Analysis of Gene Order Evolution beyond Single-Copy Genes

Nadia El-Mabrouk

Département d'Informatique

et de Recherche Opérationnelle

Université de Montréal

mabrouk@iro.umontreal.ca

David Sankoff

Department of Mathematics and Statistics

University of Ottawa

sankoff@uottawa.ca

Abstract

The purpose of this chapter is to provide a comprehensive review of the field of genome rearrangement, i.e., comparative genomics based on representation of genomes as ordered sequences of signed genes. We specifically focus on the "hard part" of genome rearrangement, how to handle duplicated genes. The main questions are: how have present-day genomes evolved from a common ancestor? What are the most realistic evolutionary scenarios explaining the observed gene orders? What was the content and structure of ancestral genomes? We aim to provide a concise but complete overview of the field, starting with the practical problem of finding an appropriate representation of a genome as a sequence of ordered genes or blocks, namely the problems of orthology, paralogy and synteny block identification. We then consider three levels of gene organization: the gene family level (evolution by duplication, loss and speciation), the cluster level (evolution by tandem duplications) and the genome level (all types of rearrangement events, including whole genome duplication).

Keywords: Comparative Genomics, Gene Order, Rearrangement, Duplication, Gene Loss, Gene Family.

1   Introduction

In comparative genomics, gene orders were originally modeled as unsigned [120] or signed [99] permutations, in order to analyze data on organellar or prokaryote genomes. This provided an alternative to the classical method of comparing the DNA sequence of single genes. These models required each genome being compared to have exactly the same set of genes, in exactly one copy each. As more and more genomes have been sequenced, it is now clear that genes are not present in single copies in each genome, and that the number of copies is highly variable from gene to gene and from species to species, preventing the application of simplistic single-gene-copy model to real datasets, and requiring the represention of a genome as a sequence of genes in one or multiple copies.

The role of duplication has long been recognized in the evolution of species [85], especially in eukaryotes, where large or small sets of homologous genes, grouped into *gene families*, can be found by applying local similarity search tools. The prevalence of gene loss can also be inferred from the distribution of the number of gene copies among species. In addition to duplication and loss, the architecture of genomes is disrupted through intra- and inter-chromosomal rearrangement events, which do not change gene content, but may radically alter gene order.

Inferring the content and structure of ancestral genomes and the evolutionary scenarios that have led to the current composition and structure of present-day genomes is a major step towards answering to numerous biological questions such as the mechanisms of evolution above the DNA sequence level, variation in rearrangement rates among the different branches of a phylogenetic tree, the rates of gene loss and gain, and the consequence of such variation on the genetic and physiological specificity of species. For all of these questions, we must be able to address the different gene content of existing genomes as well as variation in the number of copies of various genes.

A variety of automated approaches have been devised to answer these questions. After introducing the general concepts of genome rearrangement and the methodological ways and difficulties of representing genomes as sets of gene orders (Section 2), this chapter recounts the contribution of computational biology to the evolutionary study of genomes based on their overall content and organization, emphasizing the problem of multiple gene copies. We will consider three levels of gene organization: the gene family level (Section 3), the cluster level (Section 4) and finally the genome level (Section 5).

At the gene family level, the pertinent events that are taken into account are speciation, duplication and loss. Understanding the evolution of gene

2

families through these events is important in evolutionary biology, phylogenomics [94, 117], and functional genomics. In this context, reconciliation between the gene tree (obtained from gene sequences) and the phylogenetic tree representing the evolution of species, is the procedure for inferring a duplication, speciation and loss history for the gene family. In Section 3, we summarize the different algorithmic approaches and optimization criteria that have been used to obtain a reconciled tree.

Duplications of chromosomal segments cover about 5% of the human genome. When multiple segmental duplications occur at a particular genomic locus they give rise to complex gene clusters. Such genomic regions are exceedingly difficult to sequence and assemble accurately, and represent a challenge for computational biology. In Section 4, we review the computational methods developed for inferring the evolution of gene clusters, for cases both of tandem and interspersed duplications. In addition to duplication and losses, inversions and other rearrangement events can affect the shape of a gene cluster.

Of major consequence is the continual disruption of gene order at the whole genome level. This leads to the *rearrangement phylogeny problem*, seeking the ancestral gene orders at the origin of a most "plausible" evolutionary scenario. The parsimony approach is based on inferring gene orders at the internal nodes of the tree so that the sum of distances among all branches is minimized. When studying genome rearrangements, the most natural distance between two gene orders is the minimum number of rearrangements required to transform one gene order into the other. In the case of two genomes G and H with no gene duplicates and the same gene content, a key result in the field of genome rearrangement is the 1995 Hannenhalli and Pevzner (HP) formula [61, 112] for computing the minimum number of inversions and translocations (including chromosomal fusions and fissions) required to transform G into H, leading to a polynomial-time algorithm. More recently, another distance that has been extensively studied is the Double-Cut-and-Join (DCJ) distance which represents a greater repertoire of rearrangement events while giving rise to simpler formal results [11, 12, 124]. For the purpose of genome rearrangement, handling duplicated genes leads to hard problems (see [3, 18, 29] for the computation of genomic distances for example). We review the *rearrangement phylogeny problem* in Section 5 emphasizing the case of multiple gene copies. The most radical evolutionary event resulting in genomes with multiple gene copies is the Whole Genome Doubling event. We focus on this particular event in the last sections of this chapter.

2 Genome rearrangements

In contrast to prokaryotes that tend to have single, often circular chromosomes, the genes in plants, animals, yeasts and other eukaryotes are partitioned among several linear chromosomes.

The genome rearrangement approach to comparative genomics focuses on the general structure of a chromosome, rather than on the internal nucleic structure of each building block. An essential prerequisite to any genome rearrangement method is thus to represent a chromosome as a linear sequence of building blocks. Usually genes are the considered building blocks of a genome, although other genetic or non-coding elements can be considered. In many cases, a "compressed" representation is provided by clustering two or more adjacent genes, as well as the intergenic sequences, into synteny blocks (see Section 2.5). In the most realistic version of the rearrangement problem, a sign (+ or -) is associated with each gene representing its transcriptional orientation. This orientation indicates on which of the two complementary DNA strands the gene is located.

In the rest of this chapter, unless otherwise stated, we will consider the case of signed building blocks, and will consider genes as the building blocks of a genome. Note that the mathematical developments in the genome rearrangement field do not depend on the fact that the objects in a linear order describing a chromosome are genes.

2.1 Genome representation

Let Σ be a set of n genes. A *string* is a sequence of genes from Σ, where each gene is signed (+ or $-$). The *reverse* of a string $X = x_1 x_2 \ldots x_r$ is the string $-X = -x_r -x_{r-1} \ldots -x_1$. A *chromosome* is a string, and a *genome* is a collection of chromosomes. A *unichromosomal* genome has a single chromosome, and a *multichromosomal* genome has at least two non-null chromosomes $C_1, C_2, \ldots C_N$. A *circular chromosome* is a string $x_1 \ldots x_r$, where x_1 is considered to follow x_r. A chromosome that is not circular is *linear*.

As most unichromosomal genomes are formed by a circular chromosome, and most multichromosomal genomes are formed by linear chromosomes, only circular unichromosomal genomes and linear multichromosomal genomes are generally considered in genome rearrangement studies.

Let G be a genome with gene content Σ. We say that G is a *singleton genome* iff each gene in Σ is present exactly once in G.

2.2   Rearrangement events

During their evolution, genomes are subject to global movements and displacements affecting their overall organization and gene order. The following are the most studied operations affecting gene orders.

- A *reversal* (or *inversion*) is an operation that changes some proper substring of a chromosome into its reverse.

- A *transposition* is an operation that cuts a proper substring of a chromosome and inserts it somewhere else in the same chromosome.

- A *translocation* between two chromosomes $X = X_1 X_2$ and $Y = Y_1 Y_2$ is an event transforming them into the two chromosomes $X_1 Y_2$ and $Y_1 X_2$, or into $X_1(-Y_1)$ and $(-Y_2)X_2$. Two special cases of reciprocal translocations are *fusions* (if one of the two chromosomes generated by the translocation is an empty string) and *fissions* (if one of the two input chromosomes is the empty string).

- Sometimes *inverse transpositions* and/or transpositions from one chromosome to another are considered elementary operations on the same footing as the others listed.

2.3   Rearrangement distances

The *rearrangement distance* is defined between two genomes G and H with the same gene content as the minimum number of rearrangement events required in a scenario transforming G into H (see Figure 1 (middle) for an example of the inversion distance). A key result in the field of genome rearrangement is the 1995 Hannenhalli and Pevzner (HP) formula [60, 61, 62] for computing the inversion, translocation and inversion+translocation distances between two singleton genomes, leading to exact polynomial-time algorithms. They are all based on a representation of the genomes G and H as a bicoloured graph called the *breakpoint graph*. Subsequently, various improvements and alternative representations of permutations have led to other algorithms, the most efficient once running in linear time [6, 10, 112]. As for the transposition distance, although many efficient bounded heuristics have been developed [7, 63, 79, 116], the complexity status of the problem remains unknown (though conjectured NP-hard).

A related distance that has been extensively studied in the last years is the DCJ distance [11, 12, 124]. Given a genome G, a *Double-Cut-and-Join*

(DCJ) is an operation that "cuts" two adjacencies ab and cd in a genome, and replaces them by either ac and bd, or ad and bc. See Figure 1 (right) for an example of the DCJ distance. The DCJ distance is interesting from a theoretical point of view as it leads to a unifying formula including all previously studied rearrangement events, as well as transpositions, for which no polynomial-time exact method is known. Computing the DCJ distance between two signed permutations is a linear-time problem [11, 124].

A simpler distance measure is the *breakpoint distance*, which is the number of disruptions between conserved segments in G and H, that is the number of pairs of genes a, b that are adjacent in one genome (contains the segment '$a\,b$') but not in the other (contains neither '$a\,b$', nor '$-b-a$'). See Figure 1 (left) for an example of the breakpoint distance. This metric, introduced in [120], is easily computed in time linear in the length of the genomes. Notice that this is equivalent to a similarity measure, namely the number of conserved gene adjacencies between the two genomes. Different generalizations of adjacency conservation to clusters involving more than two genes have been introduced in the literature (for example common intervals and gene teams discussed in Section 2.5). Some of them have been used as alternative to distance measures between two genomes [8, 22].

G: 1 ● 4 ● 3 ● 2 ● 5 6

H: 1 2 3 4 5 6

G: 1 ⌊4⌋ 3 2 5 6

1 −4 ⌊3⌋ 2 5 6

1 −4 −3 ⌊2⌋ 5 6

1 ⌊−4 −3 −2⌋ 5 6

H: 1 2 3 4 5 6

      a b   c d

G: 1 | 4 3 | 2 5 6

      ad, bc

        a b   c d

1 2 | 5 6 , 4 | 3

      ad, bc
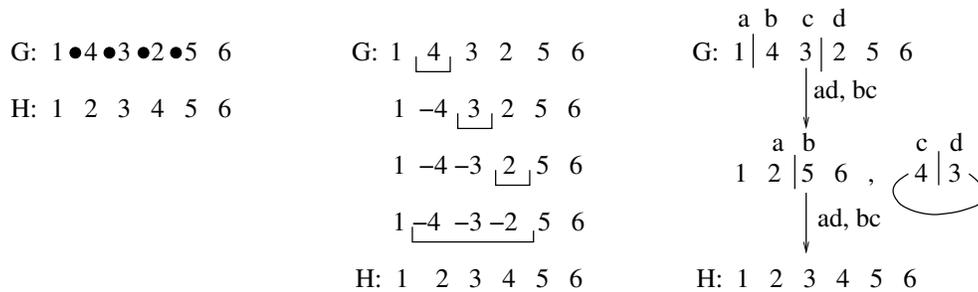
H: 1 2 3 4 5 6

Figure 1: G and H are two linear and single chromosomal genomes on the alphabet $\{1, 2, 3, 4, 5, 6\}$. (Left): The breakpoint distance between G and H is 4. Each dot represents a breakpoint in G with respect to H; (Middle): The inversion distance between G and H is 4. Each line following G is the genome obtained after applying the inversion performed on the substring underlined in the genome of the previous line; (Right): The DCJ distance between G and H is 2. Each of the two DCJ operations cuts the two adjacencies ab and cd and replaces them by ad and bc. The first DCJ creates a circular intermediate chromosome.

2.4 Gene families

Representing genomes as linear orders of genes requires a preliminary identification of the pairwise homology relationship between genes. From a conceptual evolutionary point of view, two gene copies are *homologs* if they originate through sequence divergence from the same ancestral gene. In operational practice, we attempt to identify homologs by sequence similarity. For example, using a BLAST-like method, all gene copies with a similarity score above a certain threshold would be grouped into the same *gene family* or *homology family*.

It is important to distinguish between two kinds of homology: *orthology*, the relationship between two gene copies in two genomes, where the two genes have diverged from a single gene in the most recent common ancestor of the two genomes through genome speciation (followed by independent evolution at the sequence level), and *paralogy*, the relationship between two gene copies in the same genome (due to a duplication event in that genome or in an ancestor genome) or between two gene copies in two different genomes, where the two genes have diverged from duplicate gene copies in the most recent common ancestor of these two genomes. From a functional point of view, orthologs, which are the direct descendants of a single ancestral gene copy, are more likely to be functionally related than paralogs, which originate from duplication and are a major source of gene innovation and creation of new functions [85].

In the example of Figure 2, the surviving gene copy in species 1 is orthologous to the surviving copies in genomes 2 and 3, but paralogous to the surviving copy in genome 4. A major complication in the identification of orthologs is that orthology is generally not a one-to-one relationship. Indeed, one gene copy in a phylogenetic lineage may be orthologous to a whole family of paralogs (inparalogs) in another lineage (in Figure 2, the surviving copy in genome 1 is orthologous to the two paralogous surviving copies in genome 3).

Assuming an equal rate of sequence-level evolution inside a gene family, time divergence in term of sequence similarity score can be used, at least as a first step, to discriminate orthologs from paralogs. All pairs of orthologs in two genomes should have the same divergence time, determined by the date of speciation. Paralogs are not constrained in this way. Thus, most of the existing methods for orthology assignment, such as the well-known COG system [111], the OrthoMCL [32] or INPARANOID [84] programs, just to name a few, rely mainly on sequence similarity, usually measured via BLAST scores. However, the result of a pure sequence similarity method is often questionable. Indeed, incorrect orthology assignments might be obtained if the real rates of evolution vary significantly between paralogs. Moreover, this

approach relies exclusively on local mutations, and neglects the gene order data that might provide valuable evolutionary information.

For this reason many protocols for grouping genes into families involve a second step, after filtering by sequence similarity, which takes accounts for the immediate neighbourhood of each gene copy [76, 82]. Only copies in similar neighbourhoods are kept as potential orthologs. More general methods for identifying orthologs between two genomes based on the gene order context of genes have been developed. They all begin by identifying gene families by mean of sequence similarity. The homologs are then treated as copies of the same genes, and ortholog assignment is formulated as a natural combinatorial optimization problem of rearranging one genome into another with the minimum number of events. The exemplar approach [95] selects exactly one representative of each gene family in each genome, in a way that minimizes the number of breakpoints or inversions. Other approaches maximize the number of genes matched in each family [29, 3]. A more general method allowing all gene copies to be kept, and accounting for reversals, translocations, fusions and fissions, has been developed and implemented in the MSOAR software [46, 68, 104]. Finding the most parsimonious rearrangement process transforming one genome into another constructs, as a byproduct, the list of orthologous gene pairs. (See also [125] for the implications of duplication, insertion and deletion for DCJ analyses.)

A third approach to orthology annotation in a gene family is to use pairwise sequence similarity scores to construct a gene tree for the gene family, and directly infer the duplication, speciation and loss events from this tree, by "reconciling" it with the phylogenetic tree of all the species represented. This approach will be detailed in Section 3.

2.5 Synteny blocks

An alternative for representing genomes as linear orders of building blocks is to identify sets of "conserved segments", that are not necessarily limited to single genes.

In a pioneering paper, Nadeau and Taylor [83] introduced the notion of *conserved segments*, chromosomal regions in two genomes containing the same genes in the same order. Such regions can reflect functional pressure requiring a group of genes to be close to each other on the genome. For example operons in prokaryotes, transcribed from a single messenger RNA molecule and thus required to be contiguous on the chromosome, co-expressed genes or genes part of a given biochemical pathway. Alternatively, conserved segments can

simply result from the close evolutionary relationship between two genomes: not enough time has elapsed since their speciation from a common ancestor for rearrangements to break up some groups of genes.

Based on a map of 83 mouse genes and only chromosomal assignments data for their human homologs, Nadeau and Taylor [83] estimated there to be around 180 conserved segments between the human and mouse genome. This proved to be surprisingly accurate while additional thousands of genes were added to the genetic maps [100]. As complete genomic sequences became available, however, it became clear that at higher levels of resolution, human and mouse genomes are significantly more rearranged [69]. This holds not only for "micro-rearrangement" of intergenic, non-coding DNA, but often as well for neighbouring genes within conserved regions. The complexity of genomes and the prevalence of micro-rearrangement have led to many concepts more forgiving of small rearrangements than strictly conserved segments.

In 2003, Pevzner and Tesler developed the notion of "synteny blocks" as being segments that can be converted to conserved segments by micro-rearrangements [91]. The GRIMM-Synteny algorithm they introduced by-passes the difficult issues of gene annotation and ortholog identification by constructing synteny blocks from a dot-plot of anchors, representing bidirectional best local DNA similarities between genomes (in their work, the human and mouse genomes). These anchors do not necessarily reflect similarities within genes but may also consist of similarities between non-coding regions. Synteny blocks are constructed by chaining closely located anchors, ignoring micro-rearrangements, and creating large conserved blocks on a scale similar to conserved segments predicted by Nadeau and Taylor. GRIMM-Synteny has more recently been extended to the study of multiple genomes, and to genomes exhibiting a high range of sequence duplication [90, 92]. GRIMM-Synteny is only one example of the many alignment methods that have been developed for synteny block generation.

From a combinatorial point of view, various formal models of conserved blocks of genes, also called *gene clusters* or *synteny blocks of genes* have been introduced [8, 40, 65]. In particular, the notion of *common intervals* is a first generalization of conserved segments in which we relax the conditions that genes appear in the same order or the same orientation. Formally, given K genomes represented as permutations on an alphabet Σ, a *common interval* is a subset $\mathcal{S}$ of Σ such that, in each genome, all the genes in $\mathcal{S}$ are contiguous. The notion of common intervals was first introduced by Uno and Yagiura in the case of two permutations [113], and efficient algorithms to find common intervals have been developed for K permutations [13, 64]. To avoid considering the repetitive and overlapping structure of common intervals, the

notion of a *strong common interval*, defined as a common interval that does not overlap any other common interval, has been introduced [72]. Strong intervals are likely to capture interesting biological properties as they represent a measure of maximality of gene conservation, and they allow us distinguish between local and global rearrangement events. They also have rich combinatorial properties [8]. In particular, representing them in a PQ-Tree structure allows to generate all common intervals in linear time.

The most relaxed definitions of gene clusters in permutations account for possible gaps between the conserved genes. A first formal model for max-gap clusters in permutations was introduced in [9] under the name of *gene team*, and algorithmic and statistical properties discussed in [66]. Given K genomes represented as permutations on an alphabet Σ, and given an integer $\delta \geq 0$, a *gene team* is a maximum subset $\mathcal{S}$ of Σ such that, in each genome, any gene in $\mathcal{S}$ is separated by at most δ genes from another gene of $\mathcal{S}$. Notice that a common interval is just a gene team with $\delta = 1$. The best complexity achieved to compute all the gene teams of K genomes is $O(Kn \log^2(n))$, where $n = |\Sigma|$.

Common intervals and, more generally, max-gap clusters completely abandon constraints on conservation of gene order. At the other extreme, conserved segments require complete identity of gene order. A way of introducing a degree of order conservation within gapped clusters is to require that two genes separated by at most δ genes in one genome must be separated by at most δ genes in the other [126, 135]. When $\delta = 0$, these "generalized adjacency clusters" become conserved segments, but for larger δ, common gene order becomes difficult to discern. When δ^2 approaches the number of genes in a (unichromosomal) genome, percolation occurs so that the cluster becomes the entire genome [123].

3 Reconciliation: Gene family Evolution by Duplication, Speciation and Loss

Almost all genomes which have been studied contain genes that are present in two or more copies. They may be adjacent on a single chromosome, or dispersed throughout the genome. As an example, duplicated genes account for about 15% of the protein genes in the human genome [74]. More generally, in eukaryotic genome sequences, duplicated genes account for 10% to 16% of the yeast genome, and about 20% of the worm genome [121].

Gene duplication is a fundamental process in the evolution of species [85], especially in eukaryotes [19, 35, 41, 57, 75, 117], where it is believed to play a leading role for the creation of novel gene functions. Several mechanisms

are at the origin of gene duplications: tandem repeat through slippage during recombination (see chapter 8 in [49]), gene conversion, horizontal transfer, hybridization and whole genome duplication [42, 96]. Gene loss, arising through the pseudogenization of previously functional genes or the outright deletion of chromosomal fragments, also plays a key role in the evolution of gene families [19, 35, 36, 41, 57, 75, 85].

As previously noted in Section 2.4, sequence similarity can be used to produce an initial clustering of genes into gene families. It can also be input into classical phylogenetic methods to construct a gene tree, representing the evolution of the gene family by local mutations. However, inferences about the evolution of the gene family by duplication, speciation and loss cannot be obtained directly from this gene tree alone. "Reconciliation" between the gene tree and a species tree is the most commonly used approach to infer a duplication, speciation and loss history for the gene family.

Let $\mathcal{G} = \{1, 2, \cdots, g\}$ be a set of g species. A *phylogenetic tree* or *species tree S* for $\mathcal{G}$ is a tree reflecting the evolutionary relationship among the species. More precisely, a *species tree* on $\mathcal{G}$ is a tree with exactly g leaves, where each $i \in \mathcal{G}$ is the label of a single leaf (Figure 3.(a)). A *gene tree T* on $\mathcal{G}$ is a tree where each leaf is labeled by an integer from $\mathcal{G}$ (each leaf labeled i represents a gene copy located on genome i) (Figure 3.(b)). In the presence of a strong phylogenetic signal, inferred trees are usually binary, as a speciation event usually results in the creation of two new species. Uncertainty in the phylogenetic signal can be accommodated by replacing some phylogenetic subtrees that cannot be fully resolved, by a single node, resulting in a non-binary tree. Depending on the phylogenetic reconstruction method, gene and species trees may be rooted or unrooted.

In the following sections, the input consists of a species tree S for $\mathcal{G}$ and a gene tree T for some gene family on $\mathcal{G}$, where S and T are both rooted and binary. Extensions to non-binary gene or species trees have been developed [33, 114], as well as extensions to unrooted trees [33]. Moreover, all the following developments can be directly generalized to the reconciliation of a forest of gene trees.

3.1 Incongruence between a gene tree and a species tree

Applying a classical phylogenetic method to the sequences of a family of genes generally leads to a gene tree T that is different from the species tree, mainly due to the presence of multiple gene copies in T, and that may reflect a divergence history different from S (Figure 3.(a) and (b)). Assuming no sequencing errors and a "correct" gene tree (which may be difficult to confirm), this in-

11

congruence between the two trees is a footprint of the evolution of the gene family through processes other than speciation, such as duplication, loss, gene convergence or horizontal gene transfers. It can therefore be exploited to recover the history of the gene family, and eventually decipher the orthologous and paralogous relationships among gene copies. In this section, we focus on the *duplication-loss model* of evolution, assuming an evolution of the gene family by duplications and losses only (Figure 2). The concept of reconciling a gene tree to a species tree under the duplication-loss model was pioneered by Goodman [52] and then widely accepted, utilized and also generalized to models of other processes, for example horizontal gene transfer [58].
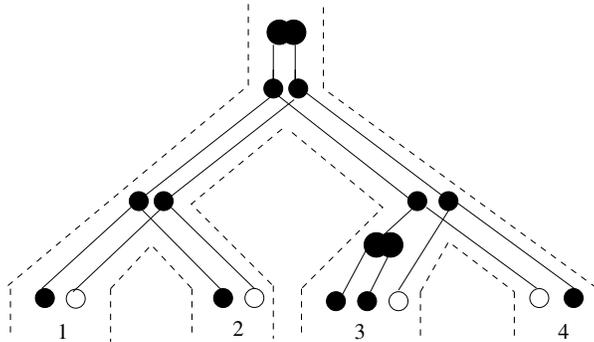
Figure 2: Evolution of a gene family by duplication, speciation and loss, embedded into the phylogenetic tree $((1,2),(3,4))$ representing the evolutionary relationship among the four species $\{1,2,3,4\}$. The double-large filled dots represent duplication events, single-small filled dots represent surviving gene copies and empty dots represent lost genes (not present in the extant species).

3.2   Definition of Reconciliation

Conceptually, a *reconciliation* between a gene tree T and a species tree S is a tree accounting for the evolutionary history of the species and all genes of the gene family, including lost and missing gene copies, by duplication, speciation and loss.

There are several formal definitions of reconciliation between a gene tree and a species tree (see section bellow). Here we define reconciliation in terms of subtree insertions, following the notation used in [27, 28, 54]. We first introduce some preliminaries:

- A *subtree insertion* in a tree T is performed by grafting a new subtree onto an existing branch of T.

- A tree T' is said to be an *extension* of T if it can be obtained from T by a sequence of subtree insertions.

- For a given vertex (or node) x of a tree T, we denote by T_x the subtree of T rooted at x and by $L(x)$ the subset of $\mathcal{G}$ defined by the labels of the leaves of T_x. $L(x)$ is called the *genome set of* x. If x is not a leaf, we denote by x_l and x_r the two children of x.

- T is said to be *DS-consistent with S* (*DS* for "*Duplication/Speciation*") if, for every vertex x of T such that $|L(x)| \geq 2$, there exists a vertex u of S such that $L(x) = L(u)$ and one of the following conditions (D) or (S) holds: (D): $L(x_r) = L(x_\ell)$; (S): $L(x_r) = L(u_r)$ and $L(x_\ell) = L(u_\ell)$.

Definition 1 *A reconciliation between a gene tree T and a species tree S is an extension $R(T, S)$ of T that is DS-consistent with S.*

For example, the tree of Figure 3.(c) is a reconciliation between the gene tree T of Figure 3.(b) and the species tree of Figure 3.(a). Such a reconciliation between T and S implies an unambiguous evolution scenario for the gene family, where a vertex that satisfies property (D) represents a duplication (duplication vertex), a vertex that satisfies property (S) represents a speciation (speciation vertex), and an inserted subtree represents a gene loss (see Figure 3.(d)).

3.3 Optimization criteria

The definition above allows for many reconciliations for given S and T. Indeed, an evolutionary model unconstrained with respect to the number of losses allows for an unbounded number of possible reconciliations. For this reason, appropriate optimization criteria, either combinatorial or probabilistic [5], should be considered. The combinatorial criteria most often considered in the literature are the number of duplications (*duplication cost*), the number of losses (*loss cost*), or both (*mutation cost*) [33, 77].

The first formal definition of a "reconciled tree" introduced by Page [86] can be reformulated as the reconciliation (following our definition of Section 3.2) of minimum size (minimum number of leaves) or, equivalently, the reconciliation minimizing the number of duplications. An equivalent constructive definition, based on a mapping, called the *LCA mapping* between the gene tree T and the species tree S, was formulated in [55, 88] and widely used [20, 39, 44, 54, 77, 86, 87, 88, 128]. The LCA mapping between T and S, denoted by M, maps every vertex x of T to the Lowest Common Ancestor (LCA) of $L(x)$ in S. This
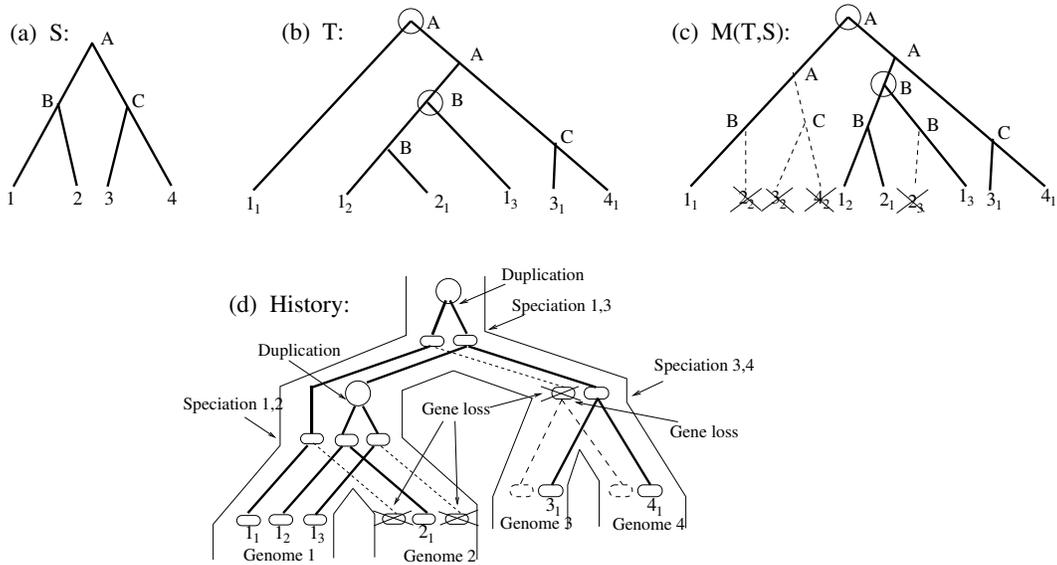
Figure 3: (a) A species tree S for $\mathcal{G} = \{1, 2, 3, 4\}$. The three internal vertices of S are named A, B and C; (b) A gene tree T. A leaf label x_y indicates the y gene copy in genome x. Internal vertices' labels are attributed according to the LCA mapping between T and S. Circles represent the duplication vertices of T with respect to S; (c) A reconciliation $M(T, S)$ of T and S. Dotted lines represent subtree insertions (3 insertions) added to construct a reconciliation, i.e. an extension of T that is DS-consistent with S. Crossed leaves represent absent gene copies that are artificially added to form the reconciliation tree. The correspondence between vertices of $M(T, S)$ and S is indicated by vertices' labels. Circles represent duplications. All other internal vertices of $M(T, S)$ are speciation vertices; (d) Evolution scenario resulting from $M(T, S)$. Each oval is a gene copy.

mapping induces a reconciliation between T and S, denoted $M(T, S)$, where an internal vertex x of T is mapped to a duplication vertex iff $M(x_\ell) = M(x)$ and/or $M(x_r) = M(x)$. See Figure 3.(a), (b) and (c) for an example.

Interestingly, $M(T, S)$ not only minimizes the duplication cost, but also minimizes the loss and mutation costs [28]. Moreover, $M(T, S)$ is the only reconciliation between T and S that minimizes the loss cost. It follows from this result that minimizing losses results in minimizing duplications. The converse is not true, as more than one reconciliation minimizing duplications may exist, in general (see Exercise 1 for an example). Stated differently, the loss cost criterion is more constraining than the duplication cost criterion for reconciliation.

Although parsimony is a convenient and widely used criterion in evolutionary inference, it is often worthwhile to investigate the wider class of near-optimal solutions. In the present context, we are thus motivated to define larger classes of reconciliations, including $M(T, S)$, but also sub-optimal solutions (with respect to the number of duplications) [38, 54, 55, 89]. This allows to explore a larger space of reconciliations and alternative evolutionary scenarios for gene families.

3.4 Algorithms

A number of algorithms have been implemented for computing $M(T, S)$ based on the LCA mapping. The two most efficient ones are those in [44, 128], the latter implemented in the program *GeneTree* [87], and both with worst-case running times of $O(n)$ for a gene tree with n leaves. An alternative, simpler algorithm running in $O(n^2)$ worst-case complexity, has also been developed in [136], for computing the LCA mapping between two trees.

From the alternative perspective of losses, [28] describes a simple algorithm for constructing the unique reconciliation tree minimizing the loss cost (which, as explained in Section 3.3, is the same tree inferred by the LCA mapping). It is based on minimizing the number of inserted subtrees required to obtain a reconciliation. As stated in Theorem 2 in [28], this algorithm can be implemented to run in $O(n)$ time and space.

Another important problem arises when the species tree S is unknown, but a number of gene trees $T_1, T_2, \cdots, T_r$ are given. The problem is to infer, from the set of gene trees, a species tree S leading to a parsimonious evolution scenario, for a chosen cost. As in the case of a known species tree, methods have been developed for the duplication and mutation cost versions of this problem [33, 59, 77]. For both criteria, inference of an optimal species tree given a forest of gene trees is an NP-hard problem [77].

3.5 Noise in Gene Trees

The main complaint about reconciliation methods is that the inferred duplication and loss history for a gene family is strongly dependent on the gene tree considered for this family. Indeed, a few misplaced leaves in the gene tree can lead to a completely different history, possibly with significantly more duplications and losses [56]. Reconciliation can therefore inspire confidence only in the case of a well-supported gene tree. Typically bootstrapping values are used as a measure of confidence in each edge of a phylogeny. How should the weak edges of a gene tree be handled? One reasonable answer is to transform the binary gene tree into an unresolved gene tree by removing each weak edge and collapsing its two incident vertices into one. Chang and Eulenstein [25] present an extension of the duplication loss model to gene trees with apparent polytomies (non-binary gene trees) and develop a polynomial time algorithm for solving this version of the reconciliation problem.

Another strategy adopted in [33] is to explore the space of gene trees obtained from the original gene tree T by performing Nearest Neighbor Interchanges (NNI's) around weakly-supported edges. The problem is then to select, from this space, the tree giving rise to the minimum reconciliation cost.

Still another possibility is to ignore gene copies leading to weak edge support. Criteria for identifying, in the gene tree, potentially misplaced or misleading leaves were given in [28], where "non-apparent" duplication vertices are flagged as potentially resulting from misplacement of one leaf in the gene tree. These concepts open the door to future developments in the correction of gene trees prior to reconciliation [37].

4 Gene Cluster Evolution

Analysis of the human genome sequence revealed the presence of many regions that have been subject to repeated local duplications, giving rise to complex gene clusters. The major mechanism causing these local duplications is unequal crossing-over during meiosis. As this phenomenon is favored by the presence of repetitive sequences, a single duplication can induce a chain reaction leading to further duplications, eventually creating large repetitive regions. When those regions contain genes, the result is a *Tandemly Arrayed Gene (TAG) cluster*: a group of paralogous genes that are adjacent on a chromosome. TAGs represent about 15% of all human genes [105] and are involved in a variety of functions such as binding and receptor activities. In particular, the olfactory receptor genes constitute the largest multigene family in verte-

brate genomes, with several hundred genes per species [51]. Other examples of TAG families include the APOBEC3 genes [73], the immunoglobulin and T-cell receptor genes [4] and the zinc finger genes [103].

As gene duplication is often followed by functional diversification, gene clusters provide a particularly interesting mechanism for rapid evolution. It is noted in [108] that a substantial fraction of what distinguishes humans from other primates, as well as the genetic differences among humans, cannot be understood until we have a clear picture of the content of gene clusters and the evolutionary mechanisms that created them. However, those repeated regions are extremely difficult to study, or even to assemble correctly. Moreover, just defining what is meant by a proper alignment of a gene cluster is a matter of discussion. Indeed, during evolution, the duplication status of segments is obscured by subsequent deletions, breaks and rearrangements. Typically, the dot-plot of a cluster self-alignment produced by a standard software such as BLASTZ [102], exhibits clouds of short interleaving alignments that cannot be directly translated into an unambiguous sequence of duplicated segments.

One solution is to restrict the study to recent duplications (those appearing clearly in the dot-plot), for example those retaining over 95% identity. In this vein, Zhang *et al.* [129, 130] proposed a method for preprocessing a self-alignment or a pairwise-alignment dot-plot, whose output represents the clusters as ordered sequences of signed atomic segments. The procedure consists of filtering out weak alignments with percentage identity less than a given threshold, processing the dot-plot such that all local alignments satisfy the "transitive closure property", and finally chaining together local alignments of similar percentage identity broken by small insertions/deletions.

Using this kind of preprocessing of dot-plots, various methods have been developed for reconstructing a hypothetical ancestral sequence and a most parsimonious set of duplications (in tandem or not) and other evolutionary events leading to the observed gene clusters [108, 115, 129, 130]. In particular, Zhang *et al.* [130] developed a simple combinatorial algorithm under the assumptions of no deletions and no boundary reuse, as well as a stochastic algorithm allowing for deletions and boundary reuse. The model was then extended in [129] for the study of orthologous TAG clusters in different species. A Bayesian version has been implemented by Vinar *et al.* [115]. A combinatorial method has also been developed in [108] for a general model involving deletions, inversions and duplications, allowing any possible placement of the duplicated segment inside the cluster (including inside the duplicated segment).

While these methods are useful to infer recent evolutionary events, they are less appropriate for longer time scales, as alignment of non-functional regions becomes impossible due to mutations (such as indels and substitutions)

continuously affecting each duplicated segment. An alternative and complementary approach is to focus on the genes present in the cluster. Indeed, as coding regions are usually characterized by lower evolutionary rates than surrounding non-coding regions, they provide a phylogenetic signal that can be used in combination with gene order data to infer evolutionary histories in which duplication events are explicitly determined. In the following section, we review the algorithmic methods that have been developed for studying the evolution of TAG clusters.

## 4.1 The Tandem-Duplication Model of Evolution for TAGs
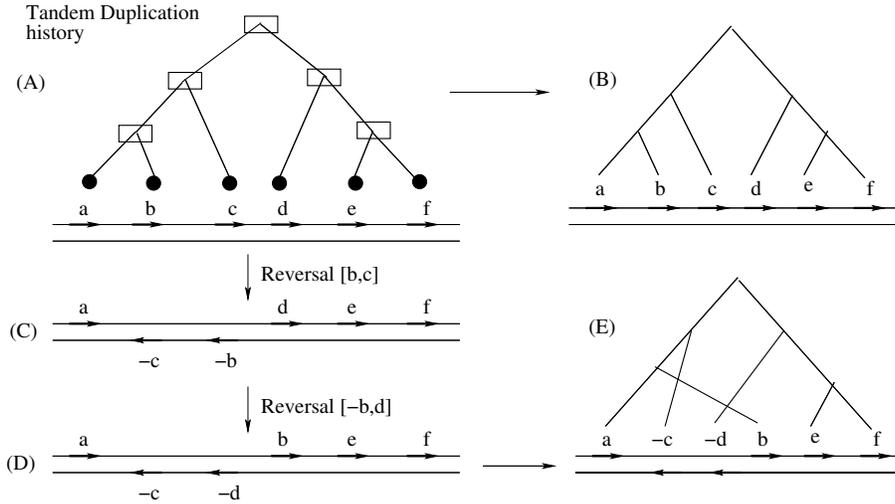
Figure 4: (A): a tandem duplication history leading to a cluster with six genes (from a to f). Each rectangle denotes a simple tandem duplication. The resulting gene order on the two DNA strands are shown below the tree; (B): the duplication tree resulting from (A); (C) and (D): The gene orders obtained after the first and second reversals indicated, respectively; (E): The gene tree resulting from the duplication and reversal history of the gene family. (E) is not a duplication tree.

The first model of evolution to consider TAGs added tandem duplications resulting from unequal recombination to the point mutations classically assumed to be the sole evolutionary mechanism acting on sequences [45]. Formally, from a single ancestral gene at a given position in the chromosome, the *tandem-duplication model* of evolution assumes that the locus grows through

18

a series of consecutive duplications placing the newly created copy next to the original one. Such tandem duplications may be *simple* (duplication of a single gene) or *multiple* (simultaneous duplication of neighboring genes).

Several studies have considered the problem of inferring an evolutionary history for a TAG cluster [15, 43, 109, 127]. These are essentially phylogenetic inference methods using the additional constraint that the resulting tree should induce a duplication history according to the given gene order. Such trees are called *duplication trees* (see Figure 4, (A) and (B)). However, it is often impossible to reconstruct a duplication history for a TAG cluster [49], even from well-supported gene trees. This is due to the occurrence of other mechanisms, such as deletions and genomic rearrangements [41], during the evolution of the gene family (Figure 4, (C), (D) and (E)).

An attempt for incorporating gene losses into the tandem-duplication model of evolution has been made by Chaudhuri *et al.* [26]. This *tandem duplication-loss model* of evolution assumes that a genome evolves through a sequence of tandem duplication-loss events, where a *tandem duplication-loss event* is a tandem duplication immediately followed by the loss of one copy of each duplicated gene. It is rather unrealistic, requiring that gene content and number remain unchanged during evolution (evolution from a permutation to another permutation).

A generalization of the tandem-duplication model allowing for inversions has been developed by Lajoie *et al.*. In [71], they present an exact branch-and-bound algorithm for the inversion distance, and a polynomial-time heuristic for the simpler breakpoint distance. The former algorithm permits the calculation of the minimum number of inversions involved in the evolutionary history of a TAG cluster in a single species, by simple tandem duplications and inversions. The model was extended in [16] to the study of orthologous TAG clusters in different species. Given the gene and species trees for a set of orthologous TAG clusters and their respective gene orders, this paper considers the problem of inferring the ancestral gene orders leading to a most parsimonious sequence of evolutionary events. The algorithm proceeds in two steps. First, ignoring gene orders, a classical gene tree/species tree reconciliation method is used to infer a "minimal" duplication, speciation and loss history in agreement with a known species tree. Second, ancestral gene orders are inferred that are consistent with minimizing the number of inversions required to obtain a valid duplication tree.

Both methods in [71] and [16] were developed under the assumption of simple tandem duplications only. However, while allowing for exact algorithmic solutions, this assumption is an important limitation to its applicability. A heuristic algorithm in [70] produces a set of optimal evolutionary histories for

a TAG cluster in a single species, allowing for tandem duplications, inverted tandem duplications, inversions and deletions, each event involving one or a set of adjacent genes. Experiments on simulated data showed that the most recent evolutionary events can be inferred accurately when the exact gene trees are used. Despite the uncertainty associated with the deeper parts of the reconstructed histories, they can be used to infer the duplication size distribution with some precision. The extension of this algorithm to consider the evolution of a cluster in multiple species is a challenging direction for future research.

5 Genome Evolution

The evolution of genomes is most often represented by a phylogenetic tree, though in some contexts, such as massive horizontal transfer of genes among prokaryotes or evolution within species, a reticulate or network representation may be required. We separate the problem of reconstructing or inferring a tree from data on present-day genomes into two parts. The "large" phylogenetic problem is one of finding the topology, or branching pattern, of the tree connecting the given genomes represented by the terminal nodes, or leaves, of the tree. The "small" problem is the inference, for a given phylogeny, of the ancestral genomes identified with each of the non-terminal nodes of the tree. This section is dedicated to the small phylogenetic problem.

5.1 The distance-based approach

We can approach the small problem by minimizing total branch length over a phylogeny while reconstructing optimal ancestral gene orders. Formally, let S be a phylogeny (i.e. a species tree) where each of the N_t terminal nodes (leaves) is labelled by a known gene order on the same n genes, and let d be a metric on the set of gene orders. Each branch of S may be incident to at most one terminal node and at least one of the N_a ancestral nodes. Each non-terminal node is of degree at least three. We want to reconstruct $R = (G_1 \ldots, G_{N_a})$, a set of gene orders at the ancestral nodes that minimize

$$L(R) = \sum_{\text{branch } XY \in S} d(XY). \tag{1}$$

The archetypical (unrooted) phylogeny has three or more leaves and exactly one non-terminal node, as on the top of the Fig. 5. The problem becomes that of reconstructing a single gene order M, the sum of whose distances to the given gene orders is minimal. An early algorithm for this "median" problem
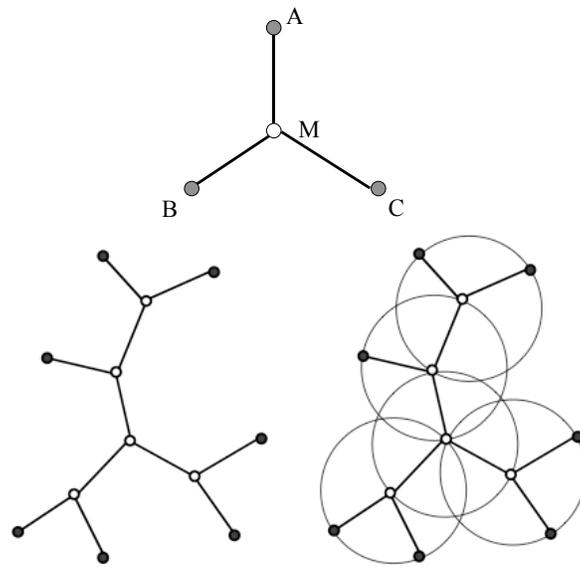
Figure 5: (top) Median problem: given genomes A, B, C, find M such that $d(A, M) + d(B, M) + d(C, M)$ is minimized. (left) Example of unrooted phylogeny with given present-day genomes at terminal nodes (dark dots) and genomes to be inferred at the ancestral nodes (white dots). (right) Inference of genomes at ancestral nodes found by iterating through the ancestral vertices, solving a median problem at each step.

[97] is based on the breakpoint median. Technical speedups were described by Cosner *et al.* [34] and incorporated into the GRAPPA software [67]. Siepel [106] and Caprara [24] gave exact median algorithms for small instances of reversal distance and Bourque [21] and Moret *et al.* [81] released heuristic web applications for this version of the problem.

For most formulations, in terms of different kinds of genome and different distances, the median problem is known (or thought) to be NP-hard; recently, however, for the case of breakpoint distance on multichromosomal genomes not restricted to be linear, Tannier *et al* [110] have given a polynomial-time algorithm, and this has been implemented [1] as a rapidly executing program.

Much progress has been made recently on exact algorithms for the DCJ distance capable of handling large or moderate size genomes [122]. As might be expected for an NP-hard problem, all exact methods encounter bad cases that require prohibitive computing time to solve. For the median problem this occurs frequently once the length of the branches approach 15 or 20 % of the number of genes. There are heuristic methods [131, 132] that are not very sensitive to the branch lengths, but when the distance becomes 25 or 30 % of the number of genes, these methods give results that may be significantly far from optimal.

For the more general small phylogeny problem with more than one ancestral node, an effective heuristic strategy is based on the ability of the median algorithm to achieve a fairly accurate solution in a reasonable time on a large proportion of instances. As illustrated at the bottom of Fig. 5, the phylogeny at the left is decomposed on the right into a set of overlapping median configurations, with one non-terminal, i.e., ancestral, node as median, and all its (three or more) co-linear nodes, terminal or non-terminal. The heuristic consists of solving each of the median problems in turn, updating the median at each step only if it diminishes the sum of the lengths of the branches incident to the median, and iterating. This eventually converges to a local minimum. The quality of the solutions may depend on the initialization of the ancestral gene orders [98], e.g., by random gene orders, or by copying some of the present-day gene orders to the ancestral nodes. It may also depend on various techniques for escaping from local minima [132].

5.2   The synteny-based approach

The also called "local" [30] or "model-free" [31] approach has three steps:

1. Inference of ancestral gene content. Assuming a model with no convergent evolution and minimum losses, the most natural is to assign a given

gene x to each internal node on the paths from the node representing the Lowest Common Ancestor (LCA) of all leaves containing x, to the leaves containing x (Figure 6.(1));

2. Inference of a set of potential ancestral syntenies (PASs) at a given internal node of interest based on the observed gene order conservation in extant species. Usually this inferred set involves a number of conflicts, i.e. pairs of syntenies that cannot co-occur in a single ancestral chromosome. For example, the two gene adjacencies xy and xz constitute a conflict. To cope with this difficulty, a weight is usually attributed to each potential ancestral synteny, reflecting its reliability and support with respect to the phylogeny;

3. Chaining ancestral syntenies in an "optimal" and non-ambiguous way, to form a set of Contiguous Ancestral Regions ($CARs$) [78].

The main difference from the distance approach is that in the absence of a complete set of ancestral syntenies, the output is a set of ancestral regions instead of a completely assembled ancestral genome. In other words, it is less ambitious than the distance approach as it does not propose a rearrangement scenario, neither does it ensure that the inferred CARs represent complete chromosomes, but the predicted ancestral syntenies are likely to be more reliable as they are more directly deduced from observed conservations in the extant species.

Steps 2. and 3. of the synteny-based approach can be implemented in several ways, and the algorithms using such approach mainly differ in: (i) the definition of synteny (adjacencies, common intervals, max-gap intervals); (ii) the method used to infer ancestral syntenies; (iii) the weight (statistical support) attributed to each potential ancestral synteny; (iv) the method used for resolving conflicts and the one used for chaining syntenies.

The first formal method based on this approach was developed by Ma *et al.* [78]. In this algorithm: (i) Syntenies are adjacencies; (ii) Sets of PASs at a given internal node are computed by the Fitch parsimony algorithm (see Figure 6.(2.1) for more details); (iii) Weights are given in an ad-hoc manner, depending on the depth of a breakpoint in the phylogeny; (iv) The set of PASs at a node is represented as a directed graph. A greedy heuristic approach is then used to output a set of paths that covers all the nodes of the graph and, at the same time, maximizes the total edge weights in the paths.

An alternative approach is considered in Bertrand *et al.* [14], involving a more general algorithm to be discussed further in Section 6.3. In this algorithm: (i) Syntenies are adjacencies; (ii) In contrast with the previous ap-
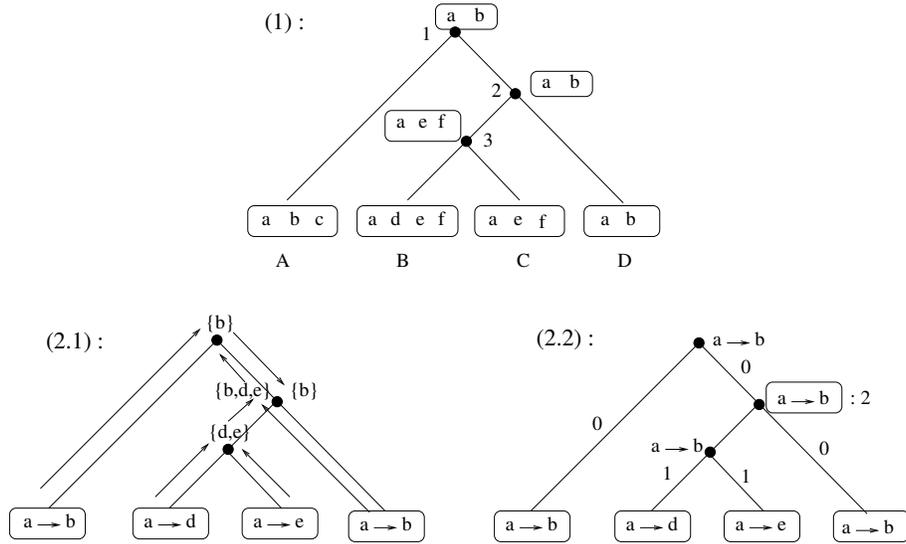
Figure 6: (1) A species tree for the species $\{A, B, C, D\}$ described by their gene content. Lower cases are gene names. Ancestral gene content is inferred as described in the text (step (1) of the synteny-based approach); (2.1) Computing the set of potential ancestral adjacencies by the algorithm of Ma *et al.*. Right adjacencies of a in extent species are indicated on leaves. For example, $a \rightarrow b$ means that ab is an adjacency in the corresponding genome. The algorithm proceeds as follows: In a bottom-up traversal (indicated by bottom-up arrows), we compute the set of potential adjacencies of a for each internal node x as follows: take the intersection of the sets computed for the two children of x if this intersection is non-empty, and the union otherwise. Then in a top-down traversal, prune the obtained set at each internal node x by taking the intersection of this set with that of x's father if this intersection is non-empty. Only the top-down arrow leading to a pruning is shown (internal node 2); (2.2) Computing the weight of an adjacency by the algorithm of Bertrand *et al.* All adjacencies are possible at each internal node. Here, the score of adjacency ab at node x is 4 as it is the maximum number of conserved right adjacencies for a in the whole tree. A possible set of ancestral adjacencies leading to this weight is shown (internal node labels).

proach, all adjacencies of a gene observed in the extant species are considered as potential adjacencies at each internal node of the tree; (iii) As the main contribution of the method, a rigorous weight is attributed to each adjacency of each gene, representing the maximum number of conserved adjacencies of that gene in the tree (see Figure 6.(2.2) for more details). An exact dynamic programming algorithm is used for this step.

Using a more relaxed definition of synteny, Chauve and Tannier [31] developed an alternative "model-free" methodology with the following properties: (i) Syntenies are defined as a combination of maximum common intervals, unsigned adjacencies and approximate common intervals; (ii) a group of genomic markers is potentially contiguous in an ancestral genome (form a PAS) if it is contiguous in at least two extant species whose evolutionary path on the phylogenetic tree goes through the ancestral node being considered; (iii) a weight is attributed to each PAS following the weighting scheme used in [78]; (iv) at a given ancestral node, all PASs are encoded by a 0/1 matrix $\mathcal{M}$, where each row i represents a given synteny S_i, each column a given marker j, and $\mathcal{M}(i,j) = 1$ if marker j belongs to S_i, and 0 otherwise. Then, an approach known in graph theory as the *Consecutive Ones problem (C1P)* [47] is used: if the matrix can be reordered to satisfy C1P, then the set of syntenies has no conflicts, and the C1P ordering of $\mathcal{M}$ can be translated directly into a set of ancestral CARs. Otherwise, the problem reduces to the one of removing the minimum number of rows from $\mathcal{M}$ leading to a C1P matrix.

6 Genome duplication

Whole genome duplication (WGD) is perhaps the most spectacular mechanism giving rise to multigene families. Normally a lethal accident of meiosis, if genome doubling can be resolved in the organism and eventually fixed as a normalized diploid state in a population, it constitutes a duplication of the entire genetic material. Right after the WGD event, the resulting genome is a perfect set of duplicated chromosomes. However, subsequent evolutionary events such as rearrangements, losses and local duplications blur this initial perfect duplicate status. Usually, a hypothesis that a given species has been subject to a whole genome duplication event during its evolution is based on the discovery of numerous pairs of syntenic regions on two different chromosomes (or regions of a single chromosome) within the same genome, covering a high proportion of the genome. Such evidence for WGD events has shown up across the whole eukaryote spectrum, from the protist *Giardia* to brewer's yeast, most flowering plant lineages, several insects, fish, amphibians, and mammalian species.

In plant lineages, those angiosperm genomes that have been completely sequenced to date all show evidence of WGD events: three ancient polyploidy events have been revealed in the *Arabidopsis thaliana* genome [17, 23], one in the rice genome that might characterize all monocots (in the grass family, maize reveals an additional WGD) [93], and others by the poplar, grape and papaya genomes [107].

In most of the cases, analyzing the duplication status of syntenies in extant species allows us to position the WGD events on the species tree. Each WGD node has a single descendant node, in contrast to the binary (at least) branching at speciation nodes (Figure 7). The content of ancestral genomes is easily inferred from that of extant species (simple extension of the method illustrated in Figure 6, taking into account gene multiplicity). However, inferring ancestral gene orders is far from being a simple task, and generalizing either the distance-based or the synteny-based approach to a phylogeny with WGD nodes rases many difficulties.
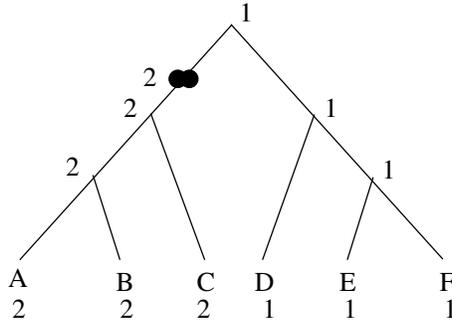
Figure 7: A phylogenetic tree exhibiting the evolution of six modern species A, B, C, D, E and F by speciation and a WGD event (the single-child node indicated by a double dot). Each number indicate gene multiplicity in the genome labeling the considered leaf or internal node.

Consider the phylogenetic tree of Figure 7, where the double-dot is a WGD node. In order to generalize the median algorithm (Section 5.1) to be applicable to such a tree, one have to be able to solve the median problem in each of the following cases:

1. three non-duplicated genomes (D, E and F);

2. two non-duplicated genomes and one duplicated genome (for example C, D and F);

26

3. one non-duplicated genome and two duplicated genomes (for example A, B and D);

4. three duplicated genomes (for example A, B and D).

While the first case is just the standard median problem, the three next cases require specific developments [134]. In Section 6.1, we first introduce the "Genome Halving Problem", that, ignoring any phylogenetic context, asks for the ancestral pre-duplicated genome of a single genome. We then introduce in Section 6.2 the "Guided genome halving" problem, which is a generalization of genome halving considering a non-duplicated outgroup. Solutions of these two problems have been used in [134] to compute the median of three genomes in case 2. (Exercise 3) and case 3.

As for the synteny-based approach, it is less problematic to generalize it to an evolution by WGD, insofar as the desired ancestral genome is the one preceding the oldest WGD event in the tree. This is developed in Section 6.3. However, inferring an ancestral genome at a node u that is a descendant of a WGD node causes difficulty in chaining PASs into CARs, as markers are present in multiple copies at u. For example, suppose we have inferred the two right adjacencies ab and ac for the gene a at u, and also one left adjacency da. Then, should this left adjacency be chained to ab, to form the PAS dab, or to ac to form the PAS dac? Clearly, other criteria than individual adjacencies should be used to handle this open problem.

6.1 Genome halving

The *Genome Halving Problem* asks, given a genome T with two copies of each gene, distributed in any manner among the chromosomes, to find the ancestral "perfectly duplicated" genome, written $A \oplus A$, consisting of two identical halves, i.e., two identical sets of chromosomes with one copy of each gene in each half, such that the rearrangement distance $d(T, A \oplus A)$ between T and $A \oplus A$ is minimal. Note that part of this problem is to find a labelling as "1" or "2" of the two genes in a pair of copies of T, so that all n copies labelled "1" are in one half of $A \oplus A$ and all those labelled "2" are in the other half. The genome A represents the ancestral genome at the moment immediately preceding the WGD event giving rise to $A \oplus A$.

For reversal and translocation distance, a linear-time solution was discovered in 1999 [42]. For reversal distance, these results have been reformulated [2] using an alternative representation of the breakpoint graph. There are also versions for DCJ [80, 118] and for breakpoint distance [110].

Generalizations of the algorithms to doubled genomes with missing gene copies have also been developed [48, 101].

6.2 Guided halving and gene order reconstruction in phylogenies with WGD

A problem with genome halving is that there are usually many, very different, perfectly duplicated genomes $A \oplus A$ leading to a minimum distance with T. For biological purposes it would be preferable to be able to use some additional, or external, information to choose amongst these solutions. Thus the *Guided Genome Halving problem* [50, 101, 133] asks, given a genome T, as well as another genome R containing only one copy of each of the n genes (a non-duplicated outgroup), find A so that $d(T, A \oplus A) + d(A, R)$ is minimal. The solution A need not be a solution to the original halving problem. The reversals and translocations version and the DCJ version of this problem are NP-hard [110].

Guided halving using the heuristic pathgroups approach [131] extends naturally to gene order reconstruction in phylogenies containing WGD events [132].

6.3 The synteny-based approach

In [53], Gordon *et al.* used a "manual" synteny-based approach to reconstruct the gene order and content of the yeast ancestor that existed immediately prior to the WGD event in the evolutionary history of many present-day yeast species, among those *S.cerevisiae*. Based on the gene set of each of the eleven available yeast species (five of them being non-duplicated species), ancestral syntenies were inferred as follows: using a sliding-window method (window of size 25), identify *Double Conserved Syntenies* (or DCS) in each post-duplicated genome. These are pairs of "syntenic" (or homologous) regions in a post-duplicated genome that are homologous to a single region in a pre-duplicated genome. One copy of each pair of DCS is then inferred to be a synteny in the pre-duplicated ancestor. The final ancestral genome obtained after chaining the ancestral syntenies has a predicted number of 8 chromosomes.

Inferring ancestral syntenies based on a formal definition of Gordon's DCSs, and using the automatic synteny-based approach developed in [31], Tannier showed that results obtained on the yeast species are very similar to those obtained with the manual approach, while avoiding repetitive and tedious work. He notes that achieving better convergence with manual processes would require us to "refine the principles of the local method" in order to take all

ancestral genes into account, and correctly weight the ancestral syntenies according to their phylogenetic signal. This is precisely the contribution of the new methodology developed in [14], where the weight of a potential ancestral adjacency (a, b) reflects the maximum number of times a and b can be adjacent in the whole tree, for any setting of ancestral genomes. The method was described above (in Section 5.2) for the case of evolution without WGDs. The dynamic programming algorithm used for computing adjacency weight generalizes to the case of WGDs. As noticed earlier in this section, chaining adjacencies into CARs is problematic at an ancestral node below a first WGD node. However, inferring the ancestral genome preceding the first WGD node is identical to the non-WGD case, as the constructed genome contains only one copy of each gene. Applying the algorithm in [14] to the data sets of the eleven yeast genomes considered in Gordon *et al.*, yields very similar results.

7   Genome Aliquoting

Whole genome doubling is not the only process that results in multiple copies of each chromosome in a genome. Hexaploidy, octoploidy, etc., are conditions where the genome has been tripled, quadrupled, etc. Warren has generalized the genome halving problem to one of genome *aliquoting* [119]:

Given a genome T with $p \geq 2$ copies of each gene, distributed in any manner among the chromosomes, to find the "ancestral" genome, written $A \oplus A \oplus \cdots \oplus A$, consisting of p identical parts, i.e., p identical sets of chromosomes with one copy of each gene in each part, such that the rearrangement distance $d(T, A \oplus A \oplus \cdots \oplus A)$ is minimal. Part of this problem is to find an optimal labelling as $1, 2, \ldots$ or p of the p copies of each gene, so that all n copies labelled "1" are in one part of $A \oplus A \oplus \cdots \oplus A$ and all those labelled "2" are in a separate part, and so on. The genome A represents the ancestral genome at the moment immediately preceding the *polyploidization* event giving rise to $A \oplus A \oplus \cdots \oplus A$.

Warren provided an efficient algorithm for the solution of genome aliquoting for DCJ [119], though the complexity of this problem has not yet been established.

8   Conclusions

The extension of genomic comparison theory to allow duplicate genes and gene families in a genome gives rise to a variety of new combinatorial optimization

problems. This has set the stage for new algorithmic results, but the difficulty in solving many of these problems ensures that a great deal of work remains to be done.

Projects for genome sequencing and analysis routinely encounter the problems due to duplication and paralogy we have discussed here. The biologists and bioinformaticians supporting them use many of the techniques we have discussed in a piecemeal way or develop *de novo* heuristics to solve the problems in ways specific to the particular genomes under study. At the same time, those working in combinatorial optimization methods use small invented problems, simulated data or the occasional full-scale real data to which they may have access. It is where these two currents intersect that the most interesting ideas emerge. More detailed characterization of biological structures and processes encourage us to relax the simplifying assumptions that lead to strong but irrelevant theoretical results, while serious attention to formal critera and analysis can avoid an unnecessary reliance on heuristics and help understand the limitations of non-unique reconstructions. Fortunately, this convergence of disciplines is on the increase.

9 Exercises

Exercise 1: Let $\mathcal{G}$ be a genome set, T be a gene tree on $\mathcal{G}$ and S be a species tree for $\mathcal{G}$.

1. Find the reconciliation $M(T, S)$ between the trees T and S of Figure 8 minimizing the loss cost.
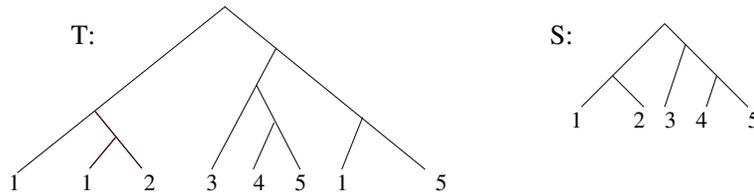
Figure 8: A gene tree T and a species tree S on $\mathcal{G} = \{1, 2, 3, 4, 5\}$.

2. As stated in Section 3.3, the reconciliation minimizing the loss cost is unique, and it is also guaranteed to minimize the duplication cost. By using the trees of Figure 8, show that the converse is not true. In other words, find a reconciliation between T and S that minimizes the duplication cost, but not the loss cost.
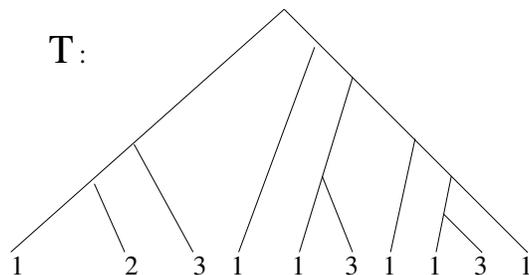
Figure 9: A gene tree T on $\mathcal{G} = \{1, 2, 3\}$.
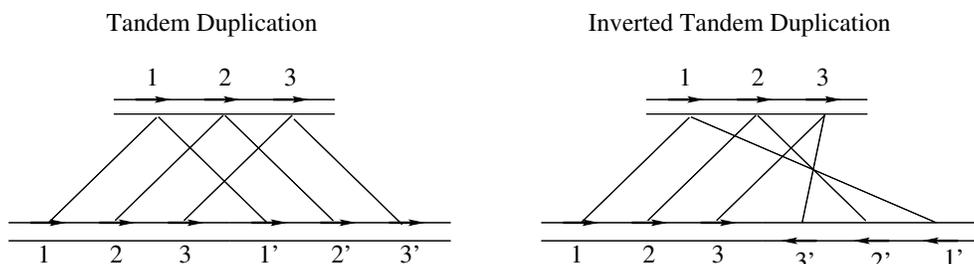
Figure 10: A chromosomal segment containing three consecutive genes $1, 2$ and 3 evolves into a segment containing 6 consecutive genes $1, 2, 3, 1', 2'$ and $3'$, where x' is the copy of x, by: A Tandem Duplication leading to the gene order $1, 2, 3, 1', 2', 3'$; An Inverted Tandem Duplication leading to the gene order $1, 2, 3, -3', -2', -1'$.

3. Suppose now that the species tree S is unknown. Let T be a gene tree. We consider the problem of finding a most parsimonious history of Duplication/Loss/Speciation explaining T. More precisely, the problem is to find a reconciliation between T and any species tree S for $\mathcal{G}$ minimizing the duplication cost (the minimum duplication problem) or minimizing the loss cost (the minimum loss problem).

   By using the gene tree of Figure 9, show that there is not always a common solution to the minimum duplication and minimum loss problem.

Exercise 2: Consider a TAG cluster, represented as an ordered sequence of n genes, and a gene tree T for these genes. We assume that the TAG cluster has evolved from a common ancestral gene through tandem duplications and inverted tandem duplications (see Figure 10). We want to find the set of all possible most recent events that have led to the observed TAG.

1. Develop an algorithm that finds, in linear time, the set of all possible

most recent tandem duplication events.

2. Develop an algorithm that finds, in linear time, the set of all possible most recent inverted tandem duplication events.

Exercise 3: Use the *Genome halving* (Section 6.1) and *Generalized genome halving* (Section 6.2) problems to develop a heuristic for computing the median of one duplicated genome (descendant from a WGD event) and two non-duplicated outgroups (case 2., Section 6).

References

[1] Z. Adam and D. Sankoff. A statistically fair comparison of ancestral genome reconstructions, based on breakpoint and rearrangement distances. *Journal of Computational Biology*, 17:1299–1314, 2010.

[2] M.A. Alekseyev and P.A. Pevzner. Colored de Bruijn graphs and the genome halving problem. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4:98 – 107, 2007.

[3] S. Angibaud, G. Fertin, I. Rusu, and S. Vialette. A general framework for computing rearrangement distances between genomes with duplicates. *Journal of Computational Biology*, 14:379–393, 2007.

[4] B. Arden, S.P. Clark, D. Kabelitz, and T.W. Mak. Human T-cell receptor variable gene segment families. *Immunogenetics*, 42:455–500, 1995.

[5] L. Arvestad, A.-C. Berglung, J. Lagergren, and B. Sennblad. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In D. Gusfield, editor, *RECOMB '04: Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology*, pages 326–335, New York, 2004. ACM.

[6] D.A. Bader, B.M.E Moret, and M. Yan. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *Journal of Computational Biology*, 8:483 – 491, 2001.

[7] V. Bafna and P. A. Pevzner. Sorting by transpositions. *SIAM Journal on Discrete Mathematics*, 11:224–240, 1998.

[8] A. Bergeron, C. Chauve, and Y. Gingras. Formal models of gene clusters. In I. Mandoiu and A. Zelikovsky, editors, *Bioinformatics algorithms: techniques and applications*, chapter 8. Wiley, 2008.

[9] A. Bergeron, S. Corteel, and M. Raffinot. The algorithmic of gene teams. In R. Guigó and D. Gusfield, editors, *Algorithms in Bioinformatics. Proceedings of WABI 2002*, volume 2452 of *Lecture Notes in Computer Science*, pages 464–476. Springer, 2002.

[10] A. Bergeron, J. Mixtacki, and J. Stoye. Reversal distance without hurdles and fortresses. In S.C. Sahinalp, S. Muthukrishnan, and U. Dogrusoz, editors, *Combinatorial Pattern Matching '04*, volume 3109 of *Lecture Notes in Computer Science*, pages 388 – 399, 2004.

[11] A. Bergeron, J. Mixtacki, and J. Stoye. A unifying view of genome rearrangements. In *Algorithms in Bioinformatics. WABI '06*, volume 4175 of *Lecture Notes in Computer Science*, pages 163 – 173, 2006.

[12] A. Bergeron, J. Mixtacki, and J. Stoye. A new linear time algorithm to compute the genomic distance via the double cut and join distance. *Theoretical Computer Science*, 410:5300 – 5316, 2009.

[13] A. Bergeron and J. Stoye. On the similarity of sets of permutations and its applications to genome comparison. *Journal of Computational Biology*, 13:1340–1354, 2003.

[14] D. Bertrand, Y. Gagnon, M. Blanchette, and N. El-Mabrouk. Reconstruction of ancestral genome subject to whole genome duplication, speciation, rearrangement and loss. In V. Moulton and M. Singh, editors, *Algorithms in Bioinformatics, WABI '10*, Lecture Notes in Computer Science, pages 78–89, 2010.

[15] D. Bertrand and O. Gascuel. Topological rearrangements and local search method for tandem duplication trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2:15–28, 2005.

[16] D. Bertrand, M. Lajoie, and N. El-Mabrouk. Inferring ancestral gene orders for a family of tandemly arrayed genes. *Journal of Computational Biology*, 15(8):1063-1077, 2008.

[17] G. Blanc, K. Hokamp, and K.H. Wolfe. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Research*, 13:137 – 144, 2003.

[18] G. Blin, C. Chauve, G. Fertin, R. Rizzi, and S. Vialette. Comparing genomes with duplications: a computational complexity point of view. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4:523–534, 2007.

[19] T. Blomme, K. Vandepoele, S. De Bodt, C. Silmillion, S. Maere, and Y. van de Peer. The gain and loss of genes during 600 millions years of vertebrate evolution. *Genome Biology*, 7:R43, 2006.

[20] P. Bonizzoni, G. Della Vedova, and R. Dondi. Reconciling a gene tree to a species tree under the duplication cost model. *Theoretical Computer Science*, 347:36–53, 2005.

[21] G. Bourque and P.A. Pevzner. Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Research*, 12:26 – 36, 2002.

[22] G. Bourque, Y. Yacef, and N. El-Mabrouk. Maximizing synteny blocks to identify ancestral homologs. In *Lecture Notes in Bioinformatics*, volume 3678 of *RECOMB-CG*, pages 21-34. Springer, 2005.

[23] J.E. Bowers, B.A. Chapman, J. Rong, and A.H. Paterson. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 422:433 – 438, 2003.

[24] A. Caprara. On the practical solution of the reversal median problem. In O. Gascuel and B.M.E. Moret, editors, *Algorithms in Bioinformatics (WABI). First International Workshop*, volume 2149 of *Lecture Notes in Computer Science*, pages 238–251. Springer, 2001.

[25] W.C. Chang and O. Eulenstein. Reconciling gene trees with apparent polytomies. In D.Z. Chen and D. T. Lee, editors, *Proceedings of the 12th Conference on Computing and Combinatorics (COCOON)*, volume 4112 of *Lecture Notes in Computer Science*, pages 235–244, 2006.

[26] K. Chaudhuri, K. Chen, R. Mihaescu, and S. Rao. On the tandem duplication-random loss model of genome rearrangement. SODA, 2006.

[27] C. Chauve, J.-P. Doyon, and N. El-Mabrouk. Gene family evolution by duplication, speciation and loss. *J. Comput. Biol.*, 15:1043-1062, 2008.

[28] C. Chauve and N. El-Mabrouk. New perspectives on gene family evolution: losses in reconciliation and a link with supertrees. In S. Batzoglou, editor, *Research in Molecular Biology (RECOMB 2009)*, volume 5541 of *Lecture Notes in Computer Science*, pages 46–58. Springer, 2009.

[29] C. Chauve, G. Fertin, R. Rizzi, and S. Vialette. Genomes containing duplicates are hard to compare. In *Computational Science (ICCS 2006)*, volume 3992 of *Lecture Notes in Computer Science*, pages 783–790, 2006.

[30] C. Chauve, H. Gavranovic, A. Ouangraoua, and E. Tannier. Yeast ancestral genome reconstructions: the possibilities of computational methods. *PloS Computational Biology*, 4(11):e1000234, 2008.

[31] C. Chauve and E. Tannier. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PloS Computational Biology*, 4:e1000234, 2008.

[32] F. Chen, A.J. Mackey, Jr C.J. Stoeckert, and D. S. Roos. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, 34:D363– D368, 2006.

[33] K. Chen, D. Durand, and M. Farach-Colton. Notung: Dating gene duplications using gene family trees. *Journal of Computational Biology*, 7:429–447, 2000.

[34] M.E. Cosner, R.K. Jansen, B.M.E. Moret, L.A. Raubeson, L.-S. Wang, T. Warnow, and S. Wyman. An empirical comparison of phylogenetic methods on chloroplast gene order data in Campanulaceae. In D.Sankoff and J. Nadeau, editors, *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment, and the Evolution of Gene Families*, pages 99–121. Kluwer Academic Publishers, Dordrecht NL, 2000.

[35] J.A. Cotton and R.D.M. Page. Rates and patterns of gene duplication and loss in the human genome. *Proceedings of the Royal Society of London. Series B*, 272:277–283, 2005.

[36] J.P. Demuth, T. De Bie, J. Stajich, N. Cristianini, and M.W. Hahn. The evolution of mammalian gene families. *PLoS ONE*, 1:e85, 2006.

[37] A. Doroftei and N. El-Mabrouk. Removing noice from gene trees. In *LNCS/LNBI*, WABI, 2011. submitted.

[38] J.-P. Doyon, C. Chauve, and S. Hamel. The space of gene tree/species tree reconciliations and parsimonious models. *Journal of Computational Biology*, 16:1399–1418, 2009.

[39] D. Durand, B.V. Haldórsson, and B. Vernot. A hybrid micro-macro-evolutionary approach to gene tree reconstruction. *Journal of Computational Biology*, 13:320–335, 2006.

[40] D. Durand and D. Sankoff. Testing for gene clusters. *Journal of Computational Biology*, 10:453–482, 2003.

[41] E.E. Eichler and D. Sankoff. Structural dynamics of eukaryotic chromosome evolution. *Science*, 301:793–797, 2003.

[42] N. El-Mabrouk and D. Sankoff. The reconstruction of doubled genomes. *SIAM Journal on Computing*, 32:754–792, 2003.

[43] O. Elemento, O. Gascuel, and M-P. Lefranc. Reconstructing the duplication history of tandemly repeated genes. *Molecular Biology and Evolution*, 19:278–288, 2002.

[44] O. Eulenstein, B. Mirkin, and M. Vingron. Duplication-based measures of difference between gene and species trees. *Journal of Computational Biology*, 5:135– 148, 1998.

[45] W.M. Fitch. Phylogenies constrained by cross-over process as illustrated by human hemoglobins and a thirteen-cycle, eleven amino-acid repeat in human apolipoprotein A-I. *Genetics*, 86:623–644, 1977.

[46] Z. Fu, X. Chen, V. Vacic, P. Nan, Y. Zhong, and T. Jiang. MSOAR: A high-throughput ortholog assignment system based on genome rearrangement. *Journal of Computational Biology*, 14:1160– 1175, 2007.

[47] D. Fulkerson and O. Gross. Incidence matricesand interval graphs. *Pac. J. Math.*, 15:835- 855, 1965.

[48] Y. Gagnon, O. Tremblay-Savard, D. Bertrand, and N. El-Mabrouk. Advances on genome duplication distances. In E. Tannier, editor, *Comparative Genomics (RECOMB CG '10)*, volume 6398 of *Lecture Notes in Computer Science*, pages 25–38, 2010.

[49] O. Gascuel, D. Bertrand, and O. Elemento. Reconstructing the duplication history of tandemly repeated sequences. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*, pages 205–235. Oxford, 2005.

[50] H. Gavranović and E. Tannier. Guided genome halving: probably optimal solutions provide good insights into the preduplication ancestral

genome of Saccharomyces cerevisiae. In *Pacific Symposium on Biocomputing*, volume 15, pages 21 – 30, 2010.

[51] G. Glusman, I. Yanai, I. Rubin, and D. Lancet. The complete human olfactory subgenome. *Genome Research*, 11:685–702, 2001.

[52] M. Goodman, J. Czelusniak, G.W. Moore, A.E. Romero-Herrera, and G. Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, 28:132–163, 1979.

[53] J.L. Gordon, K.P. Byrne, and K.H. Wolfe. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern Saccharomyces cerevisiae genome. *PloS Genetics*, 5:e1000485, 2009.

[54] P. Gorecki and J. Tiuryn. DLS-trees: a model of evolutionary scenarios. *Theoretical Computer Science*, 359:378–399, 2006.

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

[56] M.W. Hahn. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biology*, 8(R141), 2007.

[57] M.W. Hahn, M.V. Han, and S.-G. Han. Gene family evolution across 12 *drosophilia* genomes. *PLoS Genetics*, 3:e197, 2007.

[58] M. T. Hallett and J. Lagergren. Efficient algorithms for lateral gene transfer problems. In *Proceedings of the Fifth Annual International Conference on Computational Biology (RECOMB '01)*, pages 149–156, New York, 2001. ACM.

[59] M.T. Hallett and J. Lagergren. New algorithms for the duplication-loss model. In R. Shamir, S. Miyano, S. Istrail, P. Pevzner, and M. S. Waterman, editors, *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, RECOMB, pages 138–146, New York, 2000. ACM.

[60] S. Hannenhalli. Polynomial-time algorithm for computing translocation distance between genomes. In Z. Galil and E. Ukkonen, editors, *Combinatorial Pattern Matching. 6th Annual Symposium*, volume 937 of *Lecture Notes in Computer Science*, pages 162–176. Springer, 1995.

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

[62] S. Hannenhalli and P. A. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *Journal of the ACM*, 48:1–27, 1999.

[63] T. Hartman. A simpler 1.5-approximation algorithm for sorting by transpositions. In R.Baeza-Yates, E. Chávez, and M. Crochemore, editors, *Combinatorial Pattern Matching. 14th Annual Symposium.*, volume 2676 of *Lecture Notes in Computer Science*, pages 156–169, 2003.

[64] S. Heber and J. Stoye. Finding all common intervals of k permutations. In A. Amir and G. M. Landau, editors, *Combinatorial Pattern Matching. 12th Annual Symposium*, volume 2089 of *Lecture Notes in Computer Science*, pages 207– 218. Springer, 2001.

[65] R. Hoberman and D. Durand. The incompatible desiderata of gene cluster properties. In Aoife McLysaght and Daniel Huson, editors, *Comparative Genomics*, volume 3678 of *Lecture Notes in Computer Science*, pages 73–87. Springer Berlin / Heidelberg, 2005.

[66] R. Hoberman, D. Sankoff, and D. Durand. The statistical analysis of spatially clustered genes under the maximum gap criterion. *Journal of Computational Biology*, 12:1083–1102, 2005.

[67] HPCwire. Grappa runs in a record time. 9:47, November 23 2000.

[68] T. Jiang. Some algorithmic challenges in genome-wide ortholog assignment. *Journal of Computer Science and Technology*, 25, 2010.

[69] W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 100:11484–11489, 2003.

[70] M. Lajoie, D. Bertrand, and N. El-Mabrouk. Inferring the evolutionary history of gene clusters from phylogenetic and gene order data. *Molecular Biology and Evolution*, 27:761-772, 2009.

[71] M. Lajoie, D. Bertrand, N. El-Mabrouk, and O. Gascuel. Duplication and inversion history of a tandemly repeated genes family. *Journal of Computational Biology*, 14(4):462-478, 2007.

[72] G.M. Landau, L. Parida, and O. Weimann. Gene proximity analysis across whole genomes via PQ trees. *Journal of Computational Biology*, 12:1289– 1306, 2005.

[73] R.S. LaRue, S.R. Jonsson, K.A.T. Silverstein, M. Lajoie, D. Bertrand, N. El-Mabrouk, I. Hötzel, V. Andresdottir, T.P.L. Smith, and R.S. Harris. The artiodactyl APOBEC3 innate immune repertoire shows evidence for a multi-functional domain organization that existed in the ancestor of placental mammals. *BMC Molecular Biology*, 9:104, 2008.

[74] W.H. Li, Z. Gu, H. Wang, and A. Nekrutenko. Evolutionary analysis of the human genome. *Nature*, 409:847–849, 2001.

[75] M. Lynch and J.S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290:1151-1155, 2000.

[76] E. Lyons and M. Freeling. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant Journal*, 53:661–673, 2008.

[77] B. Ma, M. Li, and L. Zhang. From gene trees to species trees. *SIAM Journal on Computing*, 30:729–752, 2000.

[78] J. Ma, L. Zhang, B.B. Suh, B.J. Raney, R.C. Burhans, W.J. Kent, M. Blanchette, D. Haussler, and W. Miller. Reconstructing contiguous regions of an ancestral genome. *Genome Research*, 16:1557 – 1565, 2007.

[79] J. Meidanis, M. E. Walter, and Z. Dias. Transposition distance between a permutation and its reverse. In R. Baeza-Yates, editor, *Proceedings of the Fourth South American Workshop on String Processing (WSP '97)*, pages 70–79. Carleton University Press, 1997.

[80] J. Mixtacki. Genome halving under DCJ revisited. In X. Hu and J. Wang, editors, *Computing and Combinatorics (COCOON). Seventeenth Annual Conference*, volume 5092 of *Lecture Notes in Computer Science*, pages 276–286. Springer, 2008.

[81] B. Moret, L. Wang, T. Warnow, and S. Wyman. New approaches for re-constructing phylogenies from gene order data. *Bioinformatics*, 17:S165–S173, 2001.

[82] F. Murat, J.H. Xu, E. Tannier, M. Abrouk, N. Guilhot, C. Pont, J. Mess-ing, and J. Salse. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Research*, 2010.

[83] J.H. Nadeau and B.A. Taylor. Lengths of chromosomal segments con-served since divegence of man and mouse. *Proceedings of the National Academy of Sciences USA*, 81:814– 818, 1984.

[84] K.P. O'Brien, M. Remm, and E.L.L. Sonnhammer. Inparanoid: a com-prehensive database of eukaryotic orthologs. *Nucleic Acids Research*, 33:D476– D480, 2005.

[85] S. Ohno. *Evolution by gene duplication.* Springer, Berlin, 1970.

[86] R.D.M. Page. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, 43:58–77, 1994.

[87] R.D.M. Page. Genetree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, 14:819–820, 1998.

[88] R.D.M Page and M.A. Charleston. Reconciled trees and incongruent gene and species trees. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 37:57–70, 1997.

[89] R.D.M. Page and J. Cotton. Vertebrate phylogenomics: reconciled trees and gene duplications. In *Pacific Symposium on Biocomputing*, pages 536–547, 2002.

[90] Q. Peng, M.A. Alekseyev, G. Tesler, and P.A. Pevzner. Decoding synteny blocks and large-scale duplications in mammalian and plant genomes. In S.L. Salzberg and T. Warnow, editors, *Algorithms in Bioinformatics*, volume 5724 of *Lecture Notes in Computer Science*, pages 220–232, 2009.

[91] P.A. Pevzner and G. Tesler. Genome rearrangements in mammalian evolution: Lessons from human and mouse genomic sequences. *Genome Research*, 13:13 – 26, 2003.

[92] S.K. Pham and P.A. Pevzner. Drimm-synteny: decomposing genomes into evolutionary conserved segments. *Bioinformatics*, 26(20):2509-2516, 2010.

[93] J. Salse, S. Bolot, M. Throude, V. Jouffe, B. Piegu, U.M. Quraishi, T. Calcagno, R. Cooke, M. Delseny, and C. Feuillet. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *The Plant Cell*, 20:11 – 24, 2008.

[94] M.J. Sanderson and M.M. McMahon. Inferring angiosperm phylogeny from EST data with widespread gene duplication. *BMC Evolutionary Biology*, 7:S3, 2007.

[95] D. Sankoff. Genome rearrangements with gene families. *Bioinformatics*, 15:909–917, 1999.

[96] D. Sankoff. Gene and genome duplication. *Current Opinion in Genetics & Development*, 11:681–684, 2001.

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

[98] D. Sankoff and M. Blanchette. Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology*, 5:555–570, 1998.

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

[100] D. Sankoff, M.N. Parent, and D. Bryant. Accuracy and robustness of analyses based on numbers of genes in observed segments. In D. Sankoff and J. H. Nadeau, editors, *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families*, pages 299–306. Kluwer Academic, 2000.

[101] D. Sankoff, C. Zheng, P.K. Wall, C. dePamphilis, J. Leebens-Mack, and V.A. Albert. Towards improved reconstruction of ancestral gene order in

angiosperm phylogeny. *Journal of Computational Biology*, 16:1353–67, 2009.

[102] S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller. Human-mouse alignments with blastz. *Genome Research*, 13:103– 107, 2003.

[103] M. Shannon, A.T. Hamilton, L. Gordon, E. Branscomb, and L. Stubbs. Differential expansion of zinc finger transcription factor loci in homologous human and mouse gene clusters. *Genome Research*, 13:1097–1110, 2003.

[104] G. Shi and T. Jiang. MSOAR 2.0: Incorporating tandem duplications into ortholog assignment based on genome rearrangement. *BMC Bioinformatics*, 11:1160– 1175, 2010.

[105] V. Shoja and L. Zhang. A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. *Molecular Biology and Evolution*, 23:2134–2141, 2006.

[106] A.C. Siepel. Exact algorithms for the reversal median problem. Master's thesis, University of New Mexico, 2001.

[107] D.E. Soltis, V.A. Albert, J. Leebens-Mack, C.D. Bell, A.H. Paterson, C. Zheng, D. Sankoff, C.W. dePamphilis, P.K Wall, and P.S. Soltis. Polyploidy and angiosperm diversification. *American Journal of Botany*, 96:336 – 348, 2009.

[108] G. Song, L. Zhang, T. Vinar, and W. Miller. Inferring the recent duplication history of a gene cluster. In F.D. Ciccarelli and I. Miklós, editors, *Comparative Genomics*, volume 5817 of *Lecture Notes in Computer Science*. Springer, 2009.

[109] M. Tang, M.S. Waterman, and S. Yooseph. Zinc finger gene clusters and tandem gene duplication. In *Research in Molecular Biology (RECOMB 2001)*, pages 297–304, 2001.

[110] E. Tannier, C. Zheng, and D. Sankoff. Multichromosomal median and halving problems under different genomic distances. *BMC Bioinformatics*, 10, 2009.

[111] R.L. Tatusov, M.Y. Galperin, D.A. Natale, and E.V. Koonin. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28:33- 36, 2000.

[112] G. Tesler. Efficient algorithms for multichromosomal genome rearrangements. *Journal of Computer and System Sciences*, 65(3):587–609, 2002.

[113] T. Uno and M. Yagiura. Fast algorithms to enumerate all common intervals of two permutations. *Algorithmica*, 26:290– 309, 2000.

[114] B. Vernot, M. Stolzer, A. Goldman, and D. Durand. Reconciliation with non-binary species trees. *Journal of Computational Biology*, 15:981–1006, 2008.

[115] T. Vinař, B. Brejová, G. Song, and A. Siepel. Reconstructing histories of complex gene clusters on a phylogeny. *Journal of Computational Biology*, 17:1267–1269, 2010.

[116] M. E. Walter, Z. Dias, and J. Meidanis. Reversal and transposition distance of linear chromosomes. In *Proceedings of String Processing and Information Retrieval: A South American Symposium (SPIRE '98)*, pages 96–102, 1998.

[117] I. Wapinski, A. Pfeffer, N. Friedman, and A. Regev. Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449:54–61, 2007.

[118] R. Warren and D. Sankoff. Genome halving with double cut and join. *Journal of Bioinformatics and Computational Biology*, 7:357–371, 2009.

[119] R. Warren and D. Sankoff. Genome aliquoting revisited. In E. Tannier, editor, *Comparative Genomics (RECOMB CG). Eighth Annual Workshop*, volume 6398 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2010.

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

[121] K.H. Wolfe. Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics*, 2:333– 341, 2001.

[122] A.W. Xu. A fast and exact algorithm for the median of three problem: a graph decomposition approach. *Journal of Computational Biology*, 16:1369–1381, 2009.

[123] X. Xu and D. Sankoff. Tests for gene clusters satisfying the generalized adjacency criterion. In Ana Bazzan, Mark Craven, and Natlia Martins,

editors, *Advances in Bioinformatics and Computational Biology*, volume 5167 of *Lecture Notes in Computer Science*, pages 152–160. Springer Berlin / Heidelberg, 2008.

[124] S. Yancopoulos, O. Attie, and R. Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21:3340 – 3346, 2005.

[125] S. Yancopoulos and R. Friedberg. DCJ path formulation for genome transformations which include insertions, deletions, and duplications. *Journal of Computational Biology*, 16:1311–1338, 2009.

[126] Z. Yang and D. Sankoff. Natural parameter values for generalized gene adjacency. *Journal of Computational Biology*, 17:1113–1128, 2010.

[127] L. Zhang, B. Ma, L. Wang, and Y. Xu. Greedy method for inferring tandem duplication history. *Bioinformatics*, 19:1497–1504, 2003.

[128] L.X. Zhang. On Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *Journal of Computational Biology*, 4:177–188., 1997.

[129] Y. Zhang, G. Song, C.H Hsu, and W. Miller. Simultaneous history reconstruction for complex gene clusters in multiple species. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 162–173, 2009.

[130] Y. Zhang, G. Song, T. Vinar, E.D. Green, A. Siepel, and W. Miller. Reconstructing the evolutionary history of complex human gene clusters. In M.Vingron and L.Wong, editors, *Research in Computational Molecular Biology. (RECOMB 2008)*, volume 4955 of *Lecture Notes in Computer Science*, pages 29–49. Springer, 2008.

[131] C. Zheng. Pathgroups, a dynamic data structure for genome reconstruction problems. *Bioinformatics*, 26:1587–1594, 2010.

[132] C. Zheng and D. Sankoff. On the Pathgroups approach to rapid small phylogeny. *BMC Bioinformatics*, 12:S4, 2011.

[133] C. Zheng, Q. Zhu, Z. Adam, and D. Sankoff. Guided genome halving: hardness, heuristics and the history of the Hemiascomycetes. *Bioinformatics*, 24:i96 – i104, 2008.

[134] C. Zheng, Q. Zhu, and D. Sankoff. Descendants of whole genome duplication within gene order phylogeny. *Journal of Computational Biology*, 15:947 – 964, 2008.

[135] Q. Zhu, Z. Adam, V. Choi, and D. Sankoff. Generalized gene adjacencies, graph bandwidth, and clusters in yeast evolution. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 6:213–220, 2009.

[136] C. M. Zmasek and S. R. Eddy. A simple algorithm to infer gene duplication and speciiation events on a gene tree. *Bioinformatics*, 17:821– 828, 2001.