

# A framework for cost-constrained genome rearrangement under Double Cut and Join

Pijus Simonaitis<sup>1</sup>, Annie Chateau<sup>1,2</sup>, and Krister M. Swenson<sup>1,2</sup>

<sup>1</sup>LIRMM, CNRS – Université Montpellier, 161 rue Ada, 34392 Montpellier, France,

<sup>2</sup>Institut de Biologie Computationnelle (IBC), Montpellier, France

## Abstract

The study of genome rearrangement has many flavours, but they all are somehow tied to edit distances on variations of a multi-graph called the breakpoint graph. We study a weighted 2-break distance on Eulerian 2-edge-colored multi-graphs, which generalizes weighted versions of several Double Cut and Join problems, including those on genomes with unequal gene content. We affirm the connection between cycle decompositions and edit scenarios first discovered with the SORTING BY REVERSALS problem. Using this we show that the problem of finding a parsimonious scenario of minimum cost on an Eulerian 2-edge-colored multi-graph – with a general cost function for 2-breaks – can be solved by decomposing the problem into independent instances on simple alternating cycles. For breakpoint graphs, and a more constrained cost function, based on coloring the vertices, we give a polynomial-time algorithm for finding a parsimonious 2-break scenario of minimum cost, while showing that finding a non-parsimonious 2-break scenario of minimum cost is NP-Hard.

## 1 Introduction

Edit distance problems are a mainstay in computer science. On strings, edit distances have roots in computational linguistics, and are at the heart of many approximate string matching algorithms for signal processing and text retrieval [19]. The Levenshtein distance is the classic example, which asks for a minimum number of insertions, deletions, or substitutions of characters needed to transform one string into another. On graphs, the first edit distance to be considered is an analogue to the Levenshtein distance; insertion and deletion of vertices and edges are allowed, along with vertex and edge label substitution [22]. Graph edit distances have become an important tool in pattern recognition, which is at the heart of modern image processing and computer vision [5, 12].

From early on, there were major applications for string edit distances in computational biology, mainly due to the linear nature of DNA, RNA, and protein molecules [20, 23, 17, 13]. More recently, graph edit distances have found a role in the comparison of gene regulatory networks [18]. In the next subsection we outline the pervasive role of edit distances in genome rearrangements.

In this paper we address a graph edit distance on Eulerian 2-edge-colored multi-graphs, that is, a multi-graph with black and gray edges such that every vertex is incident to the same number of black and gray edges. When we say *graph*, we will mean Eulerian 2-edge-colored multi-graph. The edit operations that we consider are the *k-break*, where *k* black edges are replaced such that the degree of all vertices in the graph are conserved. For 2-breaks, we generalize the edit distance problem to

consider costs on the operations, posed as the MINIMUM COST PARSIMONIOUS SCENARIO (MCPS) problem. This problem asks for a minimum-length scenario of 2-breaks transforming the set of black edges into the set of gray edges.

Our main contribution is a clean formalism that facilitates simple proofs for edit distances on 2-breaks with costs. While weighted edit operations have been considered in the past, to our knowledge, this is the first study of weighted 2-breaks in this general setting [2, 10, 29, 28].

In Section 2.2 we show that a  $k$ -break scenario on a graph  $G$  partitions  $G$  into a set of Eulerian subgraphs, such that no  $k$ -break operates on edges from different subgraphs. This decomposition theorem allows us to show a strong link between MAXIMUM ALTERNATING CYCLE DECOMPOSITION and the length of a parsimonious 2-break scenario for a graph. It is also the cornerstone for Section 3, showing that MCPS can be computed on a graph  $G$  if there is a method for computing MCPS on a circle. By *circle* we mean a graph where all vertices are incident to exactly one gray and one black edge (Section 3).

These results are general in the sense that the variants of the breakpoint graph typically used for genome rearrangement problems are all specific instance of the Eulerian 2-edge-colored multi-graph. Section 1.1 gives such examples.

Section 4 is dedicated to a specific cost function that depends on a coloring of the vertices: if a 2-break replaces edges  $(a, b)$  and  $(c, d)$  with  $(a, d)$  and  $(c, b)$ , then the cost is zero if  $a$  and  $c$  have the same color, or  $b$  and  $d$  have the same color, otherwise it is of cost 1. Our tool for reasoning with this cost function is called the *color-merged graph*, which is obtained from a graph by merging all vertices of the same color. Theorem 5 states that it is NP-Hard to compute MINIMUM COST SCENARIO, even for circles. After establishing the necessary links between cycle decompositions in a graph and its merged graph (Section 4.1), we show how to compute MCPS on a circle (Sections 4.2 and 4.3). Finally, we show that MINIMUM COST PARSIMONIOUS SCENARIO for the colored cost function is computable in  $O(n^4)$  time.

## 1.1 Edit distances and breakpoint graphs

In the area of gene order comparison through gene rearrangements, both the string and graph edit distances play a central role [11]. The typical genome rearrangement problem is an edit distance problem on strings of genes called *genomes*, where each gene occurs exactly once. The first biologically motivated edit operation that was studied is the *reversal* of a substring of a genome [30, 24]. When the relative direction of gene transcription is known, this problem is called SIGNED SORTING BY REVERSALS (SSBR), and is intimately linked to the *breakpoint graph* [16, 1]. Roughly speaking, this multi-graph has a vertex for each gene extremity, and for every pair of adjacent gene extremities there is an edge. Thus, the edges for a single genome constitute a perfect matching. Say there is a gray genome and a black genome, then we have a gray matching and a black matching in the breakpoint graph. In this case the graph is 2-regular, and therefore decomposes into disjoint cycles of alternating color. A reversal on the black genome replaces two adjacencies. In this way, SSBR can be seen as a graph edit problem where the graph edit operation is a replacement of 2 edges, simulating the reversal of a substring. The SSBR problem can be solved in polynomial-time [14].

When relative transcription directions are unknown for the genes, we have the SORTING BY REVERSALS (SBR) problem. In SBR, since the orientation of gene extremities are unknown, the vertices for the two extremities of each gene are merged into a single vertex in the breakpoint graph. The result is a 4-regular graph. While finding the alternating-color cycle decomposition of the 2-regular breakpoint graph is trivial, finding the same for a 4-regular graph is not: Caprara showed that SBR is as hard as finding a MAXIMUM ALTERNATING CYCLE DECOMPOSITION (MACD), and that MACD is NP-Hard [6].

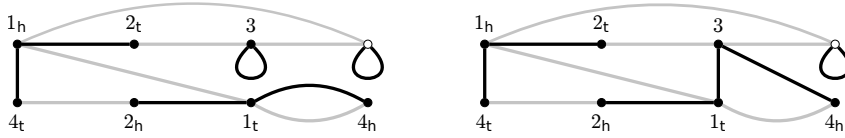


Figure 1: Examples of breakpoint graphs;  $G(A, B)$  on the left and  $G(A', B)$  on the right, where  $A = \{\{1_h, 2_t\}, \{2_h, 1_t\}, \{1_h, 4_t\}, \{4_h, 1_t\}\}$ ,  $A' = \{\{1_h, 2_t\}, \{2_h, 1_t\}, \{1_h, 4_t\}, \{4_h, 3\}, \{3, 1_t\}\}$ , and  $B = \{\{3\}, \{3, 2_t\}, \{2_h, 4_t\}, \{4_h, 1_t\}, \{1_h\}\}$ . Edges adjacent to a special vertex  $\circ$  represent the endpoints of linear chromosomes (e.g. gray edges  $\{3, \circ\}$  and  $\{1_h, \circ\}$ ), while missing genes are represented by self-loops (e.g. the black edge  $(3, 3)$  and the gray edge  $(1_h, 1_t)$ ), called *ghost adjacencies* in [25]. In the circular genomes  $A$  and  $A'$ , gene 1 is repeated twice, and the DCJ transforming  $A$  into  $A'$  is an insertion of an unoriented gene 3, corresponding to the 2-break  $G(A, B) \rightarrow G(A', B)$ . A DCJ scenario transforming  $A'$  into the linear genome  $B$  includes a deletion of one copy of a gene 1.

Currently the most mathematically clean model for genome rearrangement is called the *Double Cut and Join* (DCJ) model [31, 3]. Genome extremities that are adjacent are paired, and transformations of these pairs occur by swapping elements of the pairs. DCJ is a generalization of the SSBR paradigm, since a DCJ operation on a genome can simulate a reversal. This implies that DCJ is a less restrictive graph edit distance model, where edge pair  $\{(a, b), (c, d)\}$  is replaced by either  $\{(a, c), (b, d)\}$  or  $\{(a, d), (b, c)\}$  (in SSBR only one of the two replacements would be a valid reversal). The DCJ edit distance is inversely proportional to the number of cycles in the breakpoint graph.

While much of the work on genome rearrangement is on models where exactly one occurrence of each gene exists in each genome, content modifying operations like gene insertion, deletion, and duplication have been considered. Approaches to these problems are inextricably tied to finding cycle decompositions on breakpoint graphs [9, 4, 26, 27]. In the 2-break model, the MAXIMUM ALTERNATING CYCLE DECOMPOSITION for each of these graphs implies the 2-break distance. Figure 1 shows breakpoint graphs (see Section 5 for a definition) for genomes with unequal gene content. Our results on 2-breaks, up to and including Section 4.3, are directly applicable to the DCJ model, even in the presence of content modifying operations.

## 1.2 Distinguishing features of our model

We have previously worked on DCJ problems with weight functions [29, 28]. These papers address the 2-break model for 2-regular breakpoint graphs with a cost function on the edges instead of the vertices. The differences between the vertex-cost model from this paper, and edge-cost model of previous work are subtle. Indeed, the algorithmic results from Sections 4.2, 4.3, and 5 are similar to those in [29], but bear the hallmark of this vertex colored model: simplicity. This simplicity facilitates the much more general nature of this work.

From a practical perspective, the vertex-cost model we present here sidesteps the weakest feature of the edge-cost model: the cost function. In the edge-cost model, each DCJ on a particular pair of adjacencies has two possible states. Which state is chosen determines the cost of subsequent DCJs on these adjacencies. In this way, the edge-cost function is state dependent.

For our vertex-cost model, the natural bijection between vertices representing the same extremity in the two genomes allows for a stateless cost function. This way, the weight of a DCJ can be computed as function of both genomes instead of just one. We also speculate that the computation

of a general cost function under the vertex-cost model will be simpler. The only known polynomial time algorithm for the edge-cost model is restricted to a color based cost function.

## 2 k-breaks on a 2-edge-colored graph

### 2.1 Definitions

In our work a *graph* will be an Eulerian 2-edge-colored undirected multi-graph  $G = (V, E^b \cup E^g)$  with black and gray edges. We set  $e(G) = |E^b| = |E^g|$ .

**Definition 1** (Eulerian graphs and alternating cycles). *For a graph  $G$  and a vertex  $v$  we set  $\bar{d}(G, v) = d^b(G, v) - d^g(G, v)$  where  $d^b$  and  $d^g$  are black and gray degrees of  $v$ .  $G$  is Eulerian if  $\bar{d}(G, v) = 0$  for every vertex  $v$ . A cycle is alternating if it is Eulerian. All the cycles in our work will be alternating unless specified otherwise. The length of a cycle is its number of black edges.*

**Definition 2** (k-break). *A k-break is a transformation of a graph  $G$  into  $G'$  that replaces  $k$  black edges  $(x_1, x_2), \dots, (x_{2k-1}, x_{2k})$  by  $(x_{q_1}, x_{q_2}), \dots, (x_{q_{2k-1}}, x_{q_{2k}})$  while preserving the degree of all the vertices of  $G$ . In other words,  $d^b(G, v) = d^b(G', v)$  for all  $v$  and the multi-sets  $\{x_{q_1}, \dots, x_{q_{2k}}\}$  and  $\{x_1, \dots, x_{2k}\}$  are equal.*

Since  $G$  is Eulerian, it admits a decomposition into edge-disjoint alternating cycles. We denote the size of a MAXIMUM ALTERNATING CYCLE DECOMPOSITION (MACD) by  $c(G)$ . If  $c(G) = 1$  we say that  $G$  is a *simple cycle*. All the cycles in a MACD are simple. A graph with equal multi-sets of black and gray edges will be called *terminal*. A *scenario* will be a sequence of k-breaks transforming a graph into a terminal graph.

### 2.2 Cycle decompositions for a k-break scenario

To a k-break scenario we will associate a cycle decomposition such that all the edges replaced by a k-break belong to a single cycle. In Section 3 this will allow us to concentrate on the scenarios for the simple cycles instead of the general graphs.

We say that two edges are *l-related* according to a given scenario if among the first  $l$  k-breaks of the scenario there is one replacing these two edges.  $P_l$  will be a set of the equivalence classes of the transitive closure of this relation. We start by constructing  $P_l$ , which is a partition of the black edges of a graph, and proceed by showing how the gray edges can be incorporated to obtain a cycle decomposition of a graph.

Take a k-break scenario of length  $m$  for a graph  $G = G_0$ . We denote the graph after  $l \geq 0$  k-breaks by  $G_l$ . We label the  $G_0$  black edges by the set  $\{1, \dots, e(G)\}$  and partition them into singletons  $P_0 = \{\{1\}, \dots, \{e(G)\}\}$ . The  $l$ th k-break replaces  $k$  black edges labeled  $e_1, \dots, e_k$  of  $G_{l-1}$  by  $k$  black edges that we arbitrarily label  $e_1, \dots, e_k$ . Let  $Q_i$  be the subset of a partition  $P_{l-1}$  including the edge  $e_i$ . We obtain a new partition  $P_l$  from  $P_{l-1}$  by merging all  $Q_i$  into the set  $\bigcup_{i=1}^k Q_i$ . After applying all  $m$  merges of a scenario we obtain a partition  $P_m$  of the set  $\{1, \dots, e(G)\}$ .

We label the gray edges of  $G_m$  by  $\{e(G) + 1, \dots, 2e(G)\}$ , keeping the same labeling for all  $G_l$ . Since  $G_m$  is terminal, its multi-sets of black and gray edges are equal. This implies a one-to-one mapping of labels between  $\{1, \dots, e(G)\}$  and  $\{e(G) + 1, \dots, 2e(G)\}$  such that corresponding labels have the same endpoints in  $G_m$ . Using this mapping, we produce a partition on  $\{1, \dots, 2e(G)\}$  called  $R_l$ , that includes all black edge labels  $e_i \in P_l$  along with the gray edge label that  $e_i$  maps to. We will show that every subset of  $R_m$  defines an Eulerian subgraph of  $G$ . For a subset  $Q \subset \{1, \dots, 2e(G)\}$  and  $l \in \{0, \dots, m\}$ , we define  $s(G_l, Q)$  to be the edge-induced subgraph of

$G_l$  having only the edges of  $G_l$  labeled by elements of  $Q$ . By construction, for  $Q \in R_m$  and a vertex  $v$  we have  $\bar{d}(s(G_m, Q), v) = 0$ . Lemma 1, proven in the Appendix, establishes the equality  $\bar{d}(s(G_0, Q), v) = \bar{d}(s(G_m, Q), v)$ , which means that  $s(G, Q)$  is Eulerian. For a scenario  $\rho$   $C(\rho) = R_m$  denotes the *cycle decomposition* of the scenario  $\rho$ .

**Lemma 1.**  $\bar{d}(s(G_0, Q), v) = \bar{d}(s(G_l, Q), v)$  for a vertex  $v$  and  $Q \in R_l$  with  $l \in \{0, \dots, m\}$ .

**Theorem 1.** The minimum length of a 2-break scenario  $l(G)$  is equal to  $e(G) - c(G)$ .

*Proof.* We first show that for a 2-break scenario  $\rho$  of length  $m$  we have  $|C(\rho)| \geq e(G) - m$ . Recall that the size of  $C(\rho)$  is equal to  $|R_m| = |P_m|$  where  $P_m$  is a partition of a set  $\{1, \dots, e(G)\}$  encountered in the construction of  $C(\rho)$ . The  $l$ th 2-break of  $\rho$  replaces two edges  $e_1$  and  $e_2$  and merges the subsets  $Q_1$  and  $Q_2$  of  $P_{l-1}$  (recall that  $e_1 \in Q_1$  and  $e_2 \in Q_2$ ) to obtain a partition  $P_l$ . By construction,  $|P_l| \geq |P_{l-1}| - 1$  as at most two subsets get merged. The size of  $P_0$  is equal to  $e(G)$ , thus the size of  $P_m$  is at least  $e(G) - m$ , meaning that  $c(G) \geq e(G) - l(G)$ .

On the other hand, for any cycle  $c$  of length  $l > 1$  there is a 2-break transforming  $c$  into a union of length 1 and length  $l - 1$  cycles. In this way we obtain a scenario of length  $l - 1$  for  $c$ , and can transform every cycle of a MACD of  $G$  independently, obtaining a 2-break scenario of length  $e(G) - c(G)$ . Thus,  $l(G) \leq e(G) - c(G)$ .  $\square$

**Corollary 1.**  $C(\rho)$  for a parsimonious 2-break scenario  $\rho$  is a MAXIMUM ALTERNATING CYCLE DECOMPOSITION of  $G$ .

### 3 Minimum Cost Parsimonious Scenario

Consider a non-negative cost function  $\varphi$  for the 2-breaks on  $V$ . By *2-breaks on  $V$*  we mean the set of all the 2-breaks on the complete graph with vertices  $V$ . The cost of a scenario on a graph  $G = (V, E^b \cup E^g)$  is the sum of the costs of its 2-breaks. We provide an example of a cost function  $\varphi$  in Section 4. The MINIMUM COST PARSIMONIOUS SCENARIO (MCPS) for a graph under cost function  $\varphi$  is a minimum cost scenario among the scenarios for  $G$  of minimum length.  $\text{MCPS}_\varphi(G)$  denotes the cost of a MCPS for a graph  $G$  and a cost function  $\varphi$ .

**Theorem 2.**  $\text{MCPS}_\varphi(G)$  is the minimum over  $\{\sum_{c \in C} \text{MCPS}_\varphi(c) \mid C \text{ is a MACD of } G\}$ .

*Proof.* Take a MCPS  $\rho$  for  $G$  and  $\varphi$ .  $C(\rho)$  is a MACD of  $G$  due to Theorem 1. The subsequence of  $\rho$  consisting of the 2-breaks replacing the edges of a cycle  $c \in C(\rho)$  is a parsimonious scenario for  $c$  that we name  $\rho_c$ . A 2-break sequence  $\rho'$  obtained by performing  $\rho_c$  one by one is a parsimonious scenario for  $G$ . The costs of  $\rho$  and  $\rho'$  are equal as they consist of the same 2-breaks performed in different order. This way we know that the cost of  $\rho$  is smaller or equal to  $\sum_{c \in C(\rho)} \text{MCPS}_\varphi(c)$ . On the other hand, for a MACD  $C$  minimizing the sum  $\sum_{c \in C} \text{MCPS}_\varphi(c)$  we can construct a parsimonious scenario of cost  $\sum_{c \in C} \text{MCPS}_\varphi(c)$  transforming each cycle separately.  $\square$

#### 3.1 MCPS for a simple cycle

A simple cycle  $S$  might have a certain number  $d(S)$  of vertices  $v$  with  $d^b(S, v) = d^g(S, v) = 2$ . It is easy to check that  $d^b(G, v) = d^g(G, v) < 3$  for any vertex. If  $d(S) = 0$ , then we call  $S$  a *circle*. See Figure 2 for an example of a simple cycle that is not a circle.

Take a simple cycle  $S$  with  $d(S) > 0$ , a cost function  $\varphi$  for the 2-breaks on its vertices and a vertex  $v_0$ . For every Eulerian cycle  $(v_0, \dots, v_{m-1}, v_0)$  of  $S$  we construct a circle  $(u_0, \dots, u_{m-1}, u_0)$

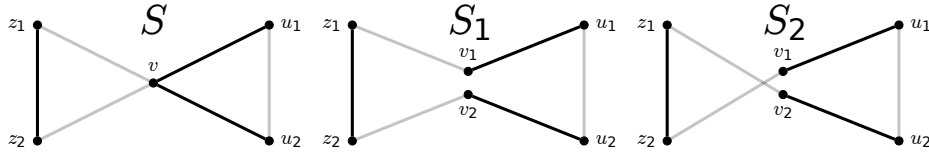


Figure 2:  $S$  and its two Eulerian circles  $S_1$  and  $S_2$ . We recover  $S$  by merging  $v_1$  and  $v_2$ .

with a cost function  $\varphi'$  defined as follows for  $i, j, k, l \in \{0, \dots, m-1\}$ :

$$\varphi'((u_i, u_j), (u_k, u_l) \rightarrow (u_i, u_k), (u_j, u_l)) = \varphi((v_i, v_j), (v_k, v_l) \rightarrow (v_i, v_k), (v_j, v_l)).$$

There are no more than  $2^{d(S)}$  of such circles. In Figure 2 a simple cycle  $S$  is given together with its two Eulerian circles. A simple cycle  $S$  is recovered from its Eulerian circle by merging  $d(S)/2$  pairs of vertices. In the Appendix we prove the following theorem:

**Theorem 3.** *The  $\text{MCPS}_\varphi$  cost of a simple cycle  $S$  is equal to the minimum over all  $\text{MCPS}_{\varphi'}$  costs of its Eulerian circles.*

Given a subroutine computing  $\text{MCPS}_\varphi$  for a circle we can compute  $\text{MCPS}_\varphi$  for every simple cycle of  $G$  using Theorem 3. Then  $\text{MCPS}_\varphi(G)$  can be computed by choosing a MACD of  $G$  maximizing the sum of the costs of its cycles using Theorem 2. In Section 4 we define a particular cost function  $\varphi$  and provide a polynomial time algorithm computing  $\text{MCPS}_\varphi$  for a circle in Section 4.3. Then in Section 5 we show how, using this algorithm as a subroutine, a minimum cost parsimonious DCJ scenario transforming genome  $A$  into  $B$  can be found in polynomial time.

## 4 A colored cost for 2-break scenarios

In this section we partition  $G$ 's vertices into subsets of different colors so as to define a *colored cost function*  $\varphi$  on 2-breaks:  $\varphi((x_1, x_2), (x_3, x_4) \rightarrow (x_{q_1}, x_{q_2}), (x_{q_3}, x_{q_4})) = 0$  if

$$\{\{col(x_1), col(x_2)\}, \{col(x_3), col(x_4)\}\} = \{\{col(x_{q_1}), col(x_{q_2})\}, \{col(x_{q_3}), col(x_{q_4})\}\},$$

and  $\varphi$  is 1 otherwise. For example a 2-break  $(x, y), (z, t) \rightarrow (x, z), (y, t)$  is of zero cost if  $col(x) = col(t)$  or  $col(y) = col(z)$ , otherwise it is of cost 1. This cost function is convenient for use with spacial proximity constraints given by the packing of the chromosomes into the nucleus [21]. See Figure 3 for an example.

In what follows, by a *graph* we will mean a graph together with a coloring  $col$  of its vertices and by *cost* we will mean the cost  $\varphi$  obtained using this coloring. We define a *color-merged graph*  $J(G, col)$  obtained by merging the vertices of  $G$  of the same color. For a 2-break  $G \rightarrow G'$  a transformation  $J(G, col) \rightarrow J(G', col)$  is also a 2-break. This means that a scenario  $\rho$  for  $G$  defines a scenario  $\rho_J$  for  $J(G, col)$ . If the cost of  $G \rightarrow G'$  is 0, then  $J(G, col) = J(G', col)$ , and such a move can be omitted from  $\rho_J$ , leaving us with a scenario of length equal to the cost of  $\rho$ . This observation leads us to the following theorems proven in the Appendix.

**Theorem 4.** *The minimum cost of a scenario for a graph  $G$  is equal to  $l(J(G, col))$ .*

**Theorem 5.** *Deciding for a circle  $O$  and a bound  $k$  whether there exists a scenario of cost at most  $k$  for  $O$  is NP-hard.*

## 4.1 Cycle decomposition of a color-merged graph

Edges of  $G$  and  $J(G, col)$  can be labeled in such a way that for the edges  $(a, b)$  and  $(x, y)$  labeled  $i$  in  $G$  and  $J(G, col)$  we have equality  $(x, y) = (col(a), col(b))$ . We say that such labelings *conform*. For a scenario  $\rho$  of cost  $w$  for  $G$  we will construct a scenario  $\rho_J$  for  $J(G, col)$  of length  $w$  that acts on the same edges as  $\rho$ . We will mimic the process of Section 2.2, taking more care this time when relabeling the edges during the scenario and mapping the black and gray edges at the end of the scenario.

We label the black edges of  $G$  and  $J(G, col)$  with  $\{1, \dots, e(G)\}$  and gray edges with  $\{e(G) + 1, \dots, 2e(G)\}$ . Given a labeling  $L_G$  we obtain a conforming labeling  $L_J$  by merging the vertices of  $G$  having the same color while keeping the labels of the edges. We set  $P_0 = \{\{1\}, \dots, \{e(G)\}\}$ ,  $G_0 = G$  and  $J_0 = J(G, col)$ . Let us fix a 2-break scenario  $\rho$  of cost  $w$  and length  $m$  for  $G$ . We take the  $l$ th 2-break  $(a, b), (c, d) \rightarrow (a, c), (b, d)$  of  $\rho$  transforming  $G_{l-1}$  into  $G_l$  with  $L_{G_{l-1}}(a, b) = i$  and  $L_{G_{l-1}}(c, d) = j$ .

If a 2-break is of cost 1, then we label the newly added edges  $(a, c)$  and  $(b, d)$  with  $i$  and  $j$  respectively to obtain  $L_{G_l}$  and merge the subsets in  $P_{l-1}$  containing  $i$  and  $j$  to obtain  $P_l$ . In  $J(G_{l-1}, col)$  we replace the edges labeled  $i$  and  $j$  by an edge  $(col(a), col(c))$  labeled  $i$  and  $(col(b), col(d))$  labeled  $j$  to obtain a labeling  $L_{J_l}$  of  $J(G_l, col)$  that conforms to  $L_{G_l}$ . If a 2-break is of cost 0, then without loss of generality we can suppose that  $col(a) = col(d)$ . In this case we label the newly added edges  $(b, d)$  and  $(a, c)$  with  $i$  and  $j$  respectively to obtain  $L_{G_l}$ .  $J(G_{l-1}, col) = J(G_l, col)$  and  $L_{G_l}$  is chosen in such a way that  $L_{J_{l-1}}$  and  $L_{G_l}$  still conform, thus we keep  $L_{J_l} = L_{J_{l-1}}$  and  $P_l = P_{l-1}$ .

At the end of the scenario we obtain a partition  $P_m$  of  $\{1, \dots, e(G)\}$ . Since  $G_m$  is terminal, we can map the subsets  $\{1, \dots, e(G)\}$  and  $\{e(G) + 1, \dots, 2e(G)\}$  one-to-one in a way that for a pair  $(i, j)$  of mapped labels the black edge  $i$  and gray edge  $j$  have the same endpoints. For  $j \in \{e(G) + 1, \dots, 2e(G)\}$  we include it into a subset of  $P_m$  containing a label to which  $j$  is mapped. This way a partition  $R_m$  of a set  $\{1, \dots, 2e(G)\}$  is obtained. Only the 2-breaks of cost 1 of  $\rho$  modify the colored-merged graph. This provides us with a 2-break scenario  $\rho_J$  of length  $w$  for  $J(G, col)$ . In addition to that  $R_m$ , when seen as a partition of the edges of  $J(G, col)$ , is exactly the cycle decomposition  $C(\rho_J)$  of the scenario  $\rho_J$ . In Section 4.2 we study the structure of  $C(\rho_J)$  for a parsimonious scenario  $\rho$ .

## 4.2 MCPS for a circle

For a circle  $O$  we take conforming labelings  $L_O$  and  $L_J$  of  $O$  and  $J(O, col)$ . For a subset  $S$  of edges of  $J(O, col)$  (resp.  $O$ ) we define a subset  $O(S)$  (resp.  $J(S)$ ) of edges of  $O$  (resp.  $J(O, col)$ ) labeled with the same labels. To a scenario  $\rho$  for  $O$  we have associated a scenario  $\rho_J$  for  $J(O, col)$  and its cycle decomposition  $C(\rho_J)$  in Section 4.1.

**Definition 3** (Crossing subsets). *Two disjoint subsets of edges  $S_1$  and  $S_2$  of a circle  $O$  cross if there are edges  $e_1, e_2$  in  $S_1$  and  $f_1, f_2$  in  $S_2$  such that a path in  $O$  joining  $e_1$  and  $e_2$  contains exactly one of the edges  $f_1$  or  $f_2$ . A cycle decomposition  $C$  of  $J(O, col)$  is said to be non-crossing if none of the subsets of  $O(C)$  cross.*

**Theorem 6.**  *$C(\rho_J)$  is non-crossing for a parsimonious scenario  $\rho$  for  $O$ .*

*Proof.* If the theorem is false, then the set of the circles for which there exists a parsimonious scenario contradicting the theorem is non-empty. Let us take a circle  $O^*$  in this set having the minimum number of edges and a scenario  $\rho^*$  for  $O^*$  such that  $O(C(\rho_J^*))$  is crossing.  $O^*$  has at least 2 black edges as otherwise  $C(\rho_J^*)$  contains a single subset.

Using Theorem 1 and the structure of a circle we get that a 2-break of a parsimonious scenario for a union of vertex-disjoint circles transforms one of its circles into a union of two vertex-disjoint circles. This means that the first 2-break of  $\rho^*$  replaces edges  $i$  and  $j$  and transforms  $O^*$  into a union of two smaller circles  $\bar{O}$  and  $\hat{O}$ . The following 2-breaks of  $\rho^*$  replace 2 edges with labels belonging to either  $\bar{O}$  or  $\hat{O}$ , which provides us with the parsimonious scenarios  $\bar{\rho}$  for  $\bar{O}$  and  $\hat{\rho}$  for  $\hat{O}$ . By the minimality of  $O^*$  we get that  $C(\bar{\rho}_{\bar{j}})$  and  $C(\hat{\rho}_{\hat{j}})$  are non-crossing.  $C(\rho_j^*)$  can be easily obtained from  $C(\bar{\rho}_{\bar{j}})$  and  $C(\hat{\rho}_{\hat{j}})$  by taking their union and then merging the subsets of edges including  $i$  and  $j$  if the first 2-break of  $\rho^*$  is of cost 1. Now it is easy to check that  $C(\rho_j^*)$  is non-crossing, a contradiction.  $\square$

**Definition 4** (Arc). *An arc is an alternating path joining two vertices of the same color, called endpoints, and having the same number of black and gray edges. Edges of an arc adjacent to its endpoints are called ends and one of them is black and another is gray.*

A circle with  $n$  vertices has  $n$  arcs if all the vertices are of different colors and  $n^2/2$  arcs if all the vertices share the same color. We say that two arcs do not *overlap* if they are edge-disjoint or one is included in another but none of their ends coincide. A set of pairwise non-overlapping arcs will be called an *independent* set of arcs. A MAXIMUM INDEPENDENT SET OF ARCS (MISA) of  $O$  is an independent subset of arcs of  $O$  of maximum cardinality. For a set of arcs  $S$  a *maximal arc* is an arc that is not included in any other arc, while a *minimal arc* is an arc that does not include any other arc. In the Appendix Lemma 2 is proven, leading us to Theorem 7.

**Lemma 2.** *The size of a maximum non-crossing cycle decomposition of  $J(O, col)$  is equal to the size of a MISA of  $O$ .*

**Theorem 7.** *The MCPS cost for a circle  $O$  is equal to  $e(O) - |I|$  for a MISA  $I$  of  $O$ .*

*Proof.* We first show that there exists a parsimonious scenario for  $O$  of cost at most  $e(O) - |I|$ . A parsimonious scenario for  $O$  is of length  $e(O) - 1$ , thus of cost at most  $e(O) - 1$ , due to Theorem 1. This means that if  $|I| = 1$ , then inequality is trivial. Otherwise take a minimal arc  $U$  in  $I$ . Perform a 2-break replacing the black end of  $U$  and the black edge not belonging to  $U$  adjacent to the gray end of  $U$  and transforming  $O$  into a union of two circles. One of these circles is of length equal to the length of  $U$  and we call it  $\hat{O}_1$ . Another is such that its MISA is of size  $|I| - 1$  and we call it  $O_1$ . The cost of this 2-break is 0 as the endpoints of an arc have the same color. We iterate this procedure for  $O_1$  until we end up with  $|I|$  circles  $O_{|I|-1}, \hat{O}_1, \dots, \hat{O}_{|I|-1}$ . A parsimonious scenario for this set of circles is of length  $e(O) - |I|$  and thus of cost at most  $e(O) - |I|$ .

On the other hand, we take a MCPS scenario  $\rho$  of cost  $w$  for a circle  $O$ . A cycle decomposition  $C(\rho_J)$  of  $J(O, col)$  is of size at least  $e(O) - w$  using Lemma 1. Using Theorem 6 we obtain that  $C(\rho_J)$  is non-crossing and using Lemma 2 we obtain an independent set of arcs of size  $|C(\rho_J)|$  and thus  $w \geq e(O) - |C(J(\rho))| \geq e(O) - |I|$  where  $I$  is a MISA.  $\square$

### 4.3 A polynomial time algorithm for MCPS on a circle

Fix a vertex  $v$  of a circle  $O$  having  $n$  vertices. We say that an arc  $U$  *crosses*  $v$  if  $v$  is in  $U$  without being its endpoint. We replace  $v$  by two vertices of the same color transforming  $O$  into an alternating path  $P$ . MISA sizes of  $P = (v_0, \dots, v_l)$  and  $O$  are equal due to Lemma 3, proven in the Appendix.

**Lemma 3.** *There exists a MISA  $I'$  of  $O$  such that there is no arc crossing  $v$  in  $I'$ .*

For  $0 \leq i \leq j \leq l$   $MISA(i, j)$  denotes the size of a MISA of an alternating path  $(v_i, \dots, v_j)$ ,  $MISA(i, i) = 0$  for all  $i$ . We say that  $i$  and  $j$  are *compatible* if  $v_i$  and  $v_j$  are the endpoints of an arc. Take  $I$  a MISA of a path  $(v_i, \dots, v_j)$  for  $j \geq i$ . Vertex  $v_j$  belongs to at most one arc in  $I$  and this



creates three possibilities. If  $v_j$  does not belong to any arc in  $I$ , then  $|I| = \text{MISA}(i, j - 1)$ . If an arc with the endpoints  $v_i$  and  $v_j$  is in  $I$ , then  $|I| = \text{MISA}(i + 1, j - 1) + 1$ . If an arc with the endpoints  $v_k$  and  $v_j$  is in  $I$  with  $i < k < j$ , then  $|I| = \text{MISA}(i, k) + \text{MISA}(k, j)$ . This leads to the following recurrence:

$$\text{MISA}(i, j) = \max \begin{cases} \text{MISA}(i, j - 1), \\ \text{MISA}(i + 1, j - 1) + 1, & \text{if } i \text{ and } j \text{ compatible} \\ \text{MISA}(i, k) + \text{MISA}(k, j), & \text{for } k, i < k < j \text{ compatible with } j, \end{cases}$$

which provides us with a dynamic program with time complexity  $O(n^3)$ . It is easy to modify the algorithm to give a MISA  $I$  of  $O$ . From  $I$  it we can obtain a MCPS scenario for  $O$  using Theorem 7. Since  $I$  is of size  $O(n)$  this can easily be done in  $O(n^2)$  time.

We partition the vertices of  $P$  into the subsets  $S_1, \dots, S_m$  of pairwise compatible vertices and set  $s_i = |S_i|$ . One can show that our dynamic program computes  $\text{MISA}(0, l)$  in  $cn(s_1^2 + \dots + s_m^2)$  steps for some constant  $c$ . If the subsets are of equal sizes then the number of steps is  $cn^3/m$ . This means that the best-case time-complexity of our program is  $O(n^2)$ .

## 5 DCJ scenarios for genomes

A *genome* consists of *chromosomes* that are linear or circular orders of genes separated by potential *breakpoint* regions. In Figure 3 the tail of an arrow represents the *tail extremity*, and the head of an arrow represents the *head extremity* of a gene. We can represent a genome by a set of *adjacencies* between the gene extremities. In Figure 3 this set is  $\{\{1t\}, \{1h, 2t\}, \{2h, 3h\}, \{3t\}\}$  for genome  $A$  and  $\{\{1t\}, \{1h, 2h\}, \{2t, 3h\}, \{3t\}\}$  for genome  $B$ . An *adjacency* is either an unordered pair of the extremities that are adjacent on a chromosome, called *internal* adjacency, or a single extremity adjacent to one of the two ends of a linear chromosome, called an *external* adjacency. In what follows we will suppose two genomes  $A$  and  $B$  that share the same genes, and our goal will be to transform  $A$  into  $B$  using a sequence of DCJs.

**Definition 5** (Double cut and join). *A DCJ cuts one or two breakpoint regions and joins the resulting ends of the chromosomes back in one of the four following ways:  $\{a, b\}, \{c, d\} \rightarrow \{a, c\}, \{b, d\}$ ;  $\{a, b\}, \{c\} \rightarrow \{a, c\}$ ;  $\{a, b\} \rightarrow \{a\}, \{b\}$ ;  $\{a\}, \{b\} \rightarrow \{a, b\}$ .*

We partition the gene extremities into subsets of different colors.  $\text{col}(a)$  denotes the color of a gene extremity  $a$ . The colored internal adjacency  $\{a, b\}$  is  $\{\text{col}(a), \text{col}(b)\}$  and the colored external adjacency  $\{a\}$  is  $\{\text{col}(a), \circ\}$ , where  $\circ$  does not coincide with any of the colors of the gene extremities. A DCJ  $A \rightarrow A'$  is said to be of zero cost if the sets of colored adjacencies of  $A$  and  $A'$  are equal. It is of cost 1 otherwise. For example  $\{a, b\}, \{c\} \rightarrow \{a, c\}, \{b\}$  is of cost 0 if  $\{\{\text{col}(a), \text{col}(b)\}, \{\text{col}(c), \circ\}\} = \{\{\text{col}(a), \text{col}(c)\}, \{\text{col}(b), \circ\}\}$ , that is if  $\text{col}(b) = \text{col}(c)$ . The cost of a DCJ scenario is the sum of the costs of its rearrangements.

In [3], a linear time algorithm for finding a parsimonious DCJ scenario was proposed. The algorithm is based on the analysis of the connected components of the *adjacency graph*. Here, we use a slightly different structure associated to a genome pair  $(A, B)$  called the *breakpoint graph* [16, 1, 7].

**Definition 6** (Breakpoint graph).  *$G(A, B) = (V, E^b \cup E^g)$  for genomes  $A$  and  $B$ , sharing  $n$  genes, is a 2-edge-colored Eulerian undirected multi-graph.  $V$  consists of  $2n$  gene extremities and an additional vertex  $\circ$ . For every internal adjacency  $\{a, b\} \in A$  (resp.  $\{a, b\} \in B$ ) there is a black (resp. gray) edge  $(a, b)$  in  $G(A, B)$  and for every external adjacency  $\{a\} \in A$  (resp.  $\{a\} \in B$ ) there is a black (resp. gray) edge  $(a, \circ)$  in  $G(A, B)$ . We add additional black and gray loops  $(\circ, \circ)$  to obtain  $d^b(\circ) = d^g(\circ) = 2n$ . The breakpoint graph  $G(A, B)$  for the genomes from Figure 3 is given in Figure 4.*

**Lemma 4.** For the DCJ scenarios transforming genome  $A$  into  $B$  the minimum length is  $l(G(A, B)) = e(G(A, B)) - c(G(A, B))$ , the minimum cost is  $l(J) = e(J) - c(J)$  with  $J = J(G(A, B), col)$  and the minimum cost of a parsimonious DCJ scenario is  $MCPS(G(A, B))$ .

**Theorem 8.** MCPS is polynomial-time solvable for a breakpoint graph  $G(A, B)$ .

*Proof.* Take genomes  $A$  and  $B$  sharing  $n$  genes. For all the vertices  $v \neq \circ$  we have  $d^g(G(A, B), v) = d^b(G(A, B), v) = 1$ . From this we obtain that for an edge belonging to a circle this is the only simple cycle in  $G(A, B)$  including this edge. Thus a MACD of  $G(A, B)$  includes all of its circles. These set aside, we are left with  $G(A, B)'$ , which is a union of alternating paths starting and ending at  $\circ$  and having the end edges of the same color. If this color is black we call a path  $AA$ , and  $BB$  otherwise. Every simple cycle of  $G(A, B)'$  is a union of a  $BB$  path and a  $AA$  path. We proceed by constructing a complete bipartite graph  $H$  having  $AA$  and  $BB$  paths as vertices. An edge joining paths  $a$  and  $b$  is assigned the weight equal to the MCPS cost of  $a \cup b$ .

**Lemma 5.** The weights of the edges of  $H$  can be assigned in  $O(n^4)$  time.

Lemma 5 is proven in the Appendix. We proceed by computing a maximum weight matching for  $H$  to obtain a MACD of  $G(A, B)'$  minimizing MCPS cost, this can be done in  $O(n^3)$  time using Hungarian algorithm. The MCPS cost of  $G(A, B)$  is obtained by adding the costs of the circles and the cost of  $G(A, B)'$ . In this way we obtain a  $O(n^4)$  time algorithm for computing the MCPS cost, which it can be easily modified to give a MCPS scenario.  $\square$

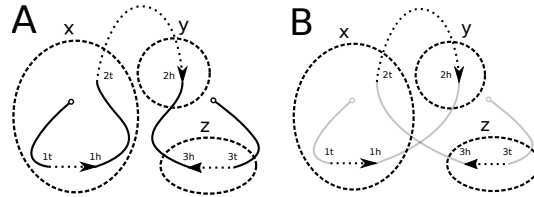


Figure 3: Genomes  $A$  and  $B$  share genes 1, 2 and 3 and both consist of a single linear chromosome. Their gene extremities are colored in three colors  $x, y$  and  $z$  according to spacial proximity. The DCJ  $\{1_h, 2_t\}, \{2_h, 3_h\} \rightarrow \{1_h, 2_h\}, \{2_t, 3_h\}$  transforming  $A$  into  $B$  is of cost 1 because  $\{\{x, x\}, \{y, z\}\} \neq \{\{x, y\}, \{x, z\}\}$ . The DCJ  $\{2_t, 1_h\}, \{1_t\} \rightarrow \{2_t, 1_t\}, \{1_h\}$  on genome  $A$  or  $B$  would be of cost 0.

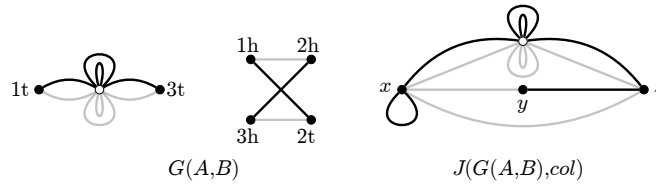


Figure 4: Breakpoint graph  $G(A, B)$  and its color-merged graph  $J = J(G(A, B), col)$  for the genomes and a coloring of their gene extremities given in Figure 3.  $e(J) = 6$  and  $c(J) = 5$ .

## 6 Conclusions and further work

### 6.1 Sorting by mathematical transpositions

Finding a parsimonious 2-break scenario on a circle is closely related to the problem of sorting a circular permutation by mathematical transpositions. For a permutation  $\pi$  of  $\{1, \dots, n\}$  we define a digraph  $D(\pi)$  with vertices  $i$  and directed edges  $(i, \pi(i))$ . A transposition on  $\pi$  defines a 2-break on  $D(\pi)$  as illustrated in Figure 5. The problem of finding a minimum cost parsimonious scenario of transpositions for permutations with a cost function  $\varphi$  defined for every pair of elements of  $V$  was treated in [10]. Such a  $\varphi$  defines a natural cost function  $\varphi'$  for the 2-breaks on  $V$  such that  $\varphi'(\{a, b\}, \{c, d\} \rightarrow \{a, c\}, \{b, d\}) = \min(\varphi(a, d), \varphi(b, c))$ . The cost function used in Section 4 is precisely of such a type. The paper described a polynomial-time algorithm for a minimum cost parsimonious scenario of transpositions for a general cost function  $\varphi$ . We speculate that this algorithm can be adapted to obtain a polynomial algorithm for  $\text{MCPS}_{\varphi'}$  on a circle. Another problem of interest is finding a minimum cost scenario among the scenarios of length smaller than some fixed length. This would be an important step towards finding more realistic evolutionary scenarios, since the most likely scenario may not always be of minimum length.



Figure 5: For a permutation  $\pi = 2341$  digraphs  $D(\pi)$  and  $D((24)\pi)$  are given. A transformation  $D(\pi) \rightarrow D((24)\pi)$  is a 2-break  $(1, 2), (3, 4) \rightarrow (1, 4), (3, 2)$ .

### 6.2 Conclusions

The DCJ models on genes that are oriented, or unoriented [8], have insertions and deletions [25], or have segmental duplications [27], are all intimately tied to the breakpoint graph. They all can be easily formulated in our setting of 2-edge-colored Eulerian multi-graphs and 2-break scenarios; all of our results about cycle decompositions and MCPS apply directly in these cases. We showed one example of how to use our work algorithmically, but we expect our framework to lead to further algorithmic results on general cost functions and general DCJ distances that consider unequal content.

## References

- [1] V Bafna and PA Pevzner. Genome rearrangements and sorting by reversals. In *Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on*, pages 148–157. IEEE, 1993.
- [2] M.A. Bender, D. Ge, S. He, H. Hu, R.Y. Pinter, S. Skiena, and F. Swidan. Improved bounds on sorting by length-weighted reversals. *J. of Comp. and System Sciences*, 74(5):744–774, 2008.
- [3] Anne Bergeron, Julia Mixtacki, and Jens Stoye. *A Unifying View of Genome Rearrangements*, pages 163–173. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.
- [4] Marília DV Braga, Eyla Willing, and Jens Stoye. Double cut and join with insertions and deletions. *Journal of Computational Biology*, 18(9):1167–1184, 2011.

- [5] Horst Bunke and Kaspar Riesen. Graph edit distance–optimal and suboptimal algorithms with applications. *Analysis of Complex Networks*.
- [6] Alberto Caprara. Sorting by reversals is difficult. In *Proceedings of the first annual international conference on Computational molecular biology*, pages 75–83. ACM, 1997.
- [7] Alberto Caprara. Sorting permutations by reversals and eulerian cycle decompositions. *SIAM journal on discrete mathematics*, 12(1):91–110, 1999.
- [8] Xin Chen. On sorting unsigned permutations by double-cut-and-joins. *Journal of Combinatorial Optimization*, 25(3):339–351, Apr 2013. URL: <https://doi.org/10.1007/s10878-010-9369-8>, doi:10.1007/s10878-010-9369-8.
- [9] Nadia El-Mabrouk. Genome rearrangement by reversals and insertions/deletions of contiguous segments. In *Annual Symposium on Combinatorial Pattern Matching*, pages 222–234. Springer, 2000.
- [10] Farzad Farnoud (Hassanzadeh) and Olgica Milenkovic. Sorting of permutations by cost-constrained transpositions. *IEEE Trans. Inf. Theor.*, 58(1):3–23, January 2012. URL: <http://dx.doi.org/10.1109/TIT.2011.2171532>, doi:10.1109/TIT.2011.2171532.
- [11] Guillaume Fertin, Anthony Labarre, Irena Rusu, Eric Tannier, and Stéphane Vialette. *Combinatorics of Genome Rearrangements*. The MIT Press, 2009.
- [12] Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. A survey of graph edit distance. *Pattern Analysis and applications*, 13(1):113–129, 2010.
- [13] Dan Gusfield. *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge university press, 1997.
- [14] Sridhar Hannenhalli and Pavel A Pevzner. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. *Journal of the ACM (JACM)*, 46(1):1–27, 1999.
- [15] Ian Holyer. The NP-completeness of some edge-partition problems. *SIAM Journal on Computing*, 10(4):713–717, 1981.
- [16] John Kececioğlu and David Sankoff. Exact and approximation algorithms for the inversion distance between two chromosomes. In *Annual Symposium on Combinatorial Pattern Matching*, pages 87–105. Springer, 1993.
- [17] Joseph B Kruskal and David Sankoff. *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Addison-Wesley, 1983.
- [18] Martin McGrane and Michael A Charleston. Biological network edit distance. *Journal of Computational Biology*, 23(9):776–788, 2016.
- [19] Gonzalo Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001.
- [20] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

- [21] Sylvain Pulicani, Pijus Simonaitis, Eric Rivals, and Krister M Swenson. Rearrangement scenarios guided by chromatin structure. In *RECOMB International Workshop on Comparative Genomics*, pages 141–155. Springer, 2017.
- [22] Alberto Sanfeliu and King-Sun Fu. A distance measure between attributed relational graphs for pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (3):353–362, 1983.
- [23] David Sankoff. Matching sequences under deletion/insertion constraints. *Proceedings of the National Academy of Sciences*, 69(1):4–6, 1972.
- [24] David Sankoff. Edit distance for genome comparison based on non-local operations. In *Annual Symposium on Combinatorial Pattern Matching*, pages 121–135. Springer, 1992.
- [25] Mingfu Shao and Yu Lin. Approximating the edit distance for genomes with duplicate genes under DCJ, insertion and deletion. *BMC Bioinformatics*, 13(19):S13, Dec 2012. URL: <https://doi.org/10.1186/1471-2105-13-S19-S13>, doi:10.1186/1471-2105-13-S19-S13.
- [26] Mingfu Shao, Yu Lin, and Bernard ME Moret. An exact algorithm to compute the double-cut-and-join distance for genomes with duplicate genes. *Journal of Computational Biology*, 22(5):425–435, 2015.
- [27] Mingfu Shao and Bernard ME Moret. Comparing genomes with rearrangements and segmental duplications. *Bioinformatics*, 31(12):i329–i338, 2015.
- [28] Pijus Simonaitis and Krister M. Swenson. Finding Local Genome Rearrangements. In Russell Schwartz and Knut Reinert, editors, *17th International Workshop on Algorithms in Bioinformatics (WABI 2017)*, volume 88 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 24:1–24:13, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. URL: <http://drops.dagstuhl.de/opus/volltexte/2017/7660>, doi:10.4230/LIPIcs.WABI.2017.24.
- [29] Krister M. Swenson, Pijus Simonaitis, and Mathieu Blanchette. Models and algorithms for genome rearrangement with positional constraints. *Algorithms for Molecular Biology*, 11(1):13, 2016. URL: <http://dx.doi.org/10.1186/s13015-016-0065-9>, doi:10.1186/s13015-016-0065-9.
- [30] G.A. Watterson, W.J. Ewens, T.E. Hall, and A. Morgan. The chromosome inversion problem. *J. Theoretical Biology*, 99:1–7, 1982.
- [31] S. Yancopoulos, O. Attie, and R. Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346, 2005.

## .1 Lemma 1

**Lemma.** For a vertex  $v$  and a subset  $Q \in R_l$  with  $l \in \{0, \dots, m\}$  we have

$$\bar{d}(s(G_0, Q), v) = \bar{d}(s(G_l, Q), v)$$

*Proof.* Equality is true for  $l = 0$ . Suppose that equality is true for every  $Q$  and  $v$  with  $l - 1$  and proceed by induction on  $l > 0$ . Fix a vertex  $v$  and a subset  $Q \in R_l$ . The  $l$ th k-break replaces the edges labeled  $e_1, \dots, e_k$ . By construction, these edges belong to the same subset  $Q' \in R_l$ . There are two possibilities:

- ( $Q' \neq Q$ ) In this case  $Q \in R_{l-1}$  and  $s(G_l, Q) = s(G_{l-1}, Q)$ , as the edges in  $Q$  are unaffected by the  $l$ th k-break. Using the inductive hypothesis we obtain

$$\bar{d}(s(G_l, Q), v) = \bar{d}(s(G_{l-1}, Q), v) = \bar{d}(s(G_0, Q), v)$$

- ( $Q = Q'$ ) In this case  $Q' = \bigcup_{i=1}^k Q_i$  where  $e_i \in Q_i \in R_{l-1}$ . There may exist  $i, j \in \{1, \dots, k\}$  such that  $Q_i = Q_j$ , thus we select such  $q_1, \dots, q_r \in \{1, \dots, k\}$  so that  $Q = Q' = \bigcup_{i=1}^r Q_{q_i}$  and all of these subsets are different. To begin with,

$$\bar{d}(s(G_l, Q), v) = \bar{d}(s(G_{l-1}, Q), v) = \sum_{i=1}^r \bar{d}(s(G_{l-1}, Q_{q_i}), v).$$

The first equality is due to the fact that a graph  $s(G_l, Q)$  is obtained from  $s(G_{l-1}, Q)$  by a k-break and k-break does not modify the degrees of the vertices in a graph. The second equality is guaranteed since no two  $Q_{q_1}, \dots, Q_{q_r}$  intersect. Then

$$\sum_{i=1}^r \bar{d}(s(G_{l-1}, Q_{q_i}), v) = \sum_{i=1}^r \bar{d}(s(G_0, Q_{q_i}), v) = \bar{d}(s(G_0, Q), v).$$

With the first equality following from the inductive hypothesis and the latter once again since no two  $Q_{q_1}, \dots, Q_{q_r}$  intersect.

As equality is preserved by a k-break, and true for  $l = 0$ , we obtain the result by induction.  $\square$

## .2 Theorem 3

**Theorem.** The  $\text{MCPS}_\varphi$  cost of a simple cycle  $S$  is equal to the minimum over all  $\text{MCPS}_{\varphi'}$  costs of its Eulerian circles.

*Proof.* Take a simple cycle  $S$  with  $d(S) = k$ . If  $k = 0$ , then  $S$  is its own Eulerian circle. If  $k > 0$ , then we choose a vertex  $v$  satisfying  $d^b(S, v) = d^g(S, v) = 2$  and construct a graph with  $v$  replaced by two vertices  $v_1$  and  $v_2$  satisfying  $d^b(S, v_i) = d^g(S, v_i) = 1$ . We start with a copy of  $S$  from which we remove  $v$  with the adjacent edges and add two new vertices  $v_1$  and  $v_2$  to obtain  $S'$ . If there is a black loop  $(v, v)$  in  $S$  then we add a black edge  $(v_1, v_2)$  to  $S'$ . Otherwise there are two black edges  $(v, u_1)$  and  $(v, u_2)$  in  $S$  with possibly  $u_1 = u_2$ . We add black edges  $(v_1, u_1)$  and  $(v_2, u_2)$  to  $S'$ . If there is a gray loop  $(v, v)$  in  $S$ , then we add a gray edge  $(v_1, v_2)$  to obtain a graph  $S_0$ . If there are two gray edges  $(v, z_1)$  and  $(v, z_2)$  in  $S$ , then we construct two graphs  $S_1$  and  $S_2$ .  $S_1$  is obtained by adding gray edges  $(z_1, v_1)$  and  $(z_2, v_2)$  to  $S'$  and  $S_2$  by adding gray edges  $(z_1, v_2)$  and  $(z_2, v_1)$  to  $S'$ . In Figure 2 the graphs  $S_1$  and  $S_2$  are given for a simple cycle  $S$ .

$S_0, S_1, S_2$  have the same set of vertices  $V'$ . We set  $\bar{v}_1 = \bar{v}_2 = v$  and  $\bar{u} = u$  for other vertices in  $V'$ . We define a cost function  $\varphi'$  for the 2-breaks on  $V'$  as follows:

$$\varphi'((x_1, x_2), (x_3, x_4) \rightarrow (x_1, x_3), (x_2, x_4)) = \varphi((\bar{x}_1, \bar{x}_2), (\bar{x}_3, \bar{x}_4) \rightarrow (\bar{x}_1, \bar{x}_3), (\bar{x}_2, \bar{x}_4)).$$

From  $S_i$  we obtain  $S$  by merging the vertices  $v_1$  and  $v_2$ . We will show that  $\text{MCPS}_\varphi(S)$  of  $S$  is equal to the  $\text{MCPS}_{\varphi'}(S_0)$  or the minimum of  $\text{MCPS}_{\varphi'}(S_1)$  and  $\text{MCPS}_{\varphi'}(S_2)$  depending on whether  $S$  contains a gray loop  $(v, v)$  or not.  $c(S_i) = 1$  and  $e(S_i) = e(S)$ , thus the lengths of their parsimonious scenarios are equal due to Theorem 1 .

For every 2-break transforming  $S_i$  to  $S'_i$  there is a unique 2-break on  $S$  of the same cost transforming  $S$  into a graph that we obtain from  $S'_i$  by merging  $v_1$  and  $v_2$ . If we take a MCPS scenario for  $S_i$  we obtain a parsimonious scenario for  $S$  of the same cost. This means that  $\text{MCPS}_\varphi(S)$  is smaller or equal to  $\text{MCPS}_{\varphi'}(S_0)$  or the minimum of  $\text{MCPS}_{\varphi'}(S_1)$  and  $\text{MCPS}_{\varphi'}(S_2)$  depending on whether  $S$  has a gray loop  $(v, v)$  or not.

On the other hand for a 2-break transforming  $S$  into  $S'$  there exists a 2-break of the same cost transforming  $S_i$  into such  $S'_i$  that by merging its vertices  $v_1$  and  $v_2$  we obtain  $S'$ . A MCPS scenario for  $S$  provides us with a sequence of 2-breaks  $\rho$  on  $S_i$  of the same cost transforming it into a graph  $\hat{S}_i$  from which we obtain a terminal graph by merging  $v_1$  and  $v_2$ .

The structure of  $\hat{S}_i$  is fairly simple and the two possible cases can be checked by hand. If  $S$  contains a gray loop  $(v, v)$  then  $\hat{S}_0$  must be already terminal as it is Eulerian and there is a single gray edge adjacent to  $v_0$  or  $v_1$ . In this case a MCPS scenario for  $S$  and  $\varphi$  provides us with a parsimonious scenario of the same cost for  $S_0$  and  $\varphi'$ .

Sets of black edges of  $S_1$  and  $S_2$  are equal. This means that a MCPS scenario for  $S$  provides us with a single sequence  $\rho$  of 2-breaks of the same cost transforming  $S_1$  into  $\hat{S}_1$  and  $S_2$  into  $\hat{S}_2$  such that by merging  $v_1$  and  $v_2$  we obtain terminal graphs. The sets of black edges of  $\hat{S}_1$  and  $\hat{S}_2$  stay equal. If  $\hat{S}_1$  is already terminal, then we are done. Otherwise  $\hat{S}_1$  is a union of a terminal graph and a cycle of length 2 containing gray edges  $(z_1, v_1)$  and  $(z_2, v_2)$  and black edges  $(z_2, v_1)$  and  $(z_1, v_2)$ , however this means that  $\hat{S}_2$  is terminal. Thus in this case  $\text{MCPS}_\varphi(S)$  is equal to the minimum of  $\text{MCPS}_{\varphi'}(S_1)$  and  $\text{MCPS}_{\varphi'}(S_2)$ .

Either  $S_i$  are circles and we are done or we can proceed by choosing another vertex  $v$ . At the end we obtain Eulerian circles of  $S$  and  $\text{MCPS}_\varphi(S)$  is equal to the minimum of the MCPS costs for these circles.  $\square$

### .3 Theorem 4

We start by proving an auxiliary lemma.

**Lemma.** *If  $J(G, col)$  is terminal, then there exists a zero cost 2-break scenario for  $G$ .*

*Proof.* We denote the maximum number of length 1 cycles in a cycle decomposition of  $G$  by  $c_1(G)$ . If  $c_1(G) = e(G)$ , then  $G$  is terminal and we are done. Otherwise we demonstrate how we can transform  $G$  with a sequence of zero cost 2-breaks into  $G'$  with  $c_1(G') > c_1(G)$ . When iterated this provides us with a zero cost scenario for  $G$ .

Take a cycle decomposition  $C$  of  $G$  having  $c_1(G)$  length 1 cycles and remove these  $c_1(G)$  cycles from  $G$  obtaining  $\bar{G}$ . Take a black edge  $(u, v)$  of  $\bar{G}$ .  $J(\bar{G}, col)$  is terminal, thus there exists a gray edge  $(col(u), col(v))$  in  $J(\bar{G}, col)$ . This means that there is a gray edge  $(u', v')$  in  $\bar{G}$  with  $col(u) = col(u')$  and  $col(v) = col(v')$ .  $G$  is Eulerian thus  $\bar{G}$  is also Eulerian and there are gray edges  $(p, u')$  and  $(r, v')$  in  $\bar{G}$ .

- If  $(u, v)$  equals one of these edges, let's say  $(p, u')$ , then  $col(p) = col(v) = col(v')$ . This means that a 2-break  $(p, u'), (r, v') \rightarrow (p, r), (u', v')$  transforming  $\bar{G}$  into  $\bar{G}'$  is of zero cost and creates a length 1 cycle.
- If  $(u, v)$  does not coincide with  $(p, u')$  or  $(r, v')$ , then a 2-break  $(u, v), (p, u') \rightarrow (u, p), (v, u')$  is of zero cost as  $col(u) = col(u')$ . So is a 2-break  $(v, u'), (v', r) \rightarrow (v, r), (u', v')$  as  $col(v) = col(v')$ .

These 2-breaks transforming  $\bar{G}$  into  $\bar{G}'$  are of zero cost and create a length 1 cycle in  $\bar{G}'$ . Once the length 1 cycles deleted from  $G$  are reintroduced to  $\bar{G}'$  we obtain a graph  $G'$  with  $c_1(G') > c_1(G)$  and a sequence of zero cost 2-breaks transforming  $G$  into  $G'$ .  $\square$

**Theorem.** *The minimum cost of a scenario for a graph  $G$  is equal to  $l(J(G, col))$ .*

*Proof.* For a 2-break  $G \rightarrow G'$  a transformation  $J(G, col) \rightarrow J(G', col)$  is also a 2-break. If the cost of  $G \rightarrow G'$  is 0, then  $J(G, col) = J(G', col)$ . This means that for a scenario of cost  $w$  for  $G$  there exists a scenario of length  $w$  for  $J(G, col)$  and thus  $l(J(G, col)) \leq w$ . On the other hand, for every 2-break  $J(G, col) \rightarrow J'$  a 2-break  $G \rightarrow G'$  can be found such that  $J(G', col) = J'$ . For  $J(G, col)$  scenario of length  $l = l(J(G, col))$  we obtain a 2-break sequence of length  $l$ , thus of cost at most  $l$ , transforming  $G$  into such  $G'$  that  $J(G', col)$  is terminal. Using the previously proven lemma we get a scenario for  $G$  of cost at most  $l$  establishing  $l(J(G, col)) \geq w$ .  $\square$

#### .4 Theorem 5

**Theorem.** *Deciding for a circle  $O$  and a bound  $k$  whether there exists a scenario of cost at most  $k$  for  $O$  is NP-hard.*

*Proof.* The problem is clearly in NP. We reduce the decision version of a maximum cycle decomposition on simple Eulerian graphs, which is NP-hard [15], to our problem. Without loss of generality, take an instance  $G = (V, E)$  and a bound  $k$ , where  $G$  is Eulerian and connected. Consider an Eulerian cycle  $(u_1, u_2, \dots, u_n, u_1)$  of  $G$  and construct a circle  $O = (v_1, \dots, v_{2n}, v_1)$  with gray edges  $(v_{2i-1}, v_{2i})$ , black edges  $(v_{2i}, v_{2i+1})$  and  $col(v_{2i-1}) = col(v_{2i}) = u_i$  for  $i \in \{1, \dots, n\}$ . The gray edges of  $J(O, col)$  are loops and black edges of  $J(O, col)$  are exactly the edges of  $G$ . There is a scenario of cost at most  $e(G) - k$  for  $O$  if and only if there exists an alternating cycle packing of  $J(O, col)$  of size at least  $k$  due to Theorem 4 and Theorem 1. And the latter is true if and only if  $G$  admits a maximum cycle decomposition of size least  $k$ .  $\square$

#### .5 Lemma 2

We start by proving an auxiliary lemma.

**Lemma.** *Every edge of a circle  $O$  is included in at least one arc of its MISA  $I$ .*

*Proof.* Let us direct the edges of  $O$  to obtain a directed cycle and suppose by contradiction that an edge  $e$  entering a vertex of color  $x$  does not belong to any arc in  $I$ . Without loss of generality we can suppose that the color of  $e$  is black. The number of gray edges leaving the vertices of color  $x$  is equal to the number of black edges entering the vertices of color  $x$  in  $O$  and in any arc of  $O$ . This is true due to the constraint that an arc must have the vertices of the same color as its endpoints and the edges of different colors as its ends. This means that there is a gray edge  $f$  in  $O$  not included in any arcs of  $I$  leaving a vertex of color  $x$ . We define a new arc starting and ending at the vertices of color  $x$  with edges  $e$  and  $f$  as its ends. By construction, this new arc does not overlap with any of the arcs in  $I$ , which contradicts the maximality of  $I$ .  $\square$



Let us take a set  $D$  of edge-disjoint non-crossing cycles of  $J(O, col)$  and an independent set of arcs  $Y$  of  $O$ . We call  $(D, Y)$  a valid pair if the set of edges included in  $Y$  and  $O(D)$  partition the set of edges of  $O$ .

**Lemma.** *The size of a maximum non-crossing cycle decomposition of  $J(O, col)$  is equal to the size of a MISA of  $O$ .*

*Proof.* For a MISA  $I$  of  $O$  we have that  $(\emptyset, I)$  is a valid pair due to the previously proven lemma. We start by showing how from a given valid pair  $(D, Y)$  with a non-empty  $Y$  we can obtain a valid pair  $(D', Y')$  with  $|D'| = |D| + 1$  and  $|Y'| = |Y| - 1$ . Take a maximal arc  $U$  in  $Y$ . The set  $c$  of edges of  $U$  not belonging to any other arc in  $Y$  is non-empty as it includes its ends.  $J(V)$  is a cycle of  $J(O, col)$  for every arc  $V$  of  $O$ , thus we obtain that  $J(c)$  is also a cycle of  $J(O, col)$  and, by construction, it does not cross with the cycles in  $D$ . We include this cycle in  $D$  and eliminate  $U$  from  $Y$  to obtain  $(D', Y')$ . Iterating this step we obtain a non-crossing cycle decomposition of  $J(O, col)$  of size  $|I|$ .

Take a maximum non-crossing cycle decomposition  $C$  of  $J(O, col)$ .  $(C, \emptyset)$  is a valid pair. We proceed by showing how from a given valid pair  $(D, Y)$  with non-empty  $D$  we can obtain a valid pair  $(D', Y')$  with  $|D'| = |D| - 1$  and  $|Y'| = |Y| + 1$ . For every cycle  $c$  in  $D$  we define the minimum length path  $P(c)$  in  $O$  that contains every edge of  $O(c)$ . We choose such  $c$  for which  $P(c)$  is of minimum length. Since  $P(c)$  is the minimum length path containing  $O(c)$ , we have that the ends of  $P(c)$  belong to  $O(c)$  and thus do not belong to any arc in  $Y$ . We remove from  $P(c)$  any arcs of  $Y$  that it might contain to obtain  $P'$ , a set of edges of  $O$ , and show that  $P'$  consists entirely of edges of  $O(c)$ . Suppose by contradiction that there is an edge  $e \in P'$  such that  $e \notin O(c)$ . As it does not belong to any arc in  $Y$  and  $(D, Y)$  is valid, there must be a cycle  $c' \in C$  that contains  $e$ . However  $O(c)$  and  $O(c')$  can not cross and the endpoints of  $P(c)$  belong to  $O(c)$  and stay in  $P'$ , thus  $O(c')$  must be properly included in  $P'$  and thus in  $P(c)$ . However this means that  $P(c')$  is also properly included in  $P(c)$  and thus shorter in length which contradicts the minimality of  $P(c)$ . Thus we obtain that  $P'$  consists entirely of edges of  $O(c)$ . Now we show that  $P(c)$  is an arc non-overlapping with any arc in  $Y$ . We have already shown that  $P(c)$  is a union of  $O(c)$  where  $c$  is a cycle in  $J(O, col)$  and maybe some arcs from  $Y$ . This establishes that there is an equal number of gray and black edges in  $P(c)$  and that the colors of its endpoints are the same, which means that  $P(c)$  is an arc. By construction, the ends of this arc do not belong to any arc in  $Y$  and thus  $\{P(c)\} \cup Y$  is an independent set of arcs. We remove  $c$  from  $C$  and add  $P(c)$  to  $Y$  to obtain a valid pair  $(D', Y')$ . Iterating this step we obtain an independent set of arcs of  $O$  of cardinality equal to the size of a maximum non-crossing cycle decomposition.  $\square$

## .6 Lemma 3

**Definition 7** (Chain). *A chain  $H$  in the set of arcs  $I$  is a set of edge-disjoint arcs  $U_1, \dots, U_m$  in  $I$  with endpoints of  $U_i$  being  $u_i$  and  $v_i$  such that all these endpoints are of the same color and  $v_i = u_{i+1}$  for  $i \in \{1, \dots, m-1\}$ . We say that  $u_0$  and  $v_m$  are the endpoints of  $H$ .*

**Lemma.** *There exists a MISA  $I'$  of  $O$  such that there is no arc crossing  $v$  in  $I'$ .*

*Proof.* If there is no MISA having an arc crossing  $v$  then we are done. Otherwise take a MISA  $I$  of  $O$  having at least one arc crossing  $v$ . All such arcs include the edges adjacent to  $v$ , thus there exists the single maximal arc  $U$  in  $I$  crossing  $v$  that we remove from  $I$ . If after this  $v$  does not belong to any arc in  $I$  then we can include in  $I$  an arc whose both endpoints are  $v$  to obtain an independent set of arcs  $I'$  with no arc crossing  $v$ . Otherwise we take a maximum length chain  $H$  in  $I$  including  $v$ . Due to the maximality of  $U$  we obtain that  $H$  is included in  $U$  and does not include its ends, which means that the endpoints of  $H$  do not coincide. We take an edge  $e$  of  $O$  adjacent to an endpoint

of  $H$  but not belonging to  $H$  and show that  $e$  does not belong to any arc in  $I$ .  $e$  can not be an end of an arc as this arc would intersect with an arc in  $H$  or could extend it. If  $e$  belongs to an arc  $V$  without being its end, then all of  $H$  must belong to  $V$ , which contradicts the maximality of  $H$ . We obtain that an arc joining the endpoints of  $H$  and not crossing  $v$  can be included in  $I$  giving an independent set of arcs  $I'$  equal in size to  $I$  and having the number of arcs crossing  $v$  smaller than  $I$ . Iterating this process we obtain a MISA with no arcs crossing  $v$ .  $\square$

## .7 Lemma 4

**Lemma.** *For the DCJ scenarios transforming genome  $A$  into  $B$  the minimum length is  $l(G(A, B)) = e(G(A, B)) - c(G(A, B))$ , the minimum cost is  $l(J) = e(J) - c(J)$  with  $J = J(G(A, B), col)$  and the minimum cost of a DCJ scenario of minimum length is  $MCPS(G(A, B))$ .*

*Proof.*  $G(A, B)$  is constructed in such a way that for every DCJ  $A \rightarrow A'$  the transformation  $G(A, B) \rightarrow G(A', B)$  is a 2-break. Notably, a DCJ  $\{a, b\} \rightarrow \{a\}, \{b\}$  results in a transformation  $(a, b), (\circ, \circ) \rightarrow (a, \circ), (b, \circ)$ , as the construction of a breakpoint graph guarantees that there are enough black loops  $(\circ, \circ)$  to realize such a 2-break. For any 2-break  $G(A, B) \rightarrow G'$  with  $G' \neq G(A, B)$  there exists a DCJ  $A \rightarrow A'$  such that  $G(A', B) = G'$ . Since  $G(B, B)$  is terminal, it follows that the minimum length of a scenario transforming  $A$  into  $B$  is  $l(G(A, B))$ .

Vertices of  $G(A, B)$  are the gene extremities of  $A$  and  $B$  plus an additional vertex  $\circ$ . We color  $\circ$  with a unique color. The rest of the vertices of  $G(A, B)$  have the colors of their gene extremities. By construction, the cost of a DCJ  $A \rightarrow A'$  is equal to the cost of a 2-break  $G(A, B) \rightarrow G(A', B)$ . On the other hand for a 2-break  $G(A, B) \rightarrow G'$  there exists a DCJ  $A \rightarrow A'$  of the same cost such that  $G(A', B) = G'$ . This means that the minimum cost of a scenario transforming  $A$  into  $B$  is equal to the minimum cost of a scenario for  $G(A, B)$  and we conclude using Theorem 4 that this cost is  $l(J) = e(J) - c(J)$  with  $J = J(G(A, B), col)$ .

Let us denote the minimum cost of a parsimonious DCJ scenario transforming  $A$  into  $B$  by  $w_{min}$ . In the two previous paragraphs we have seen that for a DCJ scenario of length  $l$  and cost  $w$  transforming  $A$  into  $B$  there exists a 2-break scenario on  $G(A, B)$  of length  $l$  and cost  $w$ . This means that  $MCPS(G(A, B)) \leq w_{min}$ . On the other hand we have also seen that for a 2-break scenario on  $G(A, B)$  of length  $l$  and cost  $w$  there exists a DCJ scenario of length  $l$  and cost  $w$  transforming  $A$  into  $B$  and this establishes the equality  $MCPS(G(A, B)) = w_{min}$ .  $\square$

## .8 Lemma 5

**Lemma.** *The weights of the edges of  $H$  can be assigned in  $O(n^4)$  time.*

*Proof.*  $a_1, \dots, a_x$  denotes the sizes of  $AA$  paths and  $b_1, \dots, b_y$  denotes the sizes of  $BB$  paths with  $\sum_{i=0}^x a_i = |AA|$  and  $\sum_{j=0}^y b_j = |BB|$ . By construction,  $d^b(G(A, B), \circ) = d^g(G(A, B), \circ) = 2n$ , meaning that  $x, y \leq n$ . MCPS cost of a union of two paths having  $a_i$  and  $b_j$  vertices can be computed in  $2c(a_i + b_j)^3$  steps as MCPS cost of a circle of size  $n$  can be computed in  $cn^3$  steps for some constant  $c$  and we need to compute this cost for two Eulerian circles. We can compute MCPS cost for every pair of  $AA$  and  $BB$  path in the number of steps equal to

$$\begin{aligned} \sum_{i=0}^x \sum_{j=0}^y 2c(a_i + b_j)^3 &= 2c \sum_{i=0}^x \sum_{j=0}^y (a_i^3 + 3a_i^2 b_j + 3b_j^2 a_i + b_j^3) \\ &= 2c \left( y \sum_{i=0}^x a_i^3 + x \sum_{j=0}^y b_j^3 + 3|BB| \sum_{i=0}^x a_i^2 + 3|AA| \sum_{j=0}^y b_j^2 \right) \end{aligned}$$

The terms  $y, x, |AA|$  and  $|BB|$  are clearly  $O(n)$  and we obtain the worst-case time-complexity  $O(n^4)$  for weighting the bipartite graph.  $\square$