ of S such that the parsimony score $s(\mathcal{N}, S, \{b_i\})$ is minimum.*

The parsimony phylogenetic network problem was proposed by J. Hein in early 1990s [75, 76]. With more and more genomic sequences available, researchers redraw attention to the problem recently. Nakhleh *et al.* presented a heuristic algorithm for the parsimony phylogenetic network problem in [117]. It is easy to see the small parsimony phylogenetic network problem is polynomial-time solvable for the given network has a constant number of reticulation nodes. But, it is NP-hard in general [121]. In addition, it is not clear whether the first and third problems are polynomial time solvable for galled phylogenetic networks or not.

Other heuristic parsimony methods include statistical parsimony [153], median networks [6, 7] and the netting method [55].

Distance-based reconstruction methods Different distance-based methods have also been proposed for reconstructing phylogenetic networks [5, 28, 43]. Split decomposition was proposed to decompose the distance matrix into weighted splits (a bipartition of the given taxon set) [5]. When these splits are compatible, they induce a phylogenetic tree, in which each split corresponds an edge. Otherwise, a network (called *split graph*) is used to realize them. Split decomposition is implemented in the package SplitsTree [81]. Strictly speaking, split graphs are not phylogenetic networks. They are just used to visualize the possible recombination events.

Neighbor-Net is a kind of combination of the NJ method and the split decomposition method [28]. It first constructs a collection of weighted splits using a generalization of the NJ method, then realizes these splits using a splits graph.

The Pyramid Clustering works agglomeratively like the UPGMA method for phylogeny reconstruction [43]. The UPGMA method generates a binary tree, whose internal nodes correspond the nested, non-overlapping clusters of taxa. In contrast to this, the Pyramid Clustering constructs overlapping clusters, forming a network.

Combinatorial aspect of phylogenetic networks Phylogenetic network reconstruction also raises some interesting combinatorial problems. One of such problems is, given a set of trees, to estimate the number of reticulation nodes in any phylogenetic network that contains the given trees as induced trees. Another problem is to study the combinatorial properties of special classes of phylogenetic networks such as unicyclic and galled networks. The readers are referred to [15, 16, 82] for recent results.

5 Gene clusters

5.1 Positional gene clusters in eukaryotic genomes

Study of gene order within a genome is one of the key areas of genetics and genomics [126, 80]. It is important in terms of understanding how genomes have evolved and how they function. For example, recent analysis indicates that genomic regions with the most actively expressed genes are those of highest gene density [157]. It also has important medical implications. An intact gene in a novel location could lead to a pathological phenotype [90].

It has long been known that genes are organized into operons in prokaryotic genomes such as bacteria genomes. However, gene order seems not random neither in eukaryotic genomes [80]. Recent analyses suggest that tissue(or function)-specific genes often cluster together in eukaryotic genomes (Table 3). As a result, information about co-localised genes can be used for functional inferences of unknown genes through the ‘guilt by association’ principle [2].

5.2 Statistical tests

In most of all the literatures, testing for non-random clustering of specific genes is done by simulation. The simulation process starts with formulating a test function. Then, generate a random genome and calculate the test function for many times. The whole process generates the null distribution of the test function. The real value of the test function is then compared with the null distribution.

However, the results of different simulation studies are often difficult to compare. This motivates researchers to seek alternative analytic test methods.

| Species | Clusters observed |
|------------------------|---|
| <i>P. falciparum</i> | Clusters of co-expressed proteins [56] |
| <i>S. cerevisiae</i> | Clusters of cell-cycle-dependent genes [36] Pairs of co-expressed neighboring genes, independent of orientation [39, 92] |
| <i>A. thaliana</i> | Clusters of co-expressed genes [18, 160] |
| <i>D. melanogaster</i> | Clusters of adjacent co-expressed or function-specific genes [148] Clusters of tissue-specific genes [23] |
| <i>C. elegans</i> | Operons that contain about 15% genes [20] Clusters of muscle-specific genes [136] Clusters of co-expressed neighboring genes [96] |
| <i>M. musculus</i> | Clusters of tissue-specific genes [91, 135, 100] |
| <i>H. sapiens</i> | Clusters of tissue-specific genes [107, 22, 164] Housekeeping genes [97, 98] |

Table 3. Genome-wide analyses on gene clusters.

Neighborhood model In study of testes-specific gene clustering in the mouse genome, Li, Lee and Zhang used the neighborhood model [100]. Under this model, two testis-specific genes are in a cluster if and only if there is a series of the testis-specific genes locating between them such that the distance between any two successive testis-specific genes in the series is less than a specified threshold (D). To incorporate the variance of gene density in different regions on a chromosome, each chromosome is divided into disjoint regions of a fixed length (L). Consider a length L -region containing N genes in total. By Poisson approximation theory, the p-value of a cluster with n tissue-specific genes in that region is about $(1 - e^{-ND/L})^n$, the probability that a cluster has more than n genes in that region.

The neighborhood model was also studied in earlier works [45, 78]. A cluster in the neighborhood model is called a *max-gap* cluster [78]. Hoberman, Durand and Sankoff showed that the exact probability that all the m interesting genes form a max-gap cluster with distance threshold D in a genome with N genes is

$$P(N, m, D) = \frac{\max(0, N - w + 1) \cdot (D + 1)^{m-1} + d_o(m, D, \min(n, w - 1))}{\binom{n}{m}}$$

where $w = m + D(m - 1)$ and

$$d_o(m, D, \min(n, w - 1)) = \sum_{r=m}^{\min(n, w-1)} \sum_{i=0}^{\lfloor \frac{r-m}{D+1} \rfloor} (-1)^i \binom{m-1}{i} \binom{r-i(D+1)-1}{m-1}.$$

They also presented a dynamic programming algorithm for computing the probability of observing a cluster of h (out of m) interesting genes in a chromosome that contains N genes.

Adjacent gene clustering Order statistics can be a very powerful tool for removing the effect of non-uniform distribution of genes on statistical test although its power in clustering test has not been fully investigated. For instance, Li, Lee and Zhang considered the positional rank of a gene rather than its specific position by ordering all the genes according to their positions on a chromosome [100]. By treating the set of testis-specific genes and the set of other genes as two types of identical objects, a gene distribution on a chromosome is modeled as a binary string with 0 represents a tissue-specific gene. An *adjacent gene cluster* is a max-gap cluster with distance threshold $D = 0$.

Assume there are M genes in a chromosome and T of them are the testis-specific genes. Then, the probability that a random chromosome has a r -adjacent testis-specific gene cluster is

$$P_r = \binom{M - T + 1}{r} \binom{T - 1}{r - 1} / \binom{M}{T}.$$

Hence, the mean number of adjacent testis-specific clusters in a random chromosome is

$$\mu = \sum_{r=1}^T r P_r = (M - T + 1)T/M$$

and the standard deviation is

$$\sigma = \sqrt{(M - T + 1)(M - T)T(T - 1)/(M^2(M - 1))}.$$

Using these values, one can estimate the significance of a real testis-specific gene distribution on a chromosome.

6 Conclusion

We have briefly introduced the current research status in genome rearrangements, gene duplication, phylogenetic network and positional gene clustering. Classical results were presented in these four areas but many open problems were also highlighted. For instance, promising new developments for the analysis of genome rearrangements include: new measures of similarity that generalize simple gene order adjacencies, alternative evolutionary edit operations that facilitate the modeling of transpositions, efficient approaches for genome rearrangement with gene families and rearrangement of partially ordered genomes. Similarly, interesting future directions for the analysis of gene duplications include how to identify true orthologous genes across species using the duplication models presented in Section 3 and how to identify large-scale duplications occurring along a lineage in the evolutionary history. In the study of phylogenetic networks,

one important problem is to infer horizontal gene transfers among the bacterial genomes. Another major challenge is to develop a solid method for inferring phylogenetic network over a set of genomes given their genomic sequences. Finally, given that gene cluster analysis is a relatively new research topic, it is expected that more gene cluster testing methods based on order or other statistics will be developed. In its application end, the gene clustering analysis will probably become a routine task for every newly sequenced genome in the future.

This survey is far from being comprehensive as computational comparative genomics is a fast growing research topic. One topic which was not covered, for instance, involves studying the biological aspects around genomic structure. Other examples of important research areas missing from the current survey include: genomic sequence alignment problems and discovery of functional elements in genomic sequences. For genomic sequence alignment methods, we recommend a recent survey [8] of Batzoglou. For discovery of functional elements, we recommend the survey papers [30, 106, 27]. For further information on comparative genomics, the reader is also referred to another recent survey [108].

7 Acknowledgments

GB is supported by funds from the Agency for Science, Technology and Research (A*STAR) of Singapore. LXZ is supported by a grant from Singapore BMRC and a grant from ARF, National University of Singapore.

References

1. Addario-Berry L, Hallett M, Lagergren J. Towards identifying lateral gene transfer events. In *Proc. of Pac Symp Biocomput.* pp.279-90 2003.
2. Aravind L. Guilt by association: contextual information in genome analysis. *Genome Res.* 10(8):1074-7 2000.
3. Arvestad L, Berglund AC, Lagergren J, Sennblad B. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* 19 Suppl 1:i7-15 2003.
4. Bader DA, Moret BME, Yan M. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *WADS '01: Proceedings of the 7th Inter. Workshop on Alg. and Data Structures*,365–376 2001.
5. Bandelt HJ, Dress A. A canonical decomposition theory for metrics on a finite set. *Adv. Math.* 92: 47-105.
6. Bandelt HJ, Forster P, Sykes BC, Richards MB. Mitochondrial portraits of human populations using median networks. *Genetics* 141(2):743-53, 1995.
7. Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol.* 16(1):37-48, 1999.
8. Batzoglou S. The many faces of sequence alignment. *Brief Bioinform.* 6(1):6-22 2005.
9. Benson G, Dong L. Reconstructing the duplication history of a tandem repeat. *Proc Int Conf Intell Syst Mol Biol.*, pp. 44-53 1999.
10. Bergeron A, Stoye J. On the similarity of sets of permutations and its applications to genome comparison. In *Proceeding COCOON*, 68-79 2003.
11. Bergeron A, Mixtacki J, Stoye J. Reversal distance without hurdles and fortresses. In *proceedings CPM*, 388-399 2004.
12. Bergeron A, Blanchette M, Chateau A, Chauve C. Reconstructing ancestral gene orders using conserved intervals. In *Proceedings of WABI*, 14-25 2004.
13. Bafna V, Pevzner PA. Sorting by transpositions. *SIAM Journal on Discrete Mathematics*, 11(2):224240, 1998.
14. Bafna V., Pevzner P.A. Genome rearrangements and sorting by reversal. *SIAM Journal on Computing*, 25:272-289 1996.
15. Baroni M, Semple C, Steel M. A framework for representing reticulate evolution. *Annals of Combinatorics* 8: 391–408 2004.
16. Baroni M, Grunewald S, Moulton V, Semple C. Bounding the number of hybridisation events for a consistent evolutionary history. *J. Math. Biol.* 51: 171-182 2005.
17. Berman P, Hannenhalli S. Fast sorting by reversal. In *Proceedings Combinatorial Pattern Matching.*, LNCS 1075:168-185 1996.
18. Birnbaum K, Shasha DE, Wang JY, et al. A gene expression map of the Arabidopsis root. *Science* 302:1956-60 2003.
19. Blanchette M, Bourque G, Sankoff D. Breakpoint phylogenies. *Genome Informatics Workshop (GIW 1997)*, 25-34 1997.
20. Blumenthal T, Evans D, Link CD, et al. A global analysis of *Caenorhabditis elegans* operons. *Nature* 417:851-4 2002.
21. Bonizzoni P, Della Vedova D, Dondi R. Reconciling gene trees to a species tree. In *Proc. Italian Conference on Alg. and Complexity (CIAC2003)*, LNCS 2653, 120-131 Rome (Italy), 2003.
22. Bortoluzzi S, Rampoldi L, Simionati B, et al. A comprehensive, high-resolution genomic transcript map of human skeletal muscle. *Genome Res.* 8: 817-25 1998.

23. Boutanaev AM, Kalmykova AI, Shevelyov YY, et al. Large clusters of co-expressed genes in the *Drosophila* genome. *Nature* 420:666-9 2002.
24. Bourque G, Pevzner PA. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res*, 12:26-36 2002.
25. Bourque G, Zdobnov EM, Bork P, Pevzner PA, Tesler G. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res*. 15(1):98-110 2005.
26. Bourque G, Yacef Y, El-Mabrouk N Maximizing synteny blocks to identify ancestral homologs *In Proceedings of RECOMB Satellite meeting on Comparative Genomics*, LNBI 3678:21-34 2005.
27. Brent MR, Guigo R. Recent advances in gene structure prediction. *Curr Opin Struct Biol*. 14(3):264-72 2004.
28. Bryant D, Moulton, V. NeighborNet: an agglomerative algorithm for the construction of phylogenetic networks. *Molecular Biology and Evolution* 21(2):255-265.
29. Bryant D. The complexity of calculating exemplar distance. In *Comparative Genomics* (eds: Sankoff D and Nadeau JH), Kluwer Academic Publishers, 2000.
30. Bulyk ML. Computational prediction of transcription-factor binding site locations. *Genome Biol*. 5(1):201 2003.
31. Caprara A Sorting by reversals is difficult. *RECOMB '97: Proceedings of the first annual international conference on Computational molecular biology*, 75–83 1997.
32. Caprara A. The reversal median problem. *INFORMS J. Computing*, 15(1)93–113 2003.
33. Chen X, Zheng J, Fu Z, Nan P, Zhong Y, Lonardi S, Jiang T. Assignment of orthologous genes via genome rearrangement. *IEEE/ACM Trans. Comput. Biol. and Bioinform.* 2: 302-315 2005.
34. Chen ZX, Fu B, Zhu BH. The approximability of the exemplar breakpoint distance problem. *Manuscript*, 2005.
35. Chen K, Durand D, Farach-Colton M. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J Comput Biol*. 7(3-4):429-47 2000.
36. Cho RJ, Campbell MJ, Winzeler EA, et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell*. 2(1):65-73 1998.
37. Choy C, Jansson J, Sadakane K, et al. Computing the maximum agreement of phylogenetic networks *Theoretical Computer Science* 335 (1): 93-107 2005.
38. Christie DA. Sorting permutations by block-interchanges. *Information Processing Letters* 60(4):165-169 1996.
39. Cohen BA, Mitra RD, Hughes JD, Church GM. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet*. 26(2):183-6 2000.
40. DasGupta B, Ferrarini S, Gopalakrishnan U, Paryani NR, Inapproximability results for the lateral gene transfer problem, In *Proc. 9th Italian Confer. on Theoret. Comput. Sci. (ICTCS'05)*, LNCS vol. 3701, Springer-Verlag, pp. 182-195, 2005.
41. Darwin C. *On the origin of species by means of natural selection or the preservation of favored races in the struggle for life*. John Murray, London, 1859.
42. Daubin V, Lerat E, Perriere G. The source of laterally transferred genes in bacterial genomes. *Genome Biol*. 4(9):R57 2003.
43. Diday E. Une représentation visuelle des classes empiétantes: le pyramides. *RAIRO Autoat.-Prod. Inform. Ind.* 20: 475-526.
44. Doolittle WF. Phylogenetic classification and the universal tree. *Science* 284:2124-2149 1999.
45. Durand D, Sankoff D. Tests for gene clustering. *J. Comput. Biol*. 10: 453-482 2003.

46. Durand D, Halldorsson BV, Vernot B. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *Proc. of Recomb 2005*, pp. 250-264 2005.
47. Elemento O, Gascuel O. An efficient and accurate distance based algorithm to reconstruct tandem duplication trees. *Bioinformatics*. 2002;18 Suppl 2:S92-9.
48. Elemento O, Gascuel O. An exact and polynomial distance-based algorithm to reconstruct single copy tandem duplication trees. In *Proceedings of the 4th Annual Symp. on Combinatorial Pattern Matching (CPM R03)*, Mexico, Lecture Notes in Computer Science **2676**: 96-108, Springer, Berlin.
49. Elemento O, Gascuel O, Lefranc MP. Reconstructing the duplication history of tandemly repeated genes. *Mol Biol Evol*. 19(3):278-88 2002.
50. Elias I, Hartman T. A 1.375-approximation algorithm for sorting by transpositions. In *Proceedings WABI*, LNBI 3692:204215 2005.
51. Eulenstein O, Mirkin B, Vingron M. Duplication-based measures of difference between gene and species trees. *J Comput Biol*. 5(1):135-48 1998.
52. Fellows M, Hallett M, Stege U. On the multiple gene duplication problem, *Proc. the 9th Inter. Sympo. on Alg. and Comput (ISAAC'98)*, LNCS 1533, pp. 347-356 1998.
53. Toward defining the course of evolution: minimum change for a specified tree topology. *Syst. Zool*. 20:406-416 1971.
54. Fitch W. Phylogenies constrained by cross-over process as illustrated by human hemoglobins in a thirteen cycle, eleven amino-acid repeat in human apolipoprotein A-I. *Genetics* **86**, 623-644 1977.
55. Fitch W. Networks and viral evolution. *J. Mol. Evol*. 44: S65-75, 1997.
56. Florens L, Washburn MP, Raine JD, et al. A proteomic view of the Plasmodium falciparum life cycle. *Nature* 419:520-6 2002.
57. Foulds L, Graham R. The steiner problem in phylogeny is NP-complete. *Adv. Appl. Math*. 3: 43-49 1982.
58. Gascuel O, Hendy MD, Jean-Marie A, McLachlan R. The combinatorics of tandem duplication trees. *Systematic Biology* 52: 110-118 2003.
59. Goodman M, Czelusniak J, Moore GW, et al. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences, *Syst. Zool*. 28, 132-163 1979.
60. Gorecki P. Reconciliation problems for duplication, loss and horizontal gene transfer, In *Proc. of Recomb*, pp.316-325, 2004.
61. Gorecki P, Tiuryn T. On the structure of reconciliations. In *Proc. Recomb Comparative Genomics Workshop 2004*, LNCS vol 3388, pp42-51.
62. Grantham R, Gautier C, Gouy M, Mercier R, Pave A. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res*. 8(1):r49-r62 1980.
63. Guigó R, Muchnik I, Smith T. Reconstruction of ancient molecular phylogeny, *Mol. Phy. and Evol*. 6(1996), No. 2, 189-213.
64. Gusfield D. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge University Press, Cambridge, UK, 1997.
65. Gusfield D, Bansal V. A fundamental decomposition theory for phylogenetic networks and incompatible characters Lecture Notes in Computer Science 3500: 217-232 2005
66. Gusfield D. Optimal, efficient reconstruction of root-unknown phylogenetic networks with constrained and structured recombination Journal of Computer and System Sciences 70 (3): 381-398 MAY 2005
67. Gusfield D, Eddhu S, Langley C. Optimal, Efficient reconstruction of phylogenetic networks with constrained recombination. *J. Bioinform. Comput. Biol*. 2(1): 173-213, 2004.

68. Gusfield D, Eddhu S, Langley C. The fine structure of galls in phylogenetic networks *Inform Journal on Computing* 16 (4): 459-469 2004
69. Gusfield D. On the full-decomposition optimality conjecture for hylogenetic networks. Tech. Report CSE-2005, UC Davis, January, 2005.
70. Hallett M, Lagergren J. Efficient algorithms for lateral gene transfer. In *Proc. of Recomb'01*, pp. 149-156 2001.
71. Hallett M, Lagergren J, Tofigh A. Simultaneous identification of duplications and lateral transfer. In *Prof. of Recomb'04*, pp. 164-173 2004.
72. Hannenhalli S, Pevzner PA. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals. *Proceedings of the 27th Annual ACM-SIAM Symposium on the Theory of Computing* 178-189 1995.
73. Hannenhalli S, Pevzner PA. Transforming men into mice: polynomial algorithm for genomic distance problem. *Proceedings of the 36th IEEE Symposium on Foundations of Computer Science* 581-592 1995.
74. Hannenhalli S, Pevzner PA Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *Journal of ACM* 46:1-27 1999.
75. Hein J. Reconstructing the history of sequences subject to gene conversion and recombination. *Math. Biosci.* 98:185-200 1990.
76. Hein J. A heuristic method to reconstruct the history of sequences subject to recombination. *J.Mol.Evol.* 20: 402-411 1993.
77. Heber S, Stoye J. Finding all Common Intervals of k Permutations. *Proceedings of CPM 2001*, LNCS 2089:207-218 2001.
78. The statistical analysis of spatially clustered genes under the maximum gap criterion. *J. Comput. Biol.* 12: 1083-1102 2005.
79. Huang J, Mullapudi N, Sicheritz-Ponten T, Kissinger JC. A first glimpse into the pattern and scale of gene transfer in Apicomplexa. *Int J Parasitol.* 34(3):265-74 2004.
80. Hurst LD, Pal C, Lercher MJ. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet.* 5(4):299-310 2004.
81. Huson D. SplitsTree - a program for analyzing and visulizing evolutionary data. *Bioinformatics* 14: 68-73 1998.
82. Huson DH, Klopper T, Lockhart PJ, et al. Reconstruction of reticulate networks from gene trees Lecture Notes in Computer Science 3500: 233-249 2005
83. Huynh TND, Jansson J, Nguyen NB, et al. Constructing a smallest refining galled phylogenetic network Lecture Notes in Computer Science 3500: 265-280 2005
84. Jaitly D, Kearney P, Lin G, Ma B. Methods for reconstructing the history of tandem repeats and their application to the human genome. *Journal of Computer and System Sciences* 65(3): 494-507 2002.
85. Kaplan H, Shamir R, Tarjan RE. Faster and simpler algorithm for sorting signed permutations by reversals. *Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms.* 344-351 1997.
86. Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* 11(7):283-90 1995.
87. Kececioğlu JD, Sankoff D. Efficient Bounds for Oriented Chromosome Inversion Distance. *CPM '94: Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching* Springer-Verlag, 307-325 1994.
88. Kececioğlu J, Ravi R. Of mice and men: algorithms for evolutionary distance between genomes with translocations. In *Proceedings of Sixth ACM-SIAM Symposium on Discrete Algorithms*, 604-613 1995.

89. Kececioğlu J, Sankoff D. Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica*, 13(1- 2):180-210 1995.
90. Kleinjan DJ, van Heyningen V. Position effect in human genetic disease. *Hum Mol Genet.* 7(10):1611-8 1998.
91. Ko MS, Threat TA, Wang X, et al. Genome-wide mapping of unselected transcripts from extraembryonic tissue of 7.5-day mouse embryos reveals enrichment in the t-complex and under-representation on the X chromosome. *Hum Mol Genet.* 7:1967-78 1998.
92. Kruglyak S, Tang H. Regulation of adjacent yeast genes. *Trends Genet.* 16(3):109-11 2000.
93. Kunin V, Goldovsky L, Darzentas N, et al. The net of life: Reconstructing the microbial phylogenetic network *Genome Res* 15 (7): 954-959 2005
94. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001 Feb 15;409(6822):860-921.
95. Lawrence JG, Ochman H. Reconciling the many faces of lateral gene transfer. *Trends Microbiol.* 10(1):1-4 2002.
96. Lercher MJ, Blumenthal T, Hurst LD. Coexpression of neighboring genes in *Caenorhabditis elegans* is mostly due to operons and duplicate genes. *Genome Res.* 13(2):238-43 2003.
97. Lercher MJ, Urrutia AO, Hurst LD. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet.* 31(2):180-3 2002.
98. Lercher MJ, Urrutia AO, Pavlicek A, Hurst LD. A unification of mosaic structures in the human genome. *Hum Mol Genet.* 12(19):2411-5 2003.
99. Levine M, Tjian R. Transcription regulation and animal diversity. *Nature.* 2003 Jul 10;424(6945):147-51
100. Li Q, Lee BT, Zhang LX. Genome-scale analysis of positional clustering of mouse testis-specific genes. *BMC Genomics.* 6(1):7 2005.
101. , Lin YC, Lu CL, Chang HY, Tang CY. An efficient algorithm for sorting by block-interchanges and its application to the evolution of vibrio species. *Journal of Computational Biology*, 12:102-112 2005.
102. Lu CL, Wang TC, Lin YC, Tang CY. ROBIN: a tool for genome rearrangement of block-interchanges. *Bioinformatics*, 21(11):2780-2782 2005.
103. Ma B, Li M, Zhang LX. From gene trees to species trees. *SIAM J. Comput.* 30: 729-752 2000.
104. MacLeod D, Charlebois RL, Doolittle F, Baptiste E. Deduction of probable events of lateral gene transfer through comparison of phylogenetic trees by recursive consolidation and rearrangement. *BMC Evol Biol.* 5(1):27 2005.
105. Maddison WP Gene trees in species trees. *Systematic Biology* 46(3), 523-536 1997.
106. Mathe C, Sagot MF, Schiex T, Rouze P. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* 30(19): 4103-4117 2002.
107. Megy K, Audic S, Claverie JM. Positional clustering of differentially expressed genes on human chromosomes 20, 21 and 22. *Genome Biol.* 4(2):P1 2003.
108. Miller W, Makova KD, Nekrutenko A, Hardison RC Comparative genomics. *Annu Rev Genomics Hum Genet.* 5: 15-56 2004.
109. Mirkin B, Muchnik I, Smith TF. A biologically consistent model for comparing molecular phylogenies. *J Comput Biol.* 2(4):493-507 1995.
110. Mirkin BG, Fenner TI, Galperin MY, Koonin EV. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common

- ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol.* 3:2 2003.
111. Moret BME, Nakhleh L, Warnow T, et al. Phylogenetic networks: Modeling, reconstructibility, and accuracy. *IEEE/ACM Trans Comput. Biology and Bioinformatics* 1(1): 13-23, 2004.
 112. Moret BME, Wyman S, Bader BA, Warnow T, Yan M. A new implementation and detailed study of breakpoint analysis. *6th Pacific Symposium on Biocomputing (PSB 2001)*, 583-594 2001.
 113. , Siepel AC, Moret BME. Finding an optimal inversion median: experimental results. *Algorithms in Bioinformatics, First International Workshop, WABI 2001*, LNCS 2149:189-203 2001.
 114. Sturtevant, AH Genetic studies on *Drosophila simulans*. II. Sex-linked group of genes. *Genetics* 6:43-64 1921.
 115. Morrison DA. Networks in phylogenetic analysis: new tools for population biology. *Int J Parasitol.* 35(5):567-82 2005.
 116. Murphy WJ, Larkin DM, Everts-van der Wind A, et al. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* 309(5734): 613-617 2005.
 117. Nakhleh L, Jin G, Zhao F, et al. Reconstructing phylogenetic networks using maximum parsimony. *Proc. the 2005 IEEE Computational Systems Bioinformatics Conference*, 93-102, 2005.
 118. Nakhleh L, Warnow T, Linder CR, et al. Reconstructing reticulate evolution in species - Theory and practice *Journal of Computational Biology* 12 (6): 796-811 JUL 2005
 119. Nelson KE, Clayton RA, Gill SR, et al. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399:323-329 1999.
 120. Nordborg M and Tavaré S. Linkage disequilibrium: what history has to tell us. *Trends in Genetics*, 18, 83-90, 2002.
 121. The small phylogenetic network problem is NP-hard. Manuscript, 2005.
 122. The complexity and algorithms for the exemplar problem with gene family. Honors Thesis, School of Computing, National University of Singapore, 2005.
 123. Nguyen CT, Tay YC, Zhang LX. Divide-and-conquer approach for the exemplar breakpoint distance. *Bioinformatics.* 21(10):2171-6 2005.
 124. O'Brien SJ, Menotti-Raymond M, Murphy WJ, et al. The promise of comparative genomics in mammals. *Science.* 1999 Oct 15;286(5439):458-62, 479-81.
 125. Ohno S. *Evolution by Gene Duplication*, Springer-Verlag, 1970.
 126. Oliver B, Misteli T. A non-random walk through the genome. *Genome Biol.* 6(4):214 2005.
 127. , Ozery-Flato M, Shamir R. Two notes on genome rearrangement. *J. Bioinform. and Comput. Biol.*. 1(1):71-94 2003.
 128. Page RD. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Bio.* 43, 58-77 1994.
 129. Page RD, Charleston M From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol. Phy. and Evol.* 7, 231-240 1997.
 130. Page RD. GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics* 14(9):819-20 1998.
 131. Page RD. Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. *Mol Phylogenet Evol.* 14(1):89-106, 2000.
 132. Page RD, Cotton JA. Vertebrate phylogenomics: reconciled trees and gene duplications. In *Proc. of Pac Symp Biocomput.* pp. 536-47 2002.

133. Pevzner PA *Computational molecular biology: an algorithmic approach*, The MIT Press, Chapter 10, 2000.
134. Posada D, Crandall KA. Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol Evol.* 16(1):37-45 2001.
135. Reymond A, Marigo V, Yaylaoglu MB, et al. Human chromosome 21 gene expression atlas in the mouse. *Nature* 420(6915):582-6 2002.
136. Roy PJ, Stuart JM, Lund J, Kim SK. Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature* 418:975-9 2002.
137. Sankoff D. Edit distances for genome comparisons based on non-local operations. *CPM '92: Proc. of the 3rd Annual Symp. on Combin. Pattern Matching*, Springer-Verlag, 121-135 1992.
138. Sankoff D, Sundaram G, Kececioğlu J. Steiner points in the space of genome rearrangements *Inter. J. on Foundations of Computer Science*, 7:1-9 1996.
139. Sankoff D, Blanchette M. The median problem for breakpoints in comparative genomics. *Computing and Combinatorics, Proceedings of COCOON '97*, 251-263 1997.
140. Sankoff D. Genome rearrangement with gene families. *Bioinformatics* 15: 909-917.
141. Sankoff D, Zhang CF, Lenert A. Reversals of fortune. *Manuscript*, 2005.
142. Semple C, Steel M. Unicyclic networks: compatibility and enumeration. To appear in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.
143. Setubal J, Meidanis J. *Introduction to Computational Molecular Biology*. PWS Publishing Company, chapter 7 1997.
144. Smith GP. Evolution of repeated DNA sequences by unequal crossover. *Science* 191: 58-535 1976.
145. Smith DR, Doucette-Stamm LA, Deloughery C, et al. Complete genome sequence of *Methanobacterium thermoautotrophicum deltaH*: functional analysis and comparative genomics. *J Bacteriol.* 179(22):7135-55 1997.
146. Snyder M, Gerstein M. Genomics. Defining genes in the genomics era. *Science*. 2003 Apr 11;300(5617):258-60.
147. Song YS, Hein J Constructing minimal ancestral recombination graphs *Journal of Computational Biology* 12 (2): 147-169 2005
148. Spellman PT, Rubin GM. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J Biol.* 1:5 2002.
149. Stege U. Gene trees and species trees: The gene-duplication problem is fixed-parameter tractable, *Proc. of the 6th Inter. Workshop on Alg. and Data Structures (WADS'99)*, LNCS 1663, August 1999.
150. Stutevant AH, Novitski E. The homologies of chromosome elements in the genus *drosophila*. *Genetics*, 26:517-541.
151. Swenson KM, Pattengale ND, Moret BME. A framework for orthology assignment from gene rearrangement data. *Proceedings of RECOMB 2005 Workshop on Comparative Genomics*, LNBI 3678, 153-166, 2005.
152. Tang M, Waterman M, Yooseph S. Zinc finger gene clusters and tandem gene duplication. *J Comput Biol.* 9(2):429-46 2002.
153. Templeton A, Crandall K, Sing C. A cladistic analysis of phenotypic association with haplotype inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* 132:619-633 1992.
154. Tesler G. Efficient algorithms for multichromosomal genome rearrangements. *J. Comput. Syst. Sci.*, 65(3):587-609 2002.
155. Tesler G. GRIMM: genome rearrangements web server. *Bioinformatics*, 18(3):492-493 2002.

156. Uno T, Yagiura M. Fast algorithms to enumerate all common intervals of two permutations. *Algorithmica*, 26(2):290-309 2000.
157. Versteeg R, van Schaik BD, van Batenburg MF, et al. The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.* 13(9):1998-2004 2003.
158. Walter MET, Dias Z, Meidanis J. A new approach for approximating the transposition distance. In *Proceedings SPIRE*, 199-208 2000.
159. Waterston RH, Lindblad-Toh K, Birney E, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 2002 Dec 5;420(6915):520-62.
160. Williams EJ, Bowles DJ. Coexpression of neighboring genes in the genome of *Arabidopsis thaliana*. *Genome Res.* 14(6):1060-7 2004.
161. Wang LS, Zhang KZ, Zhang LX. Perfect phylogenetic networks with recombination *Journal of Computational Biology* 8 (1): 69-78 2001
162. Watterson GA, Ewens WJ, Hall TE, Morgan A The chromosome inversion problem, *Journal of Theoretical Biology* 99:1-7 1982.
163. Yang J, Zhang LX. On counting tandem duplication trees. *Mol Biol Evol.* 21(6):1160-3 2004.
164. Yang YS, Song HD, Shi WJ, et al. Chromosome localization analysis of genes strongly expressed in human visceral adipose tissue. *Endocrine* 18(1):57-66 2002.
165. Yancopoulos S, Attie O, Friedberg R. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340-3346 2005.
166. Zhang LX. On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *J. Comput. Biol.* 4:177-188 1997.
167. Zhang LX, Ma B, Wang L, Xu Y. Greedy method for inferring tandem duplication history. *Bioinformatics* 19:1497-1504 2003.
168. Zheng C, Lenert A, Sankoff D. Reversal distance for partially ordered genomes. *Bioinformatics*. 21 Suppl 1:i502-i508 2005.
169. Zheng C, Sankoff D. Genomoe rearrangement with partially ordered chromosomes. *Proceeding of COCOON 2005*. In press.
170. Zmasek CM, Eddy SR. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*. 17(9):821-8 2001.
171. Zhaxybayeva O, Gogarten JP. An improved probability mapping approach to assess genome mosaicism. *BMC Genomics* 4(1):37, 2003.