

A Tractable Variant of the Single Cut or Join Distance with Duplicated Genes

Pedro Feijão¹, Aniket Mane², and Cedric Chauve²(✉)

¹ School of Computing Science, Simon Fraser University,
8888 University Drive, Burnaby, BC V5A 1S6, Canada
pfeijao@sfu.ca

² Department of Mathematics, Simon Fraser University,
8888 University Drive, Burnaby, BC V5A 1S6, Canada
{amane, cedric.chauve}@sfu.ca

Abstract. In this work, we introduce a variant of the Single Cut or Join distance that accounts for duplicated genes, in the context of directed evolution from an ancestral genome to a descendant genome where orthology relations between ancestral genes and their descendant are known. Our model includes two duplication mechanisms: single-gene tandem duplication and creation of single-gene circular chromosomes. We prove that in this model, computing the distance and a parsimonious evolutionary scenario in terms of SCJ and single-gene duplication events can be done in linear time. Simulations show that the inferred number of cuts and joins scales linearly with the true number of such events even at high rates of genome rearrangements and segmental duplications. We also show that the median problem is tractable for this distance.

1 Introduction

The analysis of genome evolution by genome rearrangements has been subject to much research in the field of computational biology. Various rearrangement mechanisms explaining these variations have been proposed and studied, leading to a large corpus of algorithmic results [1]. The *pairwise genome rearrangement distance* problem aims at finding a most parsimonious or most likely sequence of genome rearrangements, within a given evolutionary model, that transforms one given genome into another given genome, thus giving a possible evolutionary scenario between the two given genomes. This problem has numerous applications toward unraveling important evolutionary mechanisms; recent examples include the nature of evolutionary breakpoints in bacteria [2] or the differentiated evolutionary mode of sex chromosomes and autosomes in *Anopheles* mosquitoes [3].

A number of rearrangement models have been studied from an algorithmic point of view, among them the reversal model was one of the first for which it was shown that the distance problem is tractable [4]; we refer to [1] for a review of the rich literature in this field up to 2009. For most evolutionary models that do not consider gene duplication, computing a rearrangement distance is tractable; additionally, some models can account for unequal gene content due to

gene gain or loss, for example [6, 7]. However, when gene duplication is allowed as an evolutionary event, most rearrangement distance problems become NP-hard. For instance, whereas the distance between two genomes can be computed in linear time for genomes without duplicate genes under the Double-Cut and Join (DCJ) model, it becomes NP-hard to compute when duplicate genes are considered [9], even when the gene content in both genomes is the same [11]. So far limited tractability results for computing distances in the presence of duplicated genes exist only for simpler genome rearrangement models, such as the breakpoint (BP) distance and the Single Cut or Join (SCJ) distance [5]. Even in these simpler models, the general problem of computing a distance with duplicated genes is difficult [12–15], although exponential algorithms have been developed for some specific problems such as the Exemplar BP distance for example [16–19], as well as polynomial time algorithms for two extensions of the SCJ model that include large-scale duplications: the double distance [5], where duplicated genes occur through a whole genome duplication, and the SCJ and whole chromosome duplication problem [20].

In the present work, we consider a problem of rearrangement distance with duplicated genes that we prove to be tractable. Our evolutionary model is also an extension of the SCJ model, that includes single-gene duplications. In the problem we consider that one genome, say A , is duplication-free, while the other one, denoted by D , can contain duplicated genes. This setting is inspired by the recent development of algorithms that reconstruct ancestral gene orders along a given species phylogeny using reconciled gene trees that provide, for each gene family within the set of considered genomes, one-to-one or one-to-many orthology relations between each ancestral gene and its descendant gene(s), if any. This general framework was introduced by Sankoff and El-Mabrouk in [21] (see also [23]) and motivated the introduction of the EBD problem [22]; it was later implemented in the DeCo* family of algorithms [24] to reconstruct ancestral gene orders in a duplication-aware evolutionary model from data including extant gene orders and reconciled gene trees. In this context, the genome A represents an ancestral genome, reconstructed for example with DeCo*, the genome D represents a descendant of A , either extant or reconstructed too, and we are interested in computing a directed distance, from an ancestor to its descendant, where all members of a same gene family present in genome A are considered as distinguishable thanks to the information provided by the reconciled gene tree of this family. In the evolutionary model we consider, rearrangements are either Single Cuts or Single Joins, while duplications can only be single gene duplications, but of two different types, Tandem Duplications (TD) or Floating Duplications (FD) in which a new copy is introduced as a circular chromosome. We show that in this model the distance problem can be simply reduced to deciding, for each gene family with duplicates in D , the length of a tandem array of duplicates to introduce in A and we provide a polynomial time algorithm for this problem.

The remaining paper is organized as follows: in Sect. 2, we recall some basic definitions, describe our evolutionary model and the variant of the SCJ distance and its relation with existing problems. We also introduce the median problem

for this distance. In Sect. 3 we present our theoretical results, a closed equation for the SCJ distance with duplications, a linear time algorithm to find an optimal scenario and an algorithm for the median problem. Finally, we provide preliminary experimental results in Sect. 4.

2 Preliminaries

2.1 Definitions and Problem Statements

Genes and Genomes. A genome consists of a set of chromosomes, each being a linear or circular ordered set of oriented genes¹. In our examples, a circular chromosome is represented using round brackets (e.g. (a, \bar{b}, c)) while a linear chromosome is represented using square brackets (e.g. $[a, \bar{b}, c]$), where a gene b in reverse orientation is denoted by \bar{b} . Alternatively, a genome can be represented by a set of *gene extremity adjacencies*. In this representation, gene x is represented using a pair of gene extremities (x_t, x_h) , x_t denotes the tail of the gene x and x_h denotes its head, and an *adjacency* is a pair of gene extremities that are adjacent in a genome. For example (a, \bar{b}, c) is encoded by the set of adjacencies $\{a_h b_h, b_t c_t, c_h a_t\}$ and $[a, \bar{b}, c]$ by $\{a_h b_h, b_t c_t\}$.

We assume that a given gene a can have multiple copies in a genome, with its number of occurrences being called its *copy number*. A genome in which every gene has copy number 1 is a *trivial genome* [18]. In this context, a non-trivial genome sometimes cannot be represented unambiguously by a set of adjacencies unless we distinguish the copies of each gene, for example by denoting the copies of a gene a with copy number k by a^1, \dots, a^k . For example, the genome $(a^1, \bar{b}, c^1, a^2), [\bar{a}^3, d, c^2]$, with two duplicated genes of respective copy numbers 3 and 2, is represented by $\{a_h^1 b_h, b_t c_t^1, c_h^1 a_t^2, a_h^2 a_t^1, a_h^3 d_t, d_h c_t^2\}$. We call a *gene family* the set of all copies of a gene that is present in both considered genomes. A gene family is trivial if it has exactly one copy in both genomes. From now, we identify a genome with its multi-set of adjacencies.

Let A and D be the respective adjacency sets for the genomes $[a, b, \bar{c}, d]$ and $[a, b, \bar{c}, \bar{d}][b, \bar{c}]$. Then the *multiset difference* between the two sets is denoted by $A - D$ (similarly $D - A$). Thus, if $A = \{a_h b_t, b_h c_h, c_t d_t\}$ and $D = \{a_h b_t, b_h c_h, c_t d_h, b_h c_h\}$ then $A - D = \{c_t d_t\}$ whereas $D - A = \{c_t d_h, b_h c_h\}$.

Evolutionary Model. We consider two types of evolutionary events: genome rearrangements and duplications. Genome rearrangements are modeled by *Single Cut or Join* (SCJ) operations, that either delete an adjacency from a genome (a cut) or join a pair of gene extremities that are not adjacent to any other gene extremity (a join), thus forming an adjacency. For duplication events, we consider two types of duplications, both creating an extra copy of a single gene: *Tandem Duplications* (TD) and *Floating Duplications* (FD). A tandem duplication of an existing gene g introduces an extra copy of g , say g' by adding an

¹ We use the generic term “gene” here to identify a genomic locus.

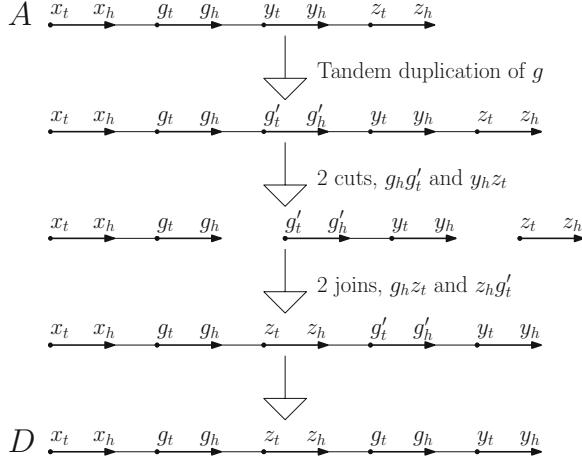


Fig. 1. In this example, a tandem duplicate of gene g is introduced. The adjacency $g_h y_t$ has been replaced by $g'_h y_t$ and an adjacency $g_h g'_t$ has been introduced. In this case the total number of operations to obtain D from A is 5. Note that the number of cut and join operations is dependent on the adjacencies of the gene g in A and D .

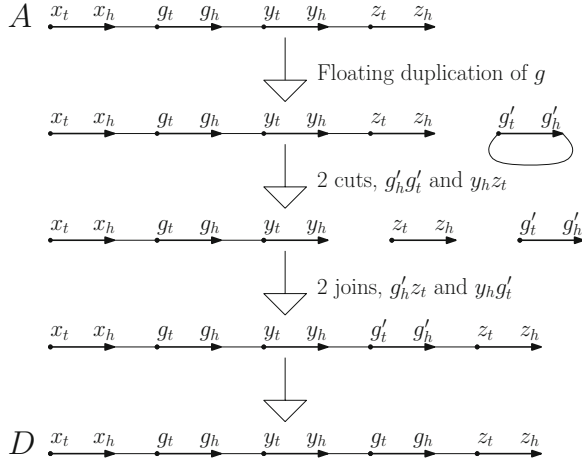


Fig. 2. In this example, a floating duplicate of gene g is introduced. An adjacency $g'_h g'_t$ has been added. In this case the total number of operations to obtain D from A is 5.

adjacency $g_h g'_t$, and, if there was an adjacency $g_h x$ by replacing it by the adjacency $g'_h x$, as shown in Fig. 1. A floating duplication introduces an extra copy g' of a gene g as a single-gene circular chromosome by adding the adjacency $g'_h g'_t$, as in Fig. 2. The motivation for this type of duplication is that gene insertions and gene deletions have been modeled with artificial circular chromosomes before, greatly simplifying how to deal with such type of operations. For instance, in the

Double-Cut and Join (DCJ) model, a deletion of a gene can be seen as a DCJ operation that applies two cuts to remove the given gene from a chromosome, followed by two joins to “repair” the broken chromosome and to circularize the deleted gene. A gene insertion is the inverse of this operation. This idea was effectively used in the DCJ indel model by Compeau [10]. We discuss the possibility of using a single-gene linear chromosome instead of a circular one at the end of Sect. 3.2.

The Pairwise Distance and Directed Median Problems. We consider the case of directed evolution from a trivial ancestral genome A to a descendant genome D . The evolutionary model excludes gene loss and de-novo gene creation, so we assume that every gene a in A has at least one descendant in D and conversely every gene D has a unique ancestor gene in A . If so, we say that A and D have the same *gene families set*.

The Directed SCJ-TD-FD (d-SCJ-TD-FD) Distance Problem. Let A be a trivial genome and D be a non-trivial genome, such that no gene family is absent from either A or D . Compute the minimum number of SCJ, TD and FD operations needed to transform A into D , denoted by $d_{\text{DSCJ}}(A, D)$.

Note that if D is a trivial genome, the usual SCJ distance, denoted by $d_{\text{SCJ}}(A, D)$ is defined by the symmetric differences of the adjacencies sets of A and D : $d_{\text{SCJ}}(A, D) = |A - D| + |D - A|$ where the first term accounts for the number of cuts and the second term for the number of joins.

We now turn to the directed median problem, that is the natural extension of the pairwise directed distance problem towards the small parsimony problem.

The Directed SCJ-TD-FD (d-SCJ-TD-FD) Median Problem. Let $k \geq 2$ and D_1, \dots, D_k (possibly) non-trivial genomes, such that no gene family is absent from any D_i . Compute a trivial genome A on the same set of gene families as the non-trivial genomes, that minimizes $\sum_{i=1}^k d_{\text{DSCJ}}(A, D_i)$.

2.2 Relation to the Exemplar Distance Framework

Sankoff [22] introduced the notion of Exemplar Breakpoint (EBP) distance, where an *exemplar* of a non-trivial genome is obtained by keeping exactly one gene copy from each gene family. In the directed evolution setting, an exemplar can be assumed to be the original gene from A having evolved into a gene now present in D , all other copies having been created by duplications. So the EBP distance problem aims to find an exemplar for each group of duplicates in D such that the trivial genome that results from deleting all non-exemplar copies minimizes the breakpoint distance to A . The notion of exemplar distance can naturally be used in conjunction with the SCJ distance instead of the BP distance, a problem we denote the ESCJ distance. The EBP distance problem has been shown to be NP-hard even in the directed evolution case where every duplicated gene has exactly two copies in D [12], and it is immediate to extend this hardness result to the directed ESCJ distance problem.

Intuitively, the directed ESCJ distance and the d-SCJ-FD-TD distance problems seem very similar. For example in the case of duplicated genes having exactly two copies in D , the later aims at deciding which copy in D is exemplar (*i.e.* evolved from the original copy in A) and then, for the second copy, if it originates from a TD or a FD, thus resulting in a matching between two genomes with two copies of each duplicate, opposed to the ESCJ setting where the matching is between genomes with one copy of each gene.

It is interesting to notice that both problems, although similar, have opposed properties in terms of tractability, and that the d-SCJ-TD-FD distance problem is tractable despite considering a larger solution space. Moreover, one can ask if there is a strong correlation between the distance obtained in both settings. It is not difficult to find examples that show that both distances can be quite different: the ESCJ distance between $A = [a, b, c, d]$ and $D = [a, c, b, d, c, a, d, b]$ is 0, whereas the SCJ-TD-FD distance between the same two genomes is 18 (4 duplications, 7 cuts, 7 joins). However, although the difference between both distances can be arbitrarily large, tight bounds can be derived.

Lemma 1. *Let A be a trivial genome, D be an arbitrary genome on the same set of gene families than A , $d_{DSCJ}(A, D)$ and $d_{ESCJ}(A, D)$ denote the d-SCJ-FD-TD and the ESCJ distances, respectively. Let k be the difference between the number of genes in D and the number of genes in A . The following bounds*

$$k \leq d_{DSCJ}(A, D) - d_{ESCJ}(A, D) \leq 5k$$

are tight.

Proof. First, we obtain the genome A' by applying $d_{ESCJ}(A, D)$ SCJ operations on A in a way that the duplicated genes in D are in the same order as the corresponding matched genes in A' , as given by an optimal exemplar matching. In the SCJ-FD-TD model, we need to apply at least k duplications on A' to obtain D , so $d_{DSCJ}(A, D) \geq d_{ESCJ}(A, D) + k$ (otherwise $d_{ESCJ}(A, D)$ would not be optimal). To show that this bound is tight, we can see that the trivial case of no duplications holds. But, whenever A and D differ by only k tandem duplications, the bound is tight, since in this case $d_{ESCJ}(A, D) = 0$ and $d_{DSCJ}(A, D) = k$.

Now, from A' , we can apply k free duplications, followed by k cuts on these duplications. Also, perform at most k cuts, between any two genes on A' if both have more than one copy on D . Since A' was ordered in relation to its corresponding copies on D , it is possible to join the “fragments” of A' that were created with the previous $2k$ cuts with $2k$ joins in a way to transform A' in D , and therefore we built a d-SCJ-FD-TD scenario from A to D with $d_{ESCJ}(A, D) + 5k$ operations. Any pair of circular genomes $A = (1, 2, \dots, n)$ and $D = (1, n, 2, 1, \dots, i, i - 1, \dots, n, n - 1)$ satisfies the tight bound. \square

3 Algorithmic Results

In this section, we show that, after a preprocessing step of removing obvious TD and FD in D , the d-SCJ-TD-FD distance can be calculated with the symmetric

difference between the adjacency (multi)sets of the input genomes, with an extra factor to account for the gene duplications. We first focus on the preprocessing. Next we describe a linear time algorithm to compute a parsimonious scenario and a polynomial time algorithm for the directed median problem.

3.1 The Directed SCJ-TD-FD Distance

An *observed duplication* in D is defined as an adjacency of the form $g_h g_t$, that defines either a single-gene circular chromosome or a *tandem array* of two (or more) copies of a gene g that occur consecutively and with the same orientation. We denote by t the number of such adjacencies in D and by D' the genome obtained from D by removing first all genes but one from each tandem arrays, and then all single-circular chromosomes for genes from non-trivial families but one if all genes of the family are in such circular chromosomes. D can obviously be obtained from D' by t duplications and the following lemma is immediate:

Lemma 2. $d_{DSCJ}(A, D) = d_{DSCJ}(A, D') + t$.

As a consequence, we assume from now on that D has been preprocessed as described above and does not contain any tandem array or any extra copy of a non-trivial family that is in a single-gene circular chromosome. We say that D is *reduced*. Note that single-gene linear chromosomes are not impacted by this preprocessing as, in our setting, if the considered gene is from a non-trivial family, the linear chromosome it forms required at least a cut to be created.

Theorem 1. *Given a trivial genome A and a reduced non-trivial genome D such that no gene family is absent from either A or D and where D has n_d more genes than A , the d -SCJ-TD-FD distance between A and D is given by*

$$d_{DSCJ}(A, D) = |A - D| + |D - A| + 2n_d.$$

Proof. First, we show that $d_{DSCJ}(A, D) \geq |A - D| + |D - A| + 2n_d$. To obtain D from A , we need exactly n_d gene duplications. Each duplication of a gene g will create the adjacency $g_h g_t$, regardless of the type of the duplication or the timing of the duplication event. Therefore, n_d adjacencies of the type $g_h g_t$ will have to be cut, as D is reduced and has no adjacency of this type. In addition, any adjacency in $A - D$ and $D - A$ defines an unavoidable cut or join respectively. Therefore, we can not transform A into D with less than $|A - D| + |D - A| + 2n_d$ operations.

Now, we show that $d_{DSCJ}(A, D) \leq |A - D| + |D - A| + 2n_d$, by induction on n_d . For the base case $n_d = 0$, the result follows immediately as both genomes are trivial and $d_{DSCJ}(A, D) = d_{SCJ}(A, D)$.

We now assume that $n_d > 0$, and pick a gene g with one copy in A and more than one copy in D . Depending on how the adjacencies of g are conserved or not in D , we have a few different subcases to consider. However, in each subcase the general strategy remains the same, as follows. We build a genome A_2 from A by applying one duplication (FD or TD) and also relabeling the original copy

g as g' , creating an adjacency $g_h g_t$ in the case of an FD or $g'_h g_t$ in the case of a TD. Then we build a genome D_2 from D by also relabeling one copy of g to g' , thus creating a new trivial gene family and an instance of the d-SCJ-TD-FD problem with exactly $n_d - 1$ duplicated gene copies. We can apply the induction hypothesis, leading to the inequality

$$d_{\text{DSCJ}}(A_2, D_2) \leq |A_2 - D_2| + |D_2 - A_2| + 2(n_d - 1).$$

Also, as D and D_2 are identical but for the relabeling of g , there is a scenario from A to D , going from A to A_2 and then to D , resulting in the upper bound

$$d_{\text{DSCJ}}(A, D) \leq d_{\text{DSCJ}}(A, A_2) + d_{\text{DSCJ}}(A_2, D_2) = 1 + d_{\text{DSCJ}}(A_2, D_2).$$

We will then show that we can build A_2 and D_2 in a way that they satisfy

$$|A - D| + |D - A| = |A_2 - D_2| + |D_2 - A_2| - 1,$$

where the -1 term is due to the extra $g_h g_t$ adjacency on A_2 created with the duplication. Together with the above inequalities this will lead to

$$d_{\text{DSCJ}}(A, D) \leq 1 + d_{\text{DSCJ}}(A_2, D_2) \leq |A - D| + |D - A| + 2n_d$$

and the result follows. To show that we can build A_2 and D_2 that satisfy the above conditions, we will consider three subcases.

Case (i): Assume that g is not a telomere (and so there are two adjacencies involving g in A , say xg_t and $g_h y$) and there is a copy of g in D whose extremities for also adjacencies xg_t and $g_h y$. We say that the context of g is *strongly conserved* between A and D . Note that x and y do not need to belong to trivial gene families and there might be several copies of x, y, g in D that conserve the context of g in A .

In this case, we build A_2 by applying an FD to create an extra copy of g and relabel the original copy of g in A as g' ; we also relabel g' an arbitrary copy of g in D that has the same context than g in A , to obtain D_2 (see Fig. 3). Comparing the adjacency sets of A and D with A_2 and D_2 , we can see that from A to A_2 two adjacencies were renamed from xg_t and $g_h y$ to xg'_t and $g'_h y$, and exactly the same change happened from D to D_2 . Also, the adjacency $g_h g_t$ was added in A_2 . As a result, $A_2 = A - \{xg_t, g_h y\} + \{xg'_t, g'_h y, g_h g_t\}$. Similarly, $D_2 = D - \{xg_t, g_h y\} + \{xg'_t, g'_h y\}$. Therefore, we have that $|A - D| + |D - A| = |A_2 - D_2| + |D_2 - A_2| - 1$. Note that this relabeling only works if we introduce an extra copy of g in A with an FD here; if instead we introduce it with a TD, it would not be possible to get adjacencies xg'_t and $g'_h y$ in D_2 , as the copy of g involved in both adjacencies would be different.

Case (ii): Assume that g is not a telomere in A , its context is not strongly conserved between A and D , but both adjacencies involving g , xg_t and $g_h y$, are present in D on different copies of g . We say that the context of g is *weakly conserved* between A and D . Again x and y need not to be trivial gene families and there might be several occurrences of adjacencies xg_t and $g_h y$ in D .

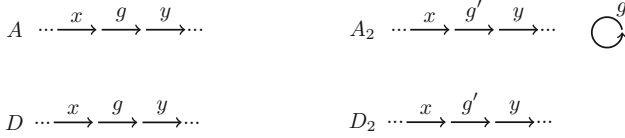


Fig. 3. The context of g is strongly conserved between A and D (Case (i)).

In this case, we build A_2 by applying a TD on g , relabeling the gene g that has the adjacency xg_t as a new gene g' in both A_2 and D_2 , as shown on Fig. 4. Comparing the adjacency sets of A and A_2 , we notice that the adjacency xg_t changes to xg'_t , and $g_h g_t$ is added. Thus, $A_2 = A - \{xg_t\} + \{xg'_t, g_h g_t\}$. From D to D_2 we also have the same change, and possibly one more, depending if g'_h is a telomere in D (no change) or if g'_h has an adjacency $g'_h w$. In the former case, $D_2 = D - \{xg_t\} + \{xg'_t\}$. Otherwise, $D_2 = D - \{xg_t, g_h w\} + \{xg'_t, g'_h w\}$. In either case, the possible adjacency $g'_h w$ does not exist in A or A_2 . Consequently, the equality $|A - D| + |D - A| = |A_2 - D_2| + |D_2 - A_2| - 1$ holds.

Note also that in this case an FD would not be optimal, because it would force the labeling of the adjacency $g_h y$ to $g'_h y$, and since the adjacency $g_h y$ on D cannot have the label $g'_h y$, this would force an extra pair of SCJ operations.

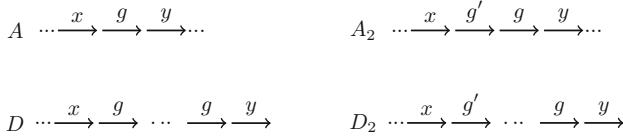


Fig. 4. The context of g is weakly conserved between A and D (Case (ii)).

Case (iii): We assume now that the context of g in A is neither strongly nor weakly conserved, and so at most one adjacency of g in A is also present in D .

This case is similar to case (i), if we assume that either xg_t or $g_h y$, are present in D , or neither. In the same way, we apply an FD on g , labeling the original copy as g' , as shown in Fig. 5. On D , we pick a gene g that has an adjacency xg_t or $g_h y$ if any or, if no adjacency involving g is conserved in D , we pick an arbitrary g , and relabel it as g' .

Now, any adjacencies that were conserved between A and D will remain conserved between A_2 and D_2 , and no new conserved adjacencies have been created. Since, as before, A_2 has a new $g_h g_t$ adjacency, the equality $|A - D| + |D - A| = |A_2 - D_2| + |D_2 - A_2| - 1$ holds.

These three cases cover all possible configurations for g , so the theorem is proved. \square

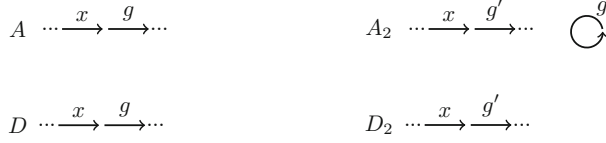


Fig. 5. At most one adjacency of g is conserved (Case (iii)).

3.2 Computing a Parsimonious Scenario

It follows from Lemma 2 and Theorem 1 that computing the d-SCJ-TD-FD distance can be done in linear time in the size of the considered genomes A and D . Moreover, they define a simple algorithm that computes a parsimonious scenario in terms of duplications, cuts and joins from A to D , described in Algorithm 1 below.

Algorithm 1. Compute an SCJ-TD-FD parsimonious scenario between a trivial genome A and a genome D

Reduce D into a reduced genome D'

Let $A' = A$ and $i = 1$

while (A', D') has a non trivial gene family **do**

Let g be an arbitrary gene from a non trivial family in A' ; relabel g by g^i .

if the context of g is strongly conserved **then**

relabel the corresponding copy of g in D' by g^i

add to A' a single-gene circular chromosome g .

else if the context of g is weakly conserved **then**

create an extra copy of g^i with a TD

relabel a copy of g involved in adjacency xg_t in D' by g^i .

else if one adjacency of g is conserved in D' **then**

relabel the corresponding copy of g in D' by g^i

add to A' a single-gene circular chromosome g^i .

else

relabel an arbitrary copy of g in D' by g^i

add to A' a single-gene circular chromosome g^i .

end if

$i = i + 1$

end while

Compute an SCJ scenario from A' to D' .

Recreate in D' , the tandem arrays and single-gene circular chromosomes removed when reducing D into D' .

Theorem 2. Given a trivial genome A with n_A genes and a possibly non-trivial genome D on the same set of gene families and with n_D genes, Algorithm 1 computes a parsimonious SCJ-TD-FD scenario that transforms A into D and can be implemented to run in time and space $O(n_D)$.

The correctness of the algorithm follows immediately from the fact that it implements exactly the rules described to compute the SCJ-TD-FD distance (Lemma 2 and Theorem 1). The linear time and space complexity follows from the fact that these rules are purely local and ask only to check for the conservation of adjacencies in both considered genomes.

Every iteration of the while loop in Algorithm 1 takes place only if there is a non-trivial gene family left in D' . The maximum number of iterations is the number of duplicate genes, $n_d = n_D - n_A$ which is $O(n_D)$ when $n_D \geq n_A$. In each iteration, we check if the context of the chosen gene g is strongly conserved, weakly conserved or not conserved. This involves trying to match the adjacencies of g in A with those in the adjacency set of D' that involve a copy of g . This can be done in constant time, with a linear time preprocessing of the data. Hence, the worst-case time complexity is $O(n_D)$.

Remark 1. We have discussed in Sect. 2.1 the rationale to create duplicate genes with a FD creating a circular single-gene chromosome. However if the evolutionary model of the FD event created a linear single-gene chromosome, this would introduce a dissymmetry between TD and FD (namely no adjacency is created with an FD), while in our model each created copy induces a cost of two due to the necessary break of the created adjacency required in the process of obtaining the reduced genome D' . We conjecture that the use of linear chromosomes would affect the choice of duplication event (FD or TD) only when the context is not conserved, which would result in a more complicated distance formula.

3.3 The Directed Median Problem

Let us remind that under the SCJ-TD-FD evolutionary model, the *directed median problem* asks, given k non-trivial genomes $D_1, \dots, D_k, k \geq 2$, with the same gene families, to find a trivial common ancestor A , such that $\sum_{i=1}^k d_{\text{DSCJ}}(A, D_i)$ is minimized.

We first assume that the genomes D_1, \dots, D_k are reduced. We define the *score* $s(A)$ of a genome A as

$$s(A) = \sum_{i=1}^k d_{\text{DSCJ}}(A, D_i) = \sum_{i=1}^k (|A - D_i| + |D_i - A| + 2n_{d_i})$$

where n_{d_i} is the number of extra gene copies in D_i compared to A , for $i = 1, \dots, k$. Using the fact that $|A - D| + |D - A| = |A| + |D| - 2|A \cap D|$ we derive

$$s(A) = N_d - \left(2 \sum_{i=1}^k |A \cap D_i| - k|A| \right)$$

where $N_d = \sum_{i=1}^k (2n_{d_i} + |D_i|)$, and does not depend from A . Therefore, minimizing $s(A)$ is equivalent to maximizing $2 \sum_{i=1}^k |A \cap D_i| - k|A|$.

For a given adjacency a , let $\delta_i(a)$ be 1 if $a \in D_i$, and 0 otherwise. The score of a genome with a single adjacency a is $s(\{a\}) = N_d - \left(2 \sum_{i=1}^k \delta_i(a) - k \right)$.

This motivates the following approach, similar to the breakpoint median algorithm of [8]. Build a graph G where the vertices are defined as the extremities (head and tail) of a unique copy for each gene family in the considered genomes D_i (so a gene family a induces two vertices a_h and a_t), and weighted edges are defined as follows: for any edge $e = (x, y)$ such that x and y form an adjacency in at least one of the genomes D_i , the weight of e is $w(e) = 2 \sum_{i=1}^k \delta_i(e) - k$. Any matching M on G defines a trivial genome A_M , having the adjacencies corresponding to the edges in the matching M . Also, if $W(M)$ denotes the weight of the matching M , that is the sum of the weights of the edges in M we have that

$$\begin{aligned} s(A_M) &= N_d - \left(2 \sum_{i=1}^k |A_M \cap D_i| - k|A_M| \right) \\ &= N_d - \sum_{e \in M} \left(2 \sum_{i=1}^k \delta_i(e) - k \right) \\ &= N_d - W(M) \end{aligned}$$

Therefore, solving a maximum weight matching problem on G solves the directed median problem. To handle the case when some D_i is not reduced, we can rely on Lemma 2 that implies that the genomes can be reduced first without impacting the optimality of a trivial genome obtained by a maximum weight matching. Combined with the tractability of computing a maximum weight matching [25], this proves our last theorem.

Theorem 3. *Let $k \geq 2$ and D_1, \dots, D_k be k genomes on the same set of n gene families, having respectively n_1, \dots, n_k adjacencies. The directed SCJ-TD-FD median problem for these genomes can be solved in time and space $O(n(n_1 + \dots + n_k) \log(n_1 + \dots + n_k))$.*

Remark 2. In the case of the median of two genomes D_1 and D_2 , note that the only edges with strictly positive weight in the graph are defined by adjacencies that appear in both D_1 and D_2 , while edges appearing just once have weight 0. So a median genome can be defined as a maximum matching over the unweighted graph defined only by adjacencies that appear in both genomes, and given such a median, it can be augmented by any subset of edges appearing just once that do not re-use a gene extremity already used in the matching.

4 Experimental Results

We ran experiments on simulated instances with the aim to evaluate the ability of the d-SCJ-TD-FD distance to correlate with the true number of syntenic events. We followed a simulation protocol inspired from [16]. The code itself was programmed in Python and is available via github². We first describe the simulation protocol, followed by the results we obtained.

² <https://github.com/acme92/SCJTDFD>.

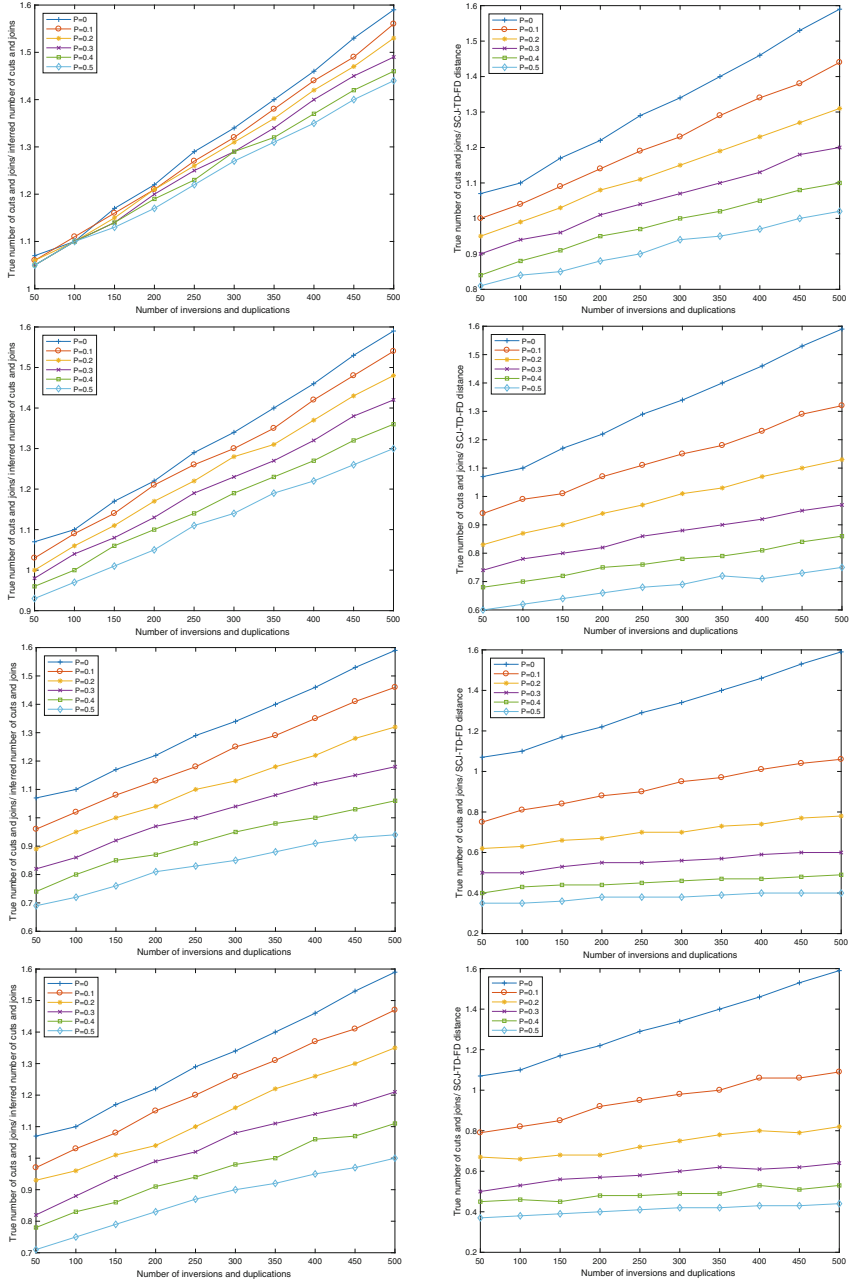


Fig. 6. Experimental results, for four duplications parameters – single-gene segmental duplication (top row), two-genes segmental duplication (second row), five-genes segmental duplications (third row), variable length segmental duplications (bottom row) – and two measured quantities – inferred cuts and joins (left column) and SCJ-TD-FD distance (right column).

We started from a genome A composed of a single linear chromosome containing 1000 single-copy genes. Then, we transformed A genome into a genome D through a sequence of random segmental duplications and inversions. We fixed the number N of evolutionary events (from 50 to 500 by steps of 50) and the probability P that a given event is a segmental duplication (from 0 to 0.5 by steps of 0.1). A segmental duplication is defined by three parameters: the position of the first gene of the duplicated segment, the length of the duplicated segment, and the breakpoint where the duplicated segment is transposed into; we considered two models of segmental duplications, one with fixed segment length L (with L taking values in $\{1, 2, 5\}$) and one where for each segment, L is picked randomly (under the uniform distribution) in $\{1, 2, 5, 10\}$. Inversion breakpoints were chosen randomly, again under the uniform distribution. For each array of parameters, we ran 50 replicates.

For each instance, we compared two quantities to the true number of cuts and joins in the scenario transforming A into D , which is roughly four times the number of inversions plus three times the number of segmental duplications: first we compared the full SCJ-TD-FD distance, defined as stated in Theorem 1 and the number of cuts and joins ($|A - D| + |D - A|$). Figure 6 illustrates the results we obtained.

We can make several observations from these results. The first one is a general trend that both measured quantities (the number of cuts and joins and the full SCJ-TD-FD distances) scale linearly with the true number of cuts and joins. The second observation is that, as expected, the slope and a y -intercept of the graphs depend from both the frequency of duplications and the length of the duplicated segments. This leaves open the question of using the SCJ-TD-FD distance as an estimator of the number of cuts and joins in an evolutionary model where the probability of duplication compared to rearrangements (that can be estimated for example from reconciled gene trees and adjacency forests [24]) is given and the length of duplicated segments is expected to follow a well defined distribution.

5 Conclusion

In this work, we introduced a simple variant of the SCJ model that accounts for duplications, and showed that, in this model, computing a directed parsimonious genomic distance from a trivial ancestral genome to a non-trivial descendant genome can be done in linear time and that a directed median can be computed in polynomial time. The tractability stems mostly from the combination of assuming that one genome is trivial and of a simplified model of duplication where gene duplication are single-gene events. However we believe it is interesting to push the tractability boundaries of the SCJ models toward augmented models of evolution (here accounting for duplications). Moreover, our work is motivated by the increasing performance of ancestral gene order reconstruction methods, that can now account for complex gene histories using reconciled gene trees and motivate the directed distance approach, and provides an additional positive result along the line of the research program outlined in [21].

For example, our algorithm will allow to extend the small parsimony algorithm PhySca introduced in [26] to a duplication-aware framework by allowing to score exactly and quickly an ancestral gene order configuration within a species phylogeny (work in progress).

There are several avenues to extend the results we presented in this paper. It will likely be easy to modify our algorithm to work in an extended the evolution model to integrate the loss of gene families and de-novo creation of genes. Our main result provides a simple algorithm that computes a parsimonious scenario, however it is likely one among a large number of parsimonious scenarios, and it is open to see if the results of [27] about counting and sampling SCJ parsimonious scenarios can be extended to our model. An important open question toward a more realistic model of evolution concerns the possibility to include larger scale duplications as unit-cost events. The case of a single whole genome duplication and of whole-chromosome duplications have been shown to be tractable [5, 20], but to the best of our knowledge there is no known result including segmental duplications in which a contiguous segment of genes is duplicated either in tandem or appearing as a single chromosome. It also remains to be seen if the directed SCJ-TD-FD distance can be used toward the computation of an estimated distance in a more realistic evolutionary distance, similarly to the use of the breakpoint distance to estimate the true DCJ distance [28]; our experimental results suggest this is a promising avenue, although it might be difficult to obtain analytical results in models mixing rearrangements and duplications. Finally, the question of the tractability of the small parsimony problem in our model is also, to the best of our knowledge, still open. It is known to be tractable in the pure SCJ model (*i.e.* with no duplications) due to the independence of adjacencies; this assumption does not hold anymore here and the small parsimony problem is thus likely more difficult in our model. Our tractability result for the directed median is a first step toward this goal as it already provides a building block for a bottom-up ancestral reconstruction algorithm.

Acknowledgments. CC is supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada. PF is supported by the Genome Canada grant PathoGiST.

References

1. Fertin, G., Labarre, A., Rusu, I., Tannier, E., Vialette, S.: *Combinatorics of Genome Rearrangements*. MIT Press, Cambridge (2009)
2. Wang, D., Li, D., Ning, K., Wang, L.: Core-genome scaffold comparison reveals the prevalence that inversion events are associated with pairs of inverted repeats. *BMC Genom.* **18**(1), 268 (2017)
3. Neafsey, D., Waterhouse, R., et al.: Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science* **347**(6217), 1258522 (2015)
4. Hannenhalli, S., Pevzner, P.: Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In: 27th Annual ACM Symposium on the Theory of Computing (STOC 1995), pp. 178–189 (1995)

5. Feijão, P., Meidanis, J.: SCJ: a breakpoint-like distance that simplifies several rearrangement problems. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8**(5), 1318–1329 (2011)
6. da Silva, P., Machado, R., Dantas, S., Braga, M.: DCJ-indel and DCJ-substitution distances with distinct operation costs. *Algorithms Mol. Biol.* **8**(1), 21 (2013)
7. Braga, M., Willing, E., Stoye, J.: Double cut and join with insertions and deletions. *J. Comput. Biol.* **18**(9), 1167–1184 (2011)
8. Tannier, E., Zheng, C., Sankoff, D.: Multichromosomal median and halving problems under various different genomic distances. *BMC Bioinform.* **10**, 120 (2009)
9. Shao, M., Lin, Y., Moret, B.: An exact algorithm to compute the double-cut-and-join distance for genomes with duplicate genes. *J. Comput. Biol.* **22**(5), 425–435 (2015)
10. Compeau, P.E.C.: DCJ-Indel sorting revisited. *Algorithms Mol. Biol.* **8**, 6 (2013)
11. Rubert, D., Feijão, P., Braga, M., Stoye, J., Martinez, F.: Approximating the DCJ distance of balanced genomes in linear time. *Algorithms Mol. Biol.* **12**, 3 (2017)
12. Bryant, D.: The complexity of calculating exemplar distances. In: Sankoff, D., Nadeau, J.H. (eds.) *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families*, vol. 1, pp. 207–211. Springer, Dordrecht (2000). doi:[10.1007/978-94-011-4309-7_19](https://doi.org/10.1007/978-94-011-4309-7_19)
13. Blin, G., Chauve, C., Fertin, G.: The breakpoint distance for signed sequences. In: *Algorithms and Computational Methods for Biochemical and Evolutionary Networks (CompBioNets 2004)*. Text in Algorithms, vol. 3, pp. 3–16 (2004)
14. Angibaud, S., Fertin, G., Rusu, I., Thevenin, A., Vialette, S.: On the approximability of comparing genomes with duplicates. *J. Graph Algorithms Appl.* **13**(1), 19–53 (2009)
15. Blin, G., Fertin, G., Sikora, F., Vialette, S.: The EXEMPLARBREAKPOINTDISTANCE for non-trivial genomes cannot be approximated. In: Das, S., Uehara, R. (eds.) *WALCOM 2009. LNCS*, vol. 5431, pp. 357–368. Springer, Heidelberg (2009). doi:[10.1007/978-3-642-00202-1_31](https://doi.org/10.1007/978-3-642-00202-1_31)
16. Shao, M., Moret, B.: A fast and exact algorithm for the exemplar breakpoint distance. *J. Comput. Biol.* **23**(5), 337–346 (2016)
17. Shao, M., Moret, B.: On computing breakpoint distances for genomes with duplicate genes. *J. Comput. Biol.* (2016, ahead of print). doi:[10.1089/cmb.2016.0149](https://doi.org/10.1089/cmb.2016.0149)
18. Wei, Z., Zhu, D., Wang, L.: A dynamic programming algorithm for (1,2)-exemplar breakpoint distance. *J. Comput. Biol.* **22**(7), 666–676 (2014)
19. Angibaud, S., Fertin, G., Rusu, I., Thévenin, A., Vialette, S.: Efficient tools for computing the number of breakpoints and the number of adjacencies between two genomes with duplicate genes. *J. Comput. Biol.* **15**(8), 1093–1115 (2008)
20. Zeira, R., Shamir, R.: Sorting by cuts, joins, and whole chromosome duplications. *J. Comput. Biol.* **24**(2), 127–137 (2017)
21. Sankoff, D., El-Mabrouk, N.: Duplication, rearrangement and reconciliation. In: Sankoff, D., Nadeau, J.H. (eds.) *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics Map, Alignment and Evolution of Gene Families*, vol. 1, pp. 537–550. Springer, Dordrecht (2000). doi:[10.1007/978-94-011-4309-7_46](https://doi.org/10.1007/978-94-011-4309-7_46)
22. Sankoff, D.: Genome rearrangement with gene families. *Bioinformatics* **15**(11), 909–917 (1999)
23. Chauve, C., El-Mabrouk, N., Guéguen, L., Semeria, M., Tannier, E.: Duplication, rearrangement and reconciliation a follow-up 13 years later. In: Chauve, C., El-Mabrouk, N., Tannier, E. (eds.) *Models and Algorithms for Genome Evolution*, vol. 19, pp. 47–62. Springer, London (2013). doi:[10.1007/978-1-4471-5298-9_4](https://doi.org/10.1007/978-1-4471-5298-9_4)

24. Duchemin, W., Anselmetti, Y., Patterson, M., Ponty, Y., Bérard, S., Chauve, C., Scornavacca, C., Daubin, V., Tannier, E.: DeCoSTAR: reconstructing the ancestral organization of genes or genomes using reconciled phylogenies. *Genome Biol. Evol.* **9**(5), 1312–1319 (2017)
25. Plummer, M.D., Lovász, L.: *Matching Theory*. Elsevier, Amsterdam (1986)
26. Luhmann N., Lafond M., Thevenin A., Ouangraoua A., Wittler R., Chauve C.: The SCJ small parsimony problem for weighted gene adjacencies. *IEEE/ACM Trans. Comput. Biol. Bioinform.* (2017, ahead of print). doi:[10.1109/TCBB.2017.2661761](https://doi.org/10.1109/TCBB.2017.2661761)
27. Miklós, I., Kiss, S., Tannier, E.: Counting and sampling SCJ small parsimony solutions. *Theoret. Comput. Sci.* **552**, 83–98 (2014)
28. Biller, P., Guéguen, L., Tannier, E.: Moments of genome evolution by Double Cut-and-Join. *BMC Bioinform.* **16**(Suppl 14), S7 (2015)

Comparative Genomics

15th International Workshop, RECOMB CG 2017,
Barcelona, Spain, October 4-6, 2017, Proceedings

Meidanis, J.; Nakhleh, L. (Eds.)

2017, VIII, 321 p. 88 illus., Softcover

ISBN: 978-3-319-67978-5