Tandem Halving Problems by DCJ

Antoine Thomas, Aïda Ouangraoua, Jean-Stéphane Varré

Tandem halving problems by DCJ

Antoine Thomas, Aïda Ouangraoua, and Jean-Stéphane Varré

LIFL, UMR 8022 CNRS, Université Lille 1
INRIA Lille, Villeneuve d'Ascq, France

Abstract. We address the problem of reconstructing a non-duplicated ancestor to a partially duplicated genome in a model where duplicated content is caused by several tandem duplications throughout its evolution and the only allowed rearrangement operations are DCJ. As a starting point, we consider a variant of the Genome Halving Problem, aiming at reconstructing a tandem duplicated genome instead of the traditional perfectly duplicated genome. We provide a distance in $\mathcal{O}(n)$ time and a scenario in $\mathcal{O}(n^2)$ time. In an attempt to enhance our model, we consider several problems related to multiple tandem reconstruction. Unfortunately we show that although the problem of reconstructing a single tandem can be solved polynomially, it is already NP-hard for 2 tandems.

1 Introduction

Studying genome architecture is of great importance. There are many applications from evolution to cancer genomics. Thanks to the growing number of sequencing projects, one has a lot of data for comparing genomes both between species but also variants within a same species. Inspection of genomes revealed a lot of duplication events during the course of evolution. It is well-known that whole genome duplications arise several times, notably among mammals. But segmental duplications also occur. Recent studies between several plant mitochondrial genomes observe that some genes are duplicated [5, 6, 4]. A hypothesis to the creation of such duplications is that tandem duplications occurred followed by other rearrangements that scrambled the duplicates. In this paper we study methods to analyse such genomes. More precisely we are interested in reconstructing a non-duplicated ancestral genome from a partially duplicated genome. Figure 1 illustrates the problem.

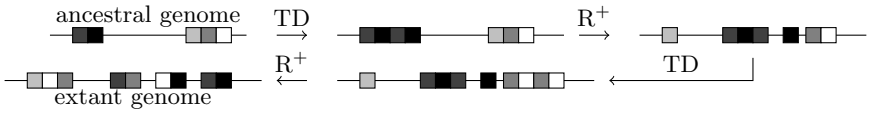

Fig. 1. A scenario from a non-duplicated ancestral genome which evolved through two tandem duplications (TD) and rearrangements (R^+). Squares denote syntenic markers.

A problem one could believe similar to the one we study in this paper is the analysis of rearrangement scenarios that use Tandem Duplication Random Loss (TDRL) operations known to occur in mt genomes of millipedes and eels [3]. However, this model differs as it supposes that one of each duplicated marker is deleted. Our problem is in fact closer to Mixtacki's model of the genome halving problem [9], although we consider tandem duplication events as an alternative to the whole genome duplication. Such model has been studied in [2] but in order to find a scenario between two given genomes through an heuristic.

Section 2 gives definitions. In Section 3 we give a distance for reconstructing a single tandem when all markers in the extant genome are duplicated. In Section 4 we provide a heuristic algorithm for reconstructing a single tandem when single markers are considered. In Section 5, we discuss the NP-hardness of various constraints on the reconstruction of more than a single tandem. We conclude in Section 6 with an application on maize mt genomes.

2 Preliminaries: duplicated genomes, rearrangement

A genome consists of linear or circular chromosomes that are composed of genomic markers. Markers are represented by signed integers such that the sign indicates the orientations of markers in chromosomes. By convention, $--x = x$. A linear chromosome is represented by an ordered sequence of signed integers surrounded by the unsigned marker $\circ$ at each end indicating the telomeres. A circular chromosome is represented by a circularly ordered sequence of signed integers. For example, $(1\ \ 2\ \ -3)\ (\circ\ \ 4\ \ -5\ \ \circ)$ is a genome composed of one circular and one linear chromosome.

Each genome contains at most two occurrences of each marker, called paralogs, arbitrarily denoted x and $\overline{x}$ (by convention $\overline{\overline{x}} = x$).

Definition 1. *A* duplicated genome *is a genome in which a subset of the markers are duplicated.*

For example, $(1\ \ 2\ \ -3\ \ -\overline{2})\ (\circ\ \ 4\ \ -5\ \ \overline{1}\ \ \overline{5}\ \ \circ)$ is a duplicated genome where markers 1, 2, and 5 are duplicated. A *non-duplicated genome* is a genome in which no marker is duplicated. A *totally duplicated genome* is a duplicated genome in which all markers are duplicated.

An *adjacency* in a genome is a pair of consecutive markers. Since a genome can be read in two directions, the adjacencies $(x\ \ y)$ and $(-y\ \ -x)$ are equivalent. For example, the genome $(1\ \ 2\ \ -\overline{2})\ (\circ\ \ -3\ \ \overline{1}\ \ \overline{3}\ \ \circ)$ has seven adjacencies, $(1\ \ 2)$, $(2\ \ -\overline{2})$, $(-\overline{2}\ \ 1)$, $(\circ\ \ -3)$, $(-3\ \ \overline{1})$, $(\overline{1}\ \ \overline{3})$, and $(\overline{3}\ \ \circ)$. When an adjacency contains a $\circ$ marker, *i.e.* a telomere, it is called a *telomeric adjacency*.

When needed, we will refer to marker extremities directly, indicating them using a dot. Thus, adjacency $(x\ \ y)$ concerns extremities $x\cdot$ and $\cdot y$.

A *double-adjacency* in a genome G is an adjacency $(a\ b)$ such that $(\overline{a}\ \overline{b})$ or $(-\overline{b}\ -\overline{a})$ is an adjacency of G as well. Note that a genome always has an even number of double-adjacencies. For example, the four double-adjacencies in the

following genome are indicated by $\diamond$:

$$G = (\circ \;\; 1 \;\; \overline{1} \;\; 3 \;\; 2 \;\diamond\; 4 \;\diamond\; 5 \;\; 6 \;\; \overline{6} \;\; 7 \;\; \overline{3} \;\; 8 \;\; \overline{2} \;\diamond\; \overline{4} \;\diamond\; \overline{5} \;\; 9 \;\; \overline{8} \;\; \overline{7} \;\; \overline{9} \;\; \circ)$$

A consecutive sequence of double-adjacencies can be rewritten as a single marker; this process is called *reduction*. For example, genome G can be reduced by rewriting $2 \;\diamond\; 4 \;\diamond\; 5$ and their paralogs as 10 and $\overline{10}$:

$$G^r = (\circ \;\; 1 \;\; \overline{1} \;\; 3 \;\; 10 \;\; 6 \;\; \overline{6} \;\; 7 \;\; \overline{3} \;\; 8 \;\; \overline{10} \;\; 9 \;\; \overline{8} \;\; \overline{7} \;\; \overline{9} \;\; \circ)$$

Definition 2. *A* single tandem duplicated genome *is a totally duplicated genome which can be reduced to a genome of the form* $(\circ \;\; x \;\; \overline{x} \;\; \circ)$.

In other words, a tandem duplicated genome is composed of a single linear chromosome where all adjacencies, except the two telomeric adjacencies and the central adjacency, are double-adjacencies. For example, the genome $(\circ\, 1 \diamond 2 \diamond 3 \diamond 4\; \overline{1} \diamond \overline{2} \diamond \overline{3} \diamond \overline{4}\, \circ)$ is a tandem-duplicated genome that can be reduced to $(\circ\; 5\; \overline{5}\; \circ)$ by rewritting $1 \diamond 2 \diamond 3 \diamond 4$ and $\overline{1} \diamond \overline{2} \diamond \overline{3} \diamond \overline{4}$ as 5 and $\overline{5}$.

Definition 3. *A* dedoubled genome *is a duplicated genome G such that for any duplicated marker x in G, either $(x \;\; \overline{x})$, or $(\overline{x} \;\; x)$ is an adjacency of G.*

One might notice that a single tandem duplicated genome, after reduction, is a unilinear dedoubled genome consisting of only one marker. Generalization of this property leads us to a short formal definition for genomes composed of several tandems, or *multiple tandem duplicated genomes*.

Definition 4. *A* k-tandem duplicated genome *is a totally duplicated genome which can be reduced to a unilinear dedoubled genome consisting of k distinct markers.*

For example, the genome $(\circ\, 1 \diamond 2 \diamond 3\; \overline{1} \diamond \overline{2} \diamond \overline{3}\; 4 \diamond 5\; \overline{4} \diamond \overline{5}\, \circ)$ is a 2-tandem duplicated genome that can be reduced to the dedoubled genome $(\circ\; 6\; \overline{6}\; 7\; \overline{7}\; \circ)$.

Naturally, following this definition, a single tandem duplicated genome is in fact a 1-tandem duplicated genome.

Definition 5. *A* perfectly duplicated genome *is a totally duplicated genome such that all adjacencies are double-adjacencies, none of them in the form $(x\; -\overline{x})$.*

For example, the genome $(1 \;\; 2 \;\; 3 \;\; 4 \;\; \overline{1} \;\; \overline{2} \;\; \overline{3} \;\; \overline{4})$ is a perfectly duplicated genome, while $(\circ \;\; 1 \;\; 2 \;\; -\overline{2} \;\; -1 \;\; \circ)$ is not. It is to note that this definition is equivalent to the one from [9].

The rearrangement operations considered in this paper will be the DCJ model, introduced in [10]. A *DCJ* operation on a genome G cuts two different adjacencies in G and glues pairs of the four exposed extremities to form two new adjacencies. A *DCJ scenario* between two genomes A and B is a sequence of DCJ operations allowing to transform A into B. The length of a scenario is the number of operations composing the scenario. The *DCJ distance* between two genomes A and B is the minimum length of a DCJ scenario between A and B.

Property 1. In the case of unichromosomal genomes, a perfectly duplicated genome is a single tandem duplicated genome which has been circularized (the perfectly duplicated genome can be reduced to $(x\; \overline{x})$, it just lacks telomeres).

3 Single Tandem Halving

We now state the first tandem halving problem considered in this paper.

Definition 6. *Given a unilinear totally duplicated genome G, the* single tandem halving problem *(or 1-tandem halving problem) consists in finding an* optimal *1-tandem duplicated genome H, such that the distance between G and H is minimal. This minimal distance is called the* 1-tandem halving distance*, and is denoted $d^t(G)$.*

Through reduction, this problem will be seen as a constraint on the well-known *DCJ genome halving problem*, as solved in [9]. We recall its definition, with slightly readapted notations.

Definition 7 ([9]). *Given a totally duplicated genome G, the* DCJ genome halving problem *consists in finding an* optimal *perfectly duplicated genome H, such that the DCJ distance between G and H is minimal. This minimal distance is called the* genome halving distance *and is denoted $d^p(G)$.*

$d^p(G)$ can be computed using a data structure called the *natural graph*, first introduced in [7]. $\texttt{NG}(G)$ is the graph whose vertices are the adjacencies of G, and 2 vertices are connected by an edge iff they share a paralogous *extremity* (see figure 2).

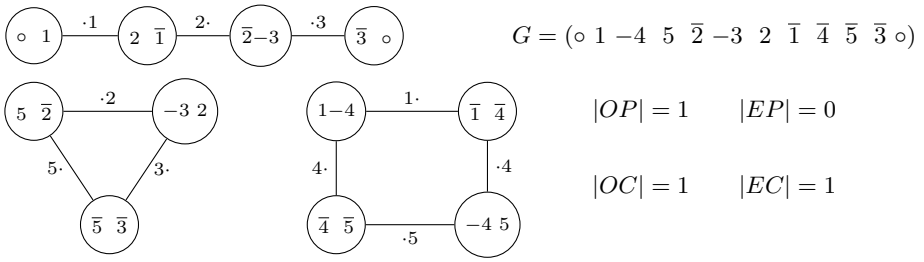

Fig. 2. The natural graph of G and the number of odd and even paths and cycles.

As an adjacency concerns a maximum of 2 markers extremities, this graph has a maximum degree of 2. Thus, it is composed of paths and cycles only. Moreover, it consists of nothing but 2-cycles and 1-paths if and only if G is a perfectly duplicated genome (a k-cycle or k-path is a cycle or path containing k edges). Using this graph, Mixtacki gave the following distance formula:

Theorem 1 ([9]). *Let G be a totally duplicated genome whose natural graph contains* EC *even cycles and* OP *odd paths. Then* $d^p(G) = n - |\mathrm{EC}| - \left\lfloor \frac{|\mathrm{OP}|}{2} \right\rfloor$.

Unlike the genome halving problem, the aim of the 1-tandem halving problem is to find a 1-tandem duplicated genome. This induces one double-adjacency not to

be reconstructed, which is inelegant to deal with. We will conveniently get rid of this concern.

From property 1, a 1-tandem genome that has been circularized is a perfectly duplicated genome and conversely. This allows us to establish a property that will reduce the 1-tandem halving problem to a constraint on genome halving.

Lemma 1. *Let G be a unilinear genome. Let G_c be the unicircular genome obtained by circularizing G. Then for any scenario that transforms G into a 1-tandem duplicated genome, there exists an equivalent scenario (of same length) transforming G_c into a unicircular perfectly duplicated genome, and vice versa.*

Proof. As G and G_c present the same breakpoints, the scenario conversion is straightforward. It suffices to apply the same DCJ on the same breakpoints. □

Thus, in the rest of this section, the focus will be on reconstructing an optimal perfectly duplicated genome such that it is unichromosomal. This is essentially a shape constraint on the genome halving solutions.

We will follow an approach a bit similar[1] to what has been done by Kováč *et al.* in [8], as they enforced another shape constraint on optimal perfectly duplicated genome configurations. It consists in taking any optimal configuration then applying a number of successive transformations (which we will refer to as *shapeshifting* in the present paper) on it, such that they preserve the distance, and that the optimal configuration converges towards the desired shape.

In the following sections G will denote a totally duplicated genome, and G_c its circularized version. H will be an optimal perfectly duplicated genome for G_c.

Following theorem 1, one can observe that circularization can alter the halving distance, depending on whether the path of $\texttt{NG}(G)$ is even or odd.

Property 2. If G is a genome such that $\texttt{NG}(G)$ contains an even path, $d^p(G_c) = d^p(G) - 1$. Else, $d^p(G_c) = d^p(G)$.

From Mixtacki's formula (Theorem 1), we know that optimal halving scenarios on circular genomes are scenarios which increase the number of even cycles at each step. There are two ways of increasing it. Either by splitting a cycle (*i.e.* extracting an even cycle from any cycle), or by merging two odd cycles.

As it can be quite complex at first sight, our shapeshifting system will first be described on a restricted class of genomes, namely those whose natural graph contains only even cycles. This way, we ensure that optimal halving scenarios consist only in cycle extractions. The restricted system will then be easily generalized to all genomes by considering merging operations.

3.1 Restricted shapeshifting system

Here we consider that $\texttt{NG}(G_c)$ has only even cycles. It follows that $\texttt{NG}(G)$ has an even path and $d^p(G_c) = d^p(G) - 1$.

[1] Although it had to be developed as a more complete system, due to the nature of our problem.

Anatomy of a multicircular perfectly duplicated genome. H is an optimal perfectly duplicated genome for G_c. Since G_c is unicircular, $\mathtt{NG}(G_c)$ contains nothing but cycles. Therefore, H consists of circular chromosomes only. For H to be a perfectly duplicated genome, circular chromosomes can be of two kinds : doubled chromosomes, which can be reduced to $(x\ \overline{x})$, and single chromosomes, which can be reduced to (x) and have a *paralog chromosome* in H, which can be reduced to $(\overline{x})$. Thus the number of single chromosomes is even.

Shapeshifting. Any optimal perfectly duplicated genome H induces a class $\mathcal{C}_H$ of optimal halving scenarios (the class of all optimal DCJ scenarios transforming G_c into H). By observing the structure of G_c and H, we will look for small changes to apply to $\mathcal{C}_H$, along two criteria : H must converge toward the desired shape, and it must preserve its optimality. Such small changes are called *shapeshifters*.

In our case, we want to end up with the least number of chromosomes in H (ideally only one), therefore we will look for ways to merge chromosomes while preserving optimality. This leads us to the following definition :

Definition 8. *A shapeshifter is an adjacency* $(x\ y)$ *such that* x *and* y *belong to* different chromosomes *of* H *(convergence towards the desired shape), and such that* $(x\ y)$ *(and therefore* $(\overline{x}\ \overline{y})$ *as well) can be reconstructed by an optimal halving scenario (preservation of optimality).*

For example, if H contains markers x and y in different chromosomes, C_x and C_y, and if $(x\ y)$ can be reconstructed by an optimal halving scenario, then such scenario induces a new shape for H such that C_x and C_y cannot be distinct chromosomes anymore.

As for now we consider genomes whose natural graph has even cycles only, shapeshifters are adjacencies reconstructible by extracting even cycles.

Property 3. Adjacencies $(x\ y)$ reconstructible by extracting even cycles are those such that there exists, in $\mathtt{NG}(G_c)$, an induced subgraph which is an *even* path, whose endpoints have outgoing edges $x\cdot$ and $\cdot y$.

Indeed, a DCJ reconstructing $(x\ y)$ will cut at the endpoints of such path and transform it into an even cycle. However, it is not necessary to consider all even paths, so w.l.o.g we shall focus only on 2-paths (ie. adjancencies $(x\ y)$ that are *present in* G_c), which correspond to 2-cycles extractions.

For example, $(1\ 4)$ in fig. 2 is a shapeshifter, as the 2-path induced by vertices $(1\ {-4})$, $(\overline{1}\ \overline{4})$, and $({-4}\ 5)$ meets the requirements.

We may proceed and show how to simply apply a shapeshifter on $\mathcal{C}_H$: Let $(x\ p)$ be the adjacency containing the extremity $x\cdot$ in H, and $(q\ y)$ the one containing the extremity $\cdot y$, it suffices to perform on H one DCJ cutting adjacencies $(x\ p)$ and $(q\ y)$ to reconstruct $(x\ y)$ (and $(p\ q)$), and the equivalent DCJ on the paralogs, cutting adjacencies $(\overline{x}\ \overline{p})$ and $(\overline{q}\ \overline{y})$ to reconstruct $(\overline{x}\ \overline{y})$ (and $(\overline{p}\ \overline{q})$).

One can easily verify that the resulting genome is still optimal (first DCJ brings H closer to G_c, second one reconstructs a perfectly duplicated genome).

Now we may proceed and study the shapeshifting induced by these DCJ.

Let $(x\ y)$ be a shapeshifter in G_c. x and y belong to different chromosomes in H, so there are only 3 possible cases depending on the types of chromosomes (C_S for single chromosomes, and C_D for doubled ones) which contain these markers: 1) $x \in C_S, y \in C_D$, 2) $x, y \in C_D$, 3) $x, y \in C_S$. The last one could lead to different shapes. Figure 3 illustrates how the genome shape can be altered, for each case.

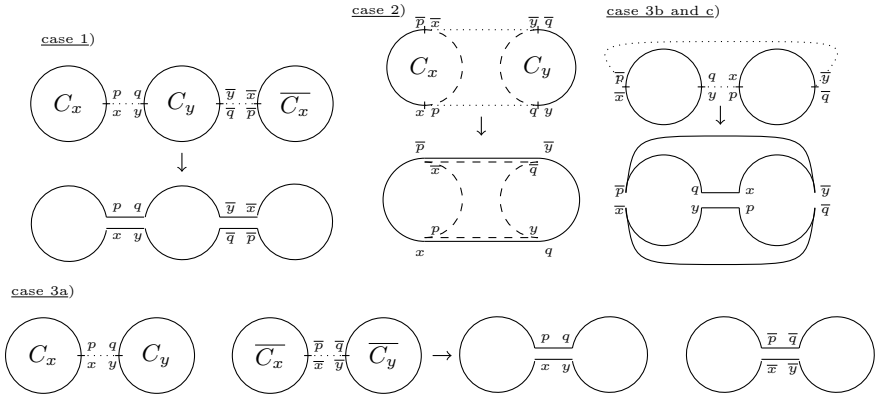

Fig. 3. The different shapes that can be obtained by applying a shapeshifter.

More formally, one can represent shapeshifting as a system of rewriting rules :

1) $2 \times C_S + C_D \to C_D$ 3.a) $4 \times C_S \to 2 \times C_S$ 3.c) $2 \times C_S \to 2 \times C_S$
2) $2 \times C_D \to 2 \times C_S$ 3.b) $2 \times C_S \to 2 \times C_D$

This is convenient as one can deduce useful properties by looking at these rules, which we are about to do, in order to study limit states of the system.

Property 4. Shapeshifting cannot increase the number of chromosomes.

Thus, any limit-cycle necessarily uses rules that do not change the number of chromosomes. Moreover, using rule 2 would eventually lead to using rule 3.b or 3.c as doubled chromosomes are changed into single chromosomes.

Property 5. Any limit-cycle of the system necessarily uses rule 3.b or 3.c.

Property 6. Parity of $|C_D|$ is invariant by shapeshifting.

Property 7. A unicircular genome (ie. one doubled chromosome) is the only steady state of the system.

Lemma 2. *By shapeshifting, the number of chromosomes in H can always be decreased under 3.*

Proof. Having 3 chromosomes or more guarantees existence of shapeshifters decreasing their number. Consider the case where H contains only 2 single chromosomes C_S and $\overline{C_S}$. Label the markers from G by the chromosome which holds them in H. Adding new chromosomes necessarily creates shapeshifters between at least one of the new chromosomes and C_S or $\overline{C_S}$. Such shapeshifter decreases the number of chromosomes. □

Lemma 3. *There exists a unicircular optimal perfectly duplicated genome for G_c if and only if H has an* odd *number of doubled chromosomes.*

Proof. Straightforward from lemma 2 and property 6. □

Lemma 4. *If H has an* even *number of doubled chromosomes, the minimum number of DCJ operations required to reconstruct a unicircular perfectly duplicated genome is $d^p(G_c) + 1$, and it can always be attained.*

Proof. From lemma 3, it is impossible to attain a unicircular genome in $d^p(G_c)$ operations. However, from lemma 2 and property 5, it is then always possible to attain two single chromosomes. Two single chromosomes can then be transformed into one doubled chromosome by one DCJ. □

In conclusion, restricted shapeshifting allows to compute the tandem distance of any genome G such that $\texttt{NG}(G)$ contains only even cycles.

Theorem 2. *Let G be a totally duplicated genome such that $\texttt{NG}(G)$ contains only even cycles. Let G_c be its circularized version, and H any optimal perfectly duplicated genome for G_c. $d^t(G) = d^p(G) - 1$ if and only if H contains an odd number of doubled chromosomes. Else $d^t(G) = d^p(G)$.*

Proof. Since $\texttt{NG}(G)$ contains only even cycles, it contains an even path. Therefore from property 2, $d^p(G_c) = d^p(G) - 1$. From lemma 1 we have that $d^t(G) = d^p(G_c)$ if and only if there exists a unicircular optimal perfectly duplicated genome. Theorem then follows from lemmas 3 and 4. □

The next step is to generalize the shapeshifting system in order to take all possible genomes into account.

3.2 Generalized shapeshifting system

As usual, G is a totally duplicated genome, G_c its circularized version, and H an optimal perfectly duplicated genome for G_c. We will also keep the same notations related to shapeshifters as in the previous section : $(x\ y)$ is a shapeshifter such that x (resp. y) is present in chromosome C_x (resp. C_y) of H, through adjacency $(x\ p)$ (resp. $(q\ y)$).

The difference with restricted shapeshifting is that, *in addition* to everything covered by restricted shapeshifting, optimal halving scenarios may now also contain cycle merges. Therefore we have to consider shapeshifters that are adjacencies which can be optimally reconstructed through merges.

Property 8. Adjacencies $(x\ y)$ reconstructible by merges are those such that extremities $x\cdot$ and $\cdot y$ are in *two distinct odd cycles* of $\texttt{NG}(G_c)$.

Corresponding shapeshifters can still allow the same shapeshifting rules depending on the types of C_x and C_y. Additionally, it is now possible to have $p = \overline{y}$ and $q = \overline{x}$. This implies that $C_y = \overline{C_x}$ and induces yet another degenerated case. The generalized shapeshifting set of rule becomes :

1) $2 \times C_S + C_D \to C_D$	3.a) $4 \times C_S \to 2 \times C_S$	3.c) $2 \times C_S \to 2 \times C_S$
2) $2 \times C_D \to 2 \times C_S$	3.b) $2 \times C_S \to 2 \times C_D$	**3.d)** $\mathbf{2 \times C_S \to C_D}$

This new rule gives generalized shapeshifting a very interesting property.

Property 9. Rule 3.d changes parity of C_D.

Lemma 5. *If $\texttt{NG}(G_c)$ contains odd cycles, and if H is made of two single chromosomes, then rule 3.d can be applied.*

Proof. As $\texttt{NG}(G_c)$ contains odd cycles, there are merges in any optimal scenario from G_c to H. Thus, there exists an adjacency $(x\ p)$ in C_x such that the adjacencies concerning extremities $x\cdot$ and $\cdot p$ are in two distinct odd cycles of $\texttt{NG}(G_c)$. By definition, the adjacency concerning extremity $\cdot\overline{p}$ is in the same cycle as the one concerning $\cdot p$. Therefore, $(x\ \overline{p})$ is a shapeshifter inducing rule 3.d. □

Corollary 1. *Presence of odd cycles in $\texttt{NG}(G_c)$ ensures a unicircular optimal perfectly duplicated genome that can always be reached, as rule 3.d can always adjust the parity of C_D if needed.*

Theorem 3. *Let G be a totally duplicated genome such that $\texttt{NG}(G)$ contains at least one odd cycle, and G_c its circularized version. Then $d^t(G) = d^p(G_c)$.*

Proof. From lemma 1 we have $d^t(G) = d^p(G_c)$ iff there exists a unicircular optimal perfectly duplicated genome. Corollary from lemma 5 ensures that there does. □

3.3 Conclusion

We finally state a definite formula for the halving distance, as well as results on computational complexity of this problem, by gathering results from the previous sections.

Theorem 4. $d^t(G) = n - |\mathrm{EC}| - |\mathrm{EP}| + f_G$

Where f_G is a parameter that is equal to 1 iff C_D is even and $|\mathrm{OC}| = 0$, and is equal to 0 otherwise. $|\mathrm{EC}|$, $|\mathrm{EP}|$ and $|\mathrm{OC}|$ are respectively the number of even cycles, even paths and odd cycles in $\texttt{NG}(G)$.

Proof. Straightforward from theorems 2 and 3. □

Theorem 5. *$d^t(G)$ can be computed in linear time.*

Proof. $\mathtt{NG}(G)$ can be computed in linear time, as well as an optimal perfectly duplicated genome. □

Theorem 6. *Computing a scenario can be done in quadratic time.*

Proof. An optimal perfectly duplicated genome can be computed in $O(n)$ time using Mixtacki's algorithm ([9]). From lemma 2, one can reduce H to the minimum number of chromosomes using $O(n)$ shapeshifters. Each shapeshifter can be found in $O(n)$ time, so we have a $O(n^2)$ time shapeshifting algorithm. An optimal DCJ scenario between G and H can then be computed in $O(n)$ time using Yancopoulos' algorithm ([10]). Thus the algorithm takes quadratic time on the whole. □

4 Disrupted Single Tandem Halving

As we could solve the 1-tandem halving problem, a first direction for generalization will be considering genomes containing both duplicated and non-duplicated markers, as it is in better accordance with real biological data.

This can be seen as a 1-tandem halving problem in which adjacencies between duplicated markers can be broken by presence of non-duplicated ones. In other words, non-duplicated markers *disrupt* the 1-tandem halving.

Definition 9. *The* disrupted 1-tandem halving problem *is a variant of the 1-tandem halving problem in which the genome contains both duplicated and non-duplicated markers. The duplicated markers have to be regrouped and arranged in tandem. The corresponding distance, the* disrupted 1-tandem halving distance*, is denoted $d^{t'}(G)$.*

Preliminary analysis. Any optimal disrupted 1-tandem halving scenario performs two tasks : it gathers duplicated markers together (gathering phase), and it reorganizes them in a tandem (tandem phase).

Definition 10. *A* break *is an interval of non-duplicated markers surrounded by duplicated markers.*

From now on, G is a duplicated genome containing n duplicated markers separated by b breaks.

Definition 11. *A* gathering operation *is a DCJ which reduces the number of breaks in G.*

Note that the presence of excisions in the gathering phase may produce a genome consisting of multiple chromosomes. Excisions and their resulting chromosomes will be categorized depending on whether said chromosomes can be reintegrated at best in their source chromosome while increasing the number of even cycles (*good* excision/chromosome), leaving it unchanged (*neutral*) or decreasing it (*bad*). As this variation in |EC| changes the tandem distance, we get the following property.

Property 10. Once the gathering phase is over in G, the remaining distance is $d^t(G) + C^0 + 2C^-$, with C^0 the number of neutral chromosomes and C^- the number of bad ones.

The key to build an optimal disrupted 1-tandem halving scenario is to find a gathering scenario that maximizes the number of even cycles while minimizing the number of neutral and bad excisions.

Optimizing the gathering scenario. A DCJ can decrease the number of breaks by at most 1.

Property 11. The minimum number of gathering operations is b.

Gathering operations are DCJ whose breakpoints are on path endpoints from $\mathtt{NG}(G)$. Breakpoints in two distinct paths will merge them, while breakpoints on the endpoints of a same path will circularize it.

Property 12. An optimal gathering operation is one that either merges two odd paths, or circularizes an even path.

We now give the maximum number of even cycles a set of b gathering operations can create.

Lemma 6. *A shortest gathering scenario can create up to* $\left\lfloor \frac{|\mathrm{OP}|}{2} \right\rfloor + |\mathrm{EP}| - 1$ *even cycles.*

Proof. sketch of proof: Any even path can be circularized by one DCJ, while any two odd paths can be turned into two even cycles with 2 DCJs. Since b breaks induce $b+1$ paths in $\mathtt{NG}(G)$, the number of gathering operations we can use is $b = |\mathrm{OP}| + |\mathrm{EP}| - 1$. □

Corollary 2. $d^{t'}(G) \geq n - |\mathrm{EC}| - 1 + \left\lceil \frac{|\mathrm{OP}|}{2} \right\rceil$.

This is assuming a shortest gathering phase produced no bad nor neutral chromosome, and that we are in the best case for the remaining tandem distance ($d^t(G) = d^p(G) - 1$).

Neutral excisions induce a penalty which is the same as performing a non-optimal gathering reversal, bad excisions are even worse. Thus our greedy heuristic will proceed as follows: Look for an optimal gathering operation which is a reversal or a good excision. When there is none, perform a non-optimal gathering reversal.

Let $C_h(G)$ be the number of even cycles produced by the heuristic, then we obtain the following upperbound : $d^{t'}(G) \leq n - |\mathrm{EC}| + |\mathrm{OP}| + |\mathrm{EP}| - 1 - C_h(G)$.

In the worst case, $C_h(G)$ can be equal to 0, however, the algorithm seems to perform pretty well on random genomes, giving values close to the lowerbound.

5 Multiple tandem halving

Unlike 1-tandem halving, k-tandem halving can be defined in various ways. We explored several constraints on the k-tandem halving (detailed studies are given as supplementary material[2]). First, when one fixes the number of tandem to be reconstructed (k) the problem is NP-hard. Fixing the content of each of the k tandem does not help and the complexity of the problem remains the same. The same result arises when one fixes the tandem order in the ancestral genome. Lastly, a "signed" version where the orientation of the tandems is fixed is also NP-hard. Approximation algorithms should be considered next, as those problems are the most interesting ones from a biological viewpoint.

6 Application

As an application, we used data from [6]. We analyzed the mitochondrial genome of *Zea mays ssp. mays CMS-C* which is made of 69 syntenic markers, 21 of them being duplicated. Figure 4 shows two optimal scenarios obtained by applying algorithm described in Section 4: a) with reversals only, b) with reversals and excision/reintegration. Those last type scenarios raises the questions about mecanisms that led to duplication in plant mitogenomes [1]. Detailed scenarios are provided in the supplementary material[2].

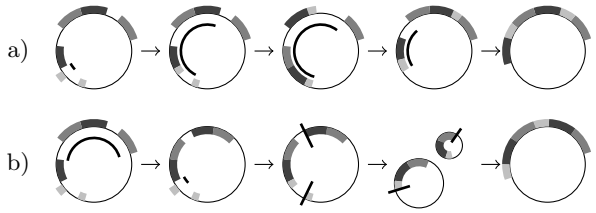

Fig. 4. Two parsimonious scenarios reconstructing a putative ancestral genome just before the tandem duplication event. Large segments show duplicated markers separated by breaks. The black line inside circles show the reversals applied while segments cutting the circles show the excision/reintegration applied.

7 Conclusion

In this paper we introduced several instances of the problem of reconstructing an ancestral genome which evolved through tandem duplications and other rearrangement operations. We obtained a distance formula for the simpliest case where all markers have been duplicated and only one tandem duplication occurred ;

[2] `http://www.lifl.fr/~varre/download/suppmatWABI2012.pdf`

which can be computed in linear time. For the case where some markers have not been duplicated we obtained an approximate algorithm. Unfortunately, all other cases we explored are NP-hard. Future work should be to design approximate algorithms allowing to go further in the analysis of biological data, in order to be able to compute phylogenetic trees and putative ancestors for a set of genomes fo which duplicates appeared through tandem duplications.

References

1. S. Backert, B. L. Nielsen, and T. Börner. The mystery of the rings: structure and replication of mitochondrial genomes from higher plants. *Trends in Plant Science*, 2(12):477–483, 1997.
2. M. Bader. Genome rearrangements with duplications. *BMC Bioinformatics*, 11(S-1):27, 2010.
3. M. Bernt, K.-Y. Chen, M.-C. Chen, A.-C. Chu, D. Merkle, H.-L. Wang, K.-M. Chao, and M. Middendorf. Finding all sorting tandem duplication random loss operations. *J. Discrete Algorithms*, 9(1):32–48, 2011.
4. S. Chang, T. Yang, T. Du, J. Chen, J. Yan, J. He, and R. Guan. Mitochondrial genome sequencing helps show the evoltutionary mechanism of mitochondrial genome formation in *Brassica*. *BMC Genomics*, 12(497), 2011.
5. A. Darracq, J.-S. Varré, L. Maréchal-Drouard, A. Courseaux, V. Castric, P. Saumitou-Laprade, S. Oztas, P. Lenoble, B. Vacherie, V. Barbe, and P. Touzet. Structural and content diversity of mitochondrial genome in beet: a comparative genomic analysis. *Genome Biology and Evolution*, 3:723–736, 2011.
6. A. Darracq, J.-S. Varré, and P. Touzet. A scenario of mitochondrial genome evolution in maize based on rearrangement events. *BMC Genomics*, 11(233), 2010.
7. N. El-Mabrouk and D. Sankoff. The reconstruction of doubled genomes. *SIAM J. Comput.*, 32(3):754–792, 2003.
8. J. Kováč, R. Warren, M. D.V. Braga, and J. Stoye. Restricted DCJ model: rearrangement problems with chromosome reincorporation. *Journal of Computational Biology*, 18(9):1231–1241, 2011.
9. J. Mixtacki. Genome halving under DCJ revisited. In Xiaodong Hu and Jie Wang, editors, *Proceedings of COCOON'08*, volume 5092 of *Lecture Notes in Computer Science*, pages 276–286. Springer, 2008.
10. S. Yancopoulos, O. Attie, and R. Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346, 2005.