SCJ: a variant of breakpoint distance for which sorting, genome median and genome halving problems are easy

Pedro Feijão[1] and João Meidanis[1,2]

[1] Institute of Computing, University of Campinas, Brazil
[2] Scylla Bioinformatics, Brazil

Abstract. The breakpoint distance is one of the most straightforward genome comparison measures. Surprisingly, when it comes to define it precisely for multichromosomal genomes with both linear and circular chromosomes, there is more than one way to go about it. In this paper we study Single-Cut-or-Join (SCJ), a breakpoint-like rearrangement event for which we present linear and polynomial time algorithms that solve several genome rearrangement problems, such as median and halving. For the multichromosomal linear genome median problem, this is the first polynomial time algorithm described, since for other breakpoint distances this problem is NP-hard. These new results may be of value as a speedily computable, first approximation to distances or phylogenies based on more realistic rearrangement models.

1 Introduction

Genome rearrangement, an evolutionary event where large, continuous pieces of the genome shuffle around, has been studied since shortly after the very advent of genetics [1, 2, 3]. With the increased availability of whole genome sequences, gene order data have been used to estimate the evolutionary distance between present-day genomes and to reconstruct the gene order of ancestral genomes. The inference of evolutionary scenarios based on gene order is a hard problem, with its simplest version being the pairwise genome rearrangement problem: given two genomes, represented as sequences of conserved segments called *syntenic blocks*, find the most parsimonious sequence of rearrangement events that transforms one genome into the other. In some applications, one is interested only in the number of events of such a sequence — the *distance* between the two genomes.

Several rearrangement events, or operations, have been proposed. Early approaches considered the case where just one operation is allowed. For some operations a polynomial solution was found (e.g., for reversals [4], translocations [5], and block-interchanges [6]), while for others the complexity is still open (e.g., for transpositions [7, 8]). Later on, polynomial algorithms for combinations of operations were discovered (e.g., for block-interchanges and reversals [9]; for fissions, fusions, and transpositions [10]; for fissions, fusions, and block-interchanges [11]). Yancopoulos et al. [12] introduced a very comprehensive model, with reversals,

transpositions, translocations, fusions, fissions, and block-interchanges modeled as compositions of the same basic operation, the double-cut-and-join (DCJ).

Different relative weights for the operations have been considered. Proposals have also differed in the number and type of allowed chromosomes (unichromosomal vs. multichromosomal genomes; linear or circular chromosomes).

When more than two genomes are considered, we have the more challenging problem of rearrangement-based phylogeny reconstruction, where we want to find a tree that minimizes the total number of rearrangement events. Early approaches were based on a breakpoint distance (e.g., BPAnalysis [13], and GRAPPA [14]). With the advances on pairwise distance algorithms, more sophisticated distances were used, with better results (e.g., reversal distance, used by MGR [15] and in an improved version of GRAPPA, and DCJ distance [16]).

Two problems are commonly used to find the gene order of ancient genomes in rearrangement-based phylogeny reconstruction: the median problem and the halving problem. These problems are NP-hard in most cases even under the simplest distances.

In this paper we propose a new way of computing breakpoint distances, based on the the *Single-Cut-or-Join* (SCJ) operation, and show that several rearrangement problems involving it are polynomial. For some problems, these will be the only polynomial results known to date. The SCJ distance is exactly twice the the breakpoint (BP) distance of Tannier et al. [17] for circular genomes, but departs from it when linear chromosomes are present, because of an alternative way of treating telomeres. In a way, SCJ is the simplest mutational event imaginable, and it may be of value as a speedily computable, first approximation to distances or phylogenies based on more realistic rearrangement models.

The rest of this paper is structured as follows. In Section 2 we present the basic definitions, including SCJ. Section 3 deals with the distance problem and compares SCJ to other distances. Sections 4 and 5 deal with genome medians and genome halving, respectively, and their generalizations. Finally, in Section 6 we present a brief discussion and future directions.

2 Representing Genomes

We will use a standard genome formulation [18, 17]. A *gene* is an oriented sequence of DNA that starts with a tail and ends with a head, called the *extremities* of the gene. The tail of a gene a is denoted by a_t, and its head by a_h. Given a set of genes $\mathcal{G}$, the extremity set is $\mathcal{E} = \{a_t : a \in \mathcal{G}\} \cup \{a_h : a \in \mathcal{G}\}$. An *adjacency* is an unordered pair of two extremities that represents the linkage between two consecutive genes in a certain orientation on a chromosome, for instance $a_h b_t$. An extremity that is not adjacent to any other extremity is called a *telomere*. A genome is represented by a set of adjacencies where the tail and head of each gene appear at most once. Telomeres will be omitted in our representation, since they are uniquely determined by the set of adjacencies and the extremity set $\mathcal{E}$.

The *graph representation* of a genome Π is a graph G_Π whose vertices are the extremities of Π and there is a grey edge connecting the extremities x and y

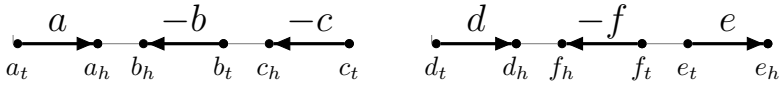

Fig. 1. Graph G_Π representing a genome with two linear chromosomes. Black directed edges represent genes, while grey edges link consecutive extremities.

when xy is an adjacency of Π or a directed black edge if x and y are head and tail of the same gene. A connected component in G_Π is a *chromosome* of Π, and it is *linear* if it is a path, and *circular* if it is a cycle. A *circular genome* is a genome whose chromosomes are all circular, and a *linear genome* is a genome whose chromosomes are all linear. A *string representation* of a genome Π, denoted by Π_S, is a set of strings corresponding to the genes of Π in the order they appear on each chromosome, with a bar over the gene if it is read from head to tail and no bar otherwise. Notice that the string representation is not unique: each chromosome can be replaced by its reverse complement.

For instance, given the set $\mathcal{G} = \{a, b, c, d, e, f\}$, and the genome $\Pi = \{a_h b_h, b_t c_h, d_h f_h, f_t e_t\}$, the graph G_Π is given in Figure 1. Notice that telomeres a_t, c_t, d_t, and e_h are omitted from the set representation without any ambiguity. A string representation of this genome is $\Pi_S = \big(a\ \overline{b}\ \overline{c}\ ,\ d\ \overline{f}\ e\big)$.

In problems where gene duplicates are allowed, a gene can have any number of homologous copies within a genome. Each copy of a gene is called a *duplicated gene* and is identified by its tail and head with the addition of an integer label, from 1 to n, identifying the copy. For instance, a gene g with three copies has extremities $g_h^1, g_t^1, g_h^2, g_t^2, g_h^3$ and g_t^3. An *n-duplicate genome* is a genome where each gene has exactly n copies. An *ordinary genome* is a genome with a single copy of each gene. We can obtain n-duplicate genomes from an ordinary genome with the following operation: for an ordinary genome Π on a set $\mathcal{G}$, Π^n represents a set of n-duplicate genomes on $\mathcal{G}^n = \{a^1, a^2, \ldots, a^n : a \in \mathcal{G}\}$ such that if the adjacency xy belongs to Π, n adjacencies of the form $x^i y^j$ belong to any genome in Π^n. The assignment of labels i and j to the duplicated adjacencies is arbitrary, with the restriction that each extremity copy has to appear exactly once. For instance, for $n = 3$ a valid choice could be $\{x^1y^2, x^2y^1, x^3y^3\}$. Since for each adjacency in Π we have $n!$ possible choices for adjacencies in Π^n, the number of genomes in the set Π^n is $|\Pi|^{n!}$, where $|\Pi|$ is the number of adjacencies of Π.

2.1 A New Rearrangement Operation

We will define two simple operations applied directly on the adjacencies and telomeres of a genome. A *cut* is an operation that breaks an adjacency in two telomeres (namely, its extremities), and a *join* is the reverse operation, pairing two telomeres into an adjacency. Any *cut* or *join* applied to a genome will be called a **Single-Cut-or-Join** (SCJ) operation. In this paper, we are interested in solving several rearrangement problems under the SCJ distance.

Since each genome is a set of adjacencies, standard set operations such as union, intersection and set difference can be applied to two (or more) genomes.

In the case of intersection and set difference, the result is a set of adjacencies contained in at least one of the genomes, and therefore it is also a genome. On the other hand, the set resulting from a union operation might not represent a genome since the same extremity could be present in more than one adjacency. We will use these operations throughout this paper in our algorithms, and whenever union is used, we will prove that the resulting set represents a valid genome. Set difference between sets A and B will be denoted by $A - B$.

3 Rearrangement by SCJ

The *rearrangement by SCJ problem* is stated as follows: given two genomes Π and Σ with the same set of genes $\mathcal{G}$, find a shortest sequence of SCJ operations that transforms Π into Σ. This problem is also called *genome sorting*. The length of such a sequence is called the *distance* between Π and Σ and is denoted by $d_{SCJ}(\Pi, \Sigma)$.

Since the only possible operations are to remove (cut) or include (join) an adjacency in a genome, the obvious way of transforming Π into Σ is to remove all adjacencies that belong to Π and not to Σ, and then include all adjacencies that belong to Σ and not to Π.

Lemma 1. *Consider the genomes Π and Σ, and let $\Gamma = \Pi - \Sigma$ and $\Lambda = \Sigma - \Pi$. Then, Γ and Λ can be found in linear time and they define a minimum set of SCJ operations that transform Π into Σ, where adjacencies in Γ define cuts and adjacencies in Λ define joins. Consequently, $d_{SCJ}(\Pi, \Sigma) = |\Pi - \Sigma| + |\Sigma - \Pi|$.*

Proof. Considering the effect an arbitrary cut or join on Π can have on the quantity $f_\Sigma(\Pi) = |\Pi - \Sigma| + |\Sigma - \Pi|$, it is straightforward to verify that $f_\Sigma(\Pi)$ can increase or decrease by at most 1. Hence, the original value is a lower bound on the distance. Given that the sequence of operations proposed in the statement does lead from Π to Σ along valid genomes in that number of steps, we have our lemma. □

3.1 SCJ Distance with the Adjacency Graph

The Adjacency Graph, introduced by Bergeron et al. [18], was used to find an easy equation for the DCJ distance. The adjacency graph $AG(\Pi, \Sigma)$ is a bipartite graph whose vertices are the adjacencies and telomeres of the genomes Π and Σ and whose edges connect two vertices that have a common extremity. Therefore, vertices representing adjacencies will have degree two and telomeres will have degree one, and this graph will be a union of path and cycles.

A formula for the SCJ distance based on the cycles and paths of $AG(\Pi, \Sigma)$ can be easily found, as we will see in the next lemma. We will use the following notation: C and P represent the number of cycles and paths of $AG(\Pi, \Sigma)$, respectively, optionally followed by a subscript to indicate the number of edges (the *length*) of the cycle or path or if the length is odd or even. For instance, P_2 is the number of paths of length two, $C_{\geq 4}$ is the number of cycles with length four or more and P_{odd} is the number of paths with an odd length.

Lemma 2. *Consider two genomes Π and Σ with the same set of genes $\mathcal{G}$. Then, we have*

$$d_{SCJ}(\Pi, \Sigma) = 2\big[N - (C_2 + P/2)\big], \tag{1}$$

where N is the number of genes, C_2 is the number of cycles of length two and P the number of paths in $AG(\Pi, \Sigma)$.

Proof. We know from the definition of SCJ distance and basic set theory that

$$d_{SCJ}(\Pi, \Sigma) = |\Pi - \Sigma| + |\Sigma - \Pi| = |\Sigma| + |\Pi| - 2|\Sigma \cap \Pi|.$$

Since the number of cycles of length two in $AG(\Pi, \Sigma)$ is the number of common adjacencies of Π and Σ, we have $|\Sigma \cap \Pi| = C_2$. For any genome Π, we know that $|\Pi| = N - t_\Pi/2$, where t_Π is the number of telomeres of Π. Since each path in $AG(\Pi, \Sigma)$ has exactly two vertices corresponding to telomeres of Π and Σ, the total number of paths in $AG(\Pi, \Sigma)$, denoted by P, is given by $P = (t_\Pi + t_\Sigma)/2$. Therefore,

$$\begin{aligned} d_{SCJ}(\Pi, \Sigma) &= |\Sigma| + |\Pi| - 2|\Sigma \cap \Pi| = \\ &= 2N - (t_\Pi + t_\Sigma)/2 - 2C_2 = 2\big[N - (C_2 + P/2)\big]. \end{aligned}$$ □

3.2 Comparing SCJ Distance to Other Distances

Based on the adjacency graph, we have the following equation for the DCJ distance [18]:

$$d_{DCJ}(\Pi, \Sigma) = N - (C + P_{odd}/2), \tag{2}$$

where N is the number of genes, C is the number of cycles, and P_{odd} is the number of odd paths in $AG(\Pi, \Sigma)$. For the Breakpoint (BP) distance, as defined by Tannier et al. [17], we have

$$d_{BP}(\Pi, \Sigma) = N - (C_2 + P_1/2), \tag{3}$$

where C_2 is the number of cycles with length two and P_1 is the number of paths with length one in $AG(\Pi, \Sigma)$.

With these equations we can find relationships between SCJ, BP, and DCJ distances. First, for the BP distance, we have

$$d_{SCJ}(\Pi, \Sigma) = 2d_{BP}(\Pi, \Sigma) - P_{\geq 2}.$$

As expected, the SCJ distance is related to the BP distance, differing only by a factor of 2 and the term $P_{\geq 2}$, the number of paths with two or more edges. For circular genomes, $P = 0$ and the SCJ distance is exactly twice the BP distance. For the general case, the following sandwich formula holds:

$$d_{BP}(\Pi, \Sigma) \leq d_{SCJ}(\Pi, \Sigma) \leq 2d_{BP}(\Pi, \Sigma).$$

For the DCJ distance, a reasonable guess is that it would be one fourth of the SCJ distance, since a DCJ operation, being formed by two cuts and two joins, should correspond to four SCJ operations. This is not true, however, for two reasons.

First, a DCJ operation may correspond to four, two, or even one SCJ operation. Examples of these three cases are shown in Figure 2, with *caps* represented by the symbol $\circ$. In each case the target genome is $(a\ b\ c\ d)$. The figure shows a reversal, a suffix reversal, and a linear fusion, all of which have the same weight under the DCJ model, but different SCJ distances, because caps do not exist in the SCJ model. Incidentally, they have different BP distances as well.

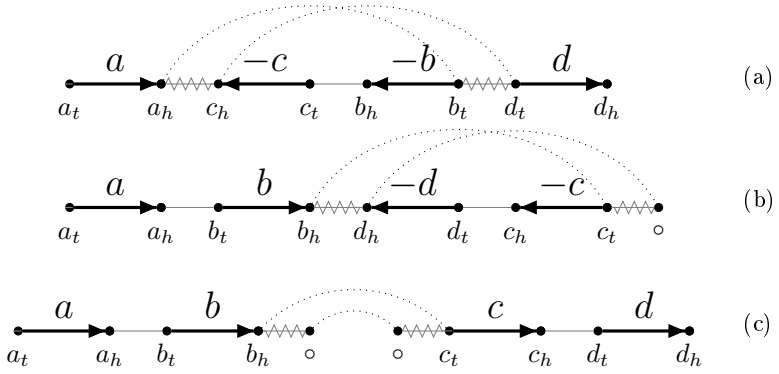

Fig. 2. Three types of single DCJ operations transforming each genome into $\Pi_S = \{a, b, c, d\}$. (a) Reversal. (b) *Suffix* Reversal. (c) Linear Fusion.

The second reason is that, when consecutive DCJ operations use common spots, the SCJ model is able to "cancel" operations, resulting in a shorter sequence. Both arguments show SCJ saving steps, which still leaves four times DCJ distance as an upper bound on SCJ distance. The complete sandwich result is

$$d_{DCJ}(\Pi, \Sigma) \leq d_{SCJ}(\Pi, \Sigma) \leq 4d_{DCJ}(\Pi, \Sigma).$$

4 The Genome Median Problem

The *Genome Median Problem* (GMP) is an important tool for phylogenetic reconstruction of trees with ancestral genomes based on rearrangement events. When genomes are unichromosomal this problem is NP-hard under the breakpoint, reversal and DCJ distances [19, 20]. In the multichromosomal general case, when there are no restrictions as to whether the genomes are linear or circular, Tannier et al. [17] recently showed that under the DCJ distance the problem is still NP-hard, but it becomes polynomial under the breakpoint distance (BP),

the first polynomial result for the median problem. The problem can be solved in linear time for SCJ, our version of the breakpoint distance.

We show this by proposing a more general problem, *Weighted Multichromosomal Genome Median Problem* (WMGMP), where we find the genome median among any number of genomes with weights for the genomes. We will give a straightforward algorithm for this problem under the SCJ distance in the general case, from which the special case of GMP follows with unique solution, and then proceed to solve it with the additional restrictions of allowing only linear or only circular chromosomes.

4.1 Weighted Multichromosomal Genome Median Problem

This problem is stated as follows: Given n genomes $\Pi_1, \ldots, \Pi_n$ with the same set of genes $\mathcal{G}$, and nonnegative weights $w_1, \ldots, w_n$, we want to find a genome Γ that minimizes $\sum_{i=1}^{n} w_i \cdot d(\Pi_i, \Gamma)$.

We know that

$$\sum_{i=1}^{n} w_i \cdot d(\Pi_i, \Gamma) = \sum_{i=1}^{n} w_i |\Pi_i| + \sum_{i=1}^{n} w_i |\Gamma| - 2 \sum_{i=1}^{n} w_i |\Gamma \cap \Pi_i|$$

and since $\sum_{i=1}^{n} w_i |\Pi_i|$ does not depend on Γ we want to minimize

$$f(\Gamma) \equiv \sum_{i=1}^{n} w_i |\Gamma| - 2 \sum_{i=1}^{n} w_i |\Gamma \cap \Pi_i| \tag{4}$$

Now, for any adjacency d, let $\mathcal{S}_d$ be the set of indices i for which Π_i has this adjacency, that is, $\mathcal{S}_d = \{i : d \in \Pi_i\}$. To simplify the notation, we will write $f(\{d\})$ as $f(d)$. Then we have

$$f(d) = \sum_{i=1}^{n} w_i |\{d\}| - 2 \sum_{i=1}^{n} w_i |\{d\} \cap \Pi_i| = \sum_{i=1}^{n} w_i - 2 \sum_{i \in \mathcal{S}_d} w_i = \sum_{i \notin \mathcal{S}_d} w_i - \sum_{i \in \mathcal{S}_d} w_i$$

and it is easy to see that for any genome Γ we have

$$f(\Gamma) = \sum_{d \in \Gamma} f(d) \tag{5}$$

Since we want to minimize $f(\Gamma)$, a valid approach would be to choose Γ as the genome with all adjacencies d such as $f(d) < 0$. As we will see from the next lemma, this strategy is optimal.

Lemma 3. *Given n genomes $\Pi_1, \ldots, \Pi_n$ and nonnegative weights $w_1, \ldots, w_n$, the genome $\Gamma = \{d : f(d) < 0\}$, where*

$$f(d) = \sum_{i \notin \mathcal{S}_d} w_i - \sum_{i \in \mathcal{S}_d} w_i$$

and $\mathcal{S}_d = \{i : d \in \Pi_i\}$, minimizes $\sum_{i=1}^{n} w_i \cdot d(\Pi_i, \Gamma)$. Furthermore, if there is no adjacency $d \in \Pi_i$ for which $f(d) = 0$, then Γ is a unique solution.

Proof. Let xy be an adjacency such that $f(xy) < 0$. For any extremity $z \neq y$, we have $xy \in \Pi_i \Rightarrow xz \notin \Pi_i$ and $xz \in \Pi_i \Rightarrow xy \notin \Pi_i$. Therefore

$$f(xz) = \sum_{i \notin \mathcal{S}_{xz}} w_i - \sum_{i \in \mathcal{S}_{xz}} w_i \geq \sum_{i \in \mathcal{S}_{xy}} w_i - \sum_{i \notin \mathcal{S}_{xy}} w_i = -f(xy) > 0$$

This means that adjacencies d with $f(d) < 0$ do not have extremities in common and it is then possible to add all those adjacencies to form a valid genome Γ, minimizing $f(\Gamma)$ and consequently $\sum_{i=1}^{n} w_i \cdot d(\Pi_i, \Gamma)$.

To prove the uniqueness of the solution, suppose there is no adjacency d such that $f(d) = 0$. Since any adjacency d belonging to Γ satisfies $f(d) < 0$ and any other adjacency d' satisfies $f(d') > 0$, for any genome $\Gamma' \neq \Gamma$ we have $f(\Gamma') > f(\Gamma)$, confirming that Γ is a unique solution. If there is d with $f(d) = 0$, then $\Gamma' = (\Gamma \cup d)$ is a valid genome (that is, the extremities of d are telomeres in Γ), which is also a solution, and uniqueness cannot be guaranteed. □

After solving the general case, we will restrict the problem to circular or linear genomes in the next two sections.

4.2 The Weighted Multichromosomal Circular Median Problem

In this section we will solve the WMGMP restricted to circular genomes: given n *circular* genomes $\Pi_1, \ldots, \Pi_n$ with the same set of genes $\mathcal{G}$, and nonnegative weights $w_1, \ldots, w_n$, we want to find a *circular* genome Γ which minimizes $\sum_{i=1}^{n} w_i \cdot d(\Pi_i, \Gamma)$.

It is easy to see that a genome is circular if and only if it has N adjacencies, where N is the number of genes. Basically we want to minimize the same function f defined in equation (4) with the additional constraint $|\Gamma| = N$. To solve this problem, let G be a complete graph where every extremity in the set $\mathcal{E}$ is a vertex and the weight of an edge connecting vertices x and y is $f(xy)$. Then a perfect matching on this graph corresponds to a circular genome Γ and the total weight of this matching is $f(\Gamma)$. Then, a minimum weight perfect matching can be found in polynomial time [21] and it is an optimum solution to the weighted circular median problem.

4.3 The Weighted Multichromosomal Linear Median Problem

The solution of this problem is found easily using the same strategy as in the WMGMP. Since we have no restrictions on the number of adjacencies, we find Γ as defined in Lemma 3, including only adjacencies for which $f > 0$. If Γ is linear, this is the optimum solution. If Γ has circular chromosomes, a linear median Γ' can be obtained by removing, in each circular chromosome of Γ, an adjacency xy with maximum $f(xy)$. Removing xy would allow the inclusion of new adjacencies of the forms xw and yz, but we know that $f(xy) < 0$ implies $f(xw) > 0$ and $f(yz) > 0$. Therefore, any genome Σ different from Γ' either has a circular chromosome or has $f(\Sigma) \geq f(\Gamma')$. Therefore, Γ' is an optimal solution.

This is the first polynomial result for this problem.

5 Genome Halving and Genome Aliquoting

The *Genome Halving Problem* (GHP) is motivated by whole genome duplication events in molecular evolution, postulated by Susumu Ohno in 1970 [22]. Whole genome duplication has been very controversial over the years, but recently, very strong evidence in its favor was discovered in yeast species [23]. The goal of a halving analysis is to reconstruct the ancestor of a 2-duplicate genome at the time of the doubling event.

The GHP is stated as follows: given a 2-duplicate genome Δ, find an ordinary genome Γ that minimizes $d(\Delta, \Gamma^2)$, where

$$d(\Delta, \Gamma^2) = \min_{\Sigma \in \Gamma^2} d(\Delta, \Sigma) \tag{6}$$

If both Δ and Γ are given, computing the right hand side of Equation (6) is known as the *Double Distance* problem, which has a polynomial solution under the breakpoint distance but is NP-hard under the DCJ distance [17]. In contrast, the GHP has a polynomial solution under the DCJ distance for unichromosomal genomes [24], and for multichromosomal genomes when both linear and circular chromosomes are allowed [25].

Warren and Sankoff recently proposed a generalization of the halving problem, the *Genome Aliquoting Problem* (GAP) [26]: Given an n-duplicate genome Δ, find an ordinary genome Γ that minimizes $d(\Delta, \Gamma^n)$. In their paper, they use the DCJ distance and develop heuristics for this problem, but a polynomial time exact solution remains open. To the best of our knowledge, this problem has never been studied under any other distance. We will show that under the SCJ distance this problem can be formulated as a special case of the WMGMP.

Lemma 4. *Given an n-duplicate genome Δ, define n ordinary genomes Π_1, $\ldots$, Π_n as follows. For each ordinary adjacency xy, add it to the k first genomes $\Pi_1, \ldots, \Pi_k$, where k is the number of adjacencies of the form $x^i y^j$ in Δ. Then, for every genome Γ, we have $d(\Delta, \Gamma^n) = \sum_{i=1}^{n} d(\Gamma, \Pi_i)$.*

Proof. We have that

$$\begin{aligned} d(\Delta, \Gamma^n) &= \min_{\Sigma \in \Gamma^n} d(\Delta, \Sigma) = \min_{\Sigma \in \Gamma^n} (|\Delta| + |\Sigma| - 2|\Delta \cap \Sigma|) \\ &= |\Delta| + n|\Gamma| - 2 \max_{\Sigma \in \Gamma^n} |\Delta \cap \Sigma| \end{aligned}$$

To maximize $|\Delta \cap \Sigma|$, let $k(xy)$ be the number of adjacencies of the form $x^i y^j$ in Δ. For each adjacency xy in Γ, we add to Σ the $k(xy)$ adjacencies of Δ plus $n - k(xy)$ arbitrarily labeled adjacencies, provided they do not collide. It is clear that this Σ maximizes $|\Delta \cap \Sigma|$ and furthermore

$$\max_{\Sigma \in \Gamma^n} |\Delta \cap \Sigma| = \sum_{xy \in \Gamma} k(xy)$$

On the other hand,

$$\sum_{i=1}^{n} d(\Gamma, \Pi_i) = n|\Gamma| + \sum_{i=1}^{n} |\Pi_i| - 2 \sum_{i=1}^{n} |\Pi_i \cap \Gamma|$$

Now $\sum_{i=1}^{n} |\Pi_i \cap \Gamma|$ is exactly $\sum_{xy \in \Gamma} k(xy)$, since any adjacency xy in Δ appears exactly $k(xy)$ times in genomes $\Pi_1, \ldots, \Pi_n$. Taking into account that $|\Delta| = \sum_{i=1}^{n} |\Pi_i|$, we have our result. □

Lemma 4 implies that the GAP is actually a special case of the WMGMP, and can be solved using the same algorithm. Another corollary is that the constrained versions of the GAP for linear or circular multichromosomal genomes are also polynomial.

5.1 Guided Genome Halving

The *Guided Genome Halving* (GGH) problem was proposed very recently, and is stated as follows: given a 2-duplicate genome Δ and an ordinary genome Γ, find an ordinary genome Π that minimizes $d(\Delta, \Pi^2) + d(\Gamma, \Pi)$. This problem is related to Genome Halving, only here an ordinary genome Γ, presumed to share a common ancestor with Π, is used to *guide* the reconstruction of the ancestral genome Π.

Under the BP distance, the GGH has a polynomial solution for general multichromosomal genomes [17] but is NP-hard when only linear chromosomes are allowed [27]. For the DCJ distance, it is NP-hard in the general case [17].

As in the Halving Problem, here we will solve a generalization of GGH, the *Guided Genome Aliquoting* problem: given an n-duplicate genome Δ and an ordinary genome Γ, find an ordinary genome Π that minimizes $d(\Delta, \Pi^n) + d(\Gamma, \Pi)$. It turns out that the version with an n-duplicate genome Δ as input is very similar to the "unguided" version with an $(n+1)$-duplicate genome.

Lemma 5. *Given an n-duplicate genome Δ and an ordinary genome Γ, let Δ' be an $(n+1)$-duplicate genome such that $\Delta' = \Delta \cup \{x^{n+1}y^{n+1} : xy \in \Gamma\}$. Then for any genome Π we have $d(\Delta', \Pi^{n+1}) = d(\Delta, \Pi^n) + d(\Gamma, \Pi)$.*

Proof. We have that

$$\begin{aligned} \min_{\Sigma \in \Pi^n} d(\Delta, \Sigma) &= \min_{\Sigma \in \Pi^n} (|\Delta| + |\Sigma| - 2|\Delta \cap \Sigma|) \\ &= |\Delta| + n|\Pi| - 2 \max_{\Sigma \in \Pi^n} |\Delta \cap \Sigma| \end{aligned}$$

and

$$\min_{\Sigma' \in \Pi^{n+1}} d(\Delta', \Sigma') = |\Delta'| + (n+1)|\Pi| - 2 \max_{\Sigma' \in \Pi^{n+1}} |\Delta' \cap \Sigma'|.$$

Since $|\Delta'| = |\Delta| + |\Gamma|$ and $|\Gamma \cap \Pi| + \max_{\Sigma \in \Pi^n} |\Delta \cap \Sigma| = \max_{\Sigma' \in \Pi^{n+1}} |\Delta' \cap \Sigma'|$, we have our result. □

The last lemma implies that GGH is a special case of GAP, which in turn is a special case of the WMGMP. Again, the constrained linear or circular versions are also polynomial for GGH in the SCJ model.

6 Discussion and Future Directions

In this paper we show that a variant of breakpoint distance, based on the Single-Cut-or-Join (SCJ) operation, allows linear- and polynomial-time solutions to some rearrangement problems that are NP-hard under the BP distance, for instance, the multichromosomal linear versions of the genome halving, guided halving and genome median problems. In addition, the SCJ approach is able to produce a rearrangement scenario between genomes, not only a distance, which is useful for phylogeny reconstruction.

The complexity of unichromosomal median and halving remain open under the SCJ distance.

From a biological point of view, we can think of a rearrangement event as an accepted mutation, that is, a mutational event involving large, continuous genome segments that was accepted by natural selection, and therefore became fixed in a population. SCJ may model the mutation part well, but a model for the acceptance part is missing. For instance, while the mutational effort of doing a fission seems to be less than that of an inversion, the latter is more frequent as a rearrangement event, probably because it has a better chance of being accepted. This may have to do with the location and movement of origins of replication, since any free segment will need one to become fixed.

Other considerations, such as the length of segments, hotspots, presence of flanking repeats, etc. are likely to play a role in genome rearrangements, and need to be taken into account in a comprehensive model.

Although crude from the standpoint of evolutionary genomics, the new distance may serve as a fast, first-order approximation for other, better founded genomic rearrangement distances, and also for reconstructed phylogenies. We intend to pursue this line or work, applying it to real datasets and comparing the results to those obtained with other methods.

References

[1] Sturtevant, A.H., Dobzhansky, T.: Inversions in the third chromosome of wild races of *Drosophila pseudoobscura*, and their use in the study of the history of the species. PNAS **22**(7) (1936) 448–450

[2] McClintock, B.: The origin and behavior of mutable loci in maize. PNAS **36**(6) (1950) 344–355

[3] Nadeau, J.H., Taylor, B.A.: Lengths of chromosomal segments conserved since divergence of man and mouse. PNAS **81**(3) (1984) 814–818

[4] Hannenhalli, S., Pevzner, P.A.: Transforming cabbage into turnip: (polynomial algorithm for sorting signed permutations by reversals). In: Proc. 27th Ann. Symp. Theory of Computing STOC 95. (1995)

[5] Hannenhalli, S.: Polynomial-time algorithm for computing translocation distance between genomes. Discrete Appl. Math **71**(1–3) (1996) 137–151

[6] Christie, D.A.: Sorting permutations by block-interchanges. Information Processing Letters **60** (1996) 165–169

[7] Bafna, V., Pevzner, P.A.: Sorting by transpositions. SIAM J. Discrete Math. **11**(2) (1998) 224–240

[8] Elias, I., Hartman, T.: A 1.375-approximation algorithm for sorting by transpositions. Computational Biology and Bioinformatics, IEEE/ACM Transactions on **3**(4) (Oct.-Dec. 2006) 369–379
[9] Mira, C., Meidanis, J.: Sorting by block-interchanges and signed reversals. In: Proc. ITNG 2007. (2007) 670–676
[10] Dias, Z., Meidanis, J.: Genome rearrangements distance by fusion, fission, and transposition is easy. In: Proc. SPIRE 2001. (2001) 250–253
[11] Lu, C.L., Huang, Y.L., Wang, T.C., Chiu, H.T.: Analysis of circular genome rearrangement by fusions, fissions and block-interchanges. BMC Bioinformatics **7** (2006) 295
[12] Yancopoulos, S., Attie, O., Friedberg, R.: Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics **21**(16) (2005) 3340–3346
[13] Blanchette, M., Bourque, G., Sankoff, D.: Breakpoint phylogenies. Genome Inform Ser Workshop Genome Inform **8** (1997) 25–34
[14] Moret, B.M., Wang, L.S., Warnow, T., Wyman, S.K.: New approaches for reconstructing phylogenies from gene order data. Bioinformatics **17 Suppl 1** (2001) S165–S173
[15] Bourque, G., Pevzner, P.A.: Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Res **12**(1) (2002) 26–36
[16] Adam, Z., Sankoff, D.: The ABCs of MGR with DCJ. Evol Bioinform Online **4** (2008) 69–74
[17] Tannier, E., Zheng, C., Sankoff, D.: Multichromosomal genome median and halving problems. In: Proc. WABI 2008. Volume 5251 of LNCS. (2008) 1–13
[18] Bergeron, A., Mixtacki, J., Stoye, J.: A unifying view of genome rearrangements. In: Proc. WABI 2006. Volume 4175 of LNCS. (2006) 163–173
[19] Bryant, D.: The complexity of the breakpoint median problem. Technical Report CRM-2579, Centre de recherches mathematiques, Université de Montréal (1998)
[20] Caprara, A.: The reversal median problem. INFORMS J. Comput. **15** (2003) 93–113
[21] Lovász, L., Plummer, M.D.: Matching theory. In: Annals of Discrete Mathematics. Volume 29. North-Holland (1986)
[22] Ohno, S.: Evolution by gene duplication. Springer-Verlag (1970)
[23] Kellis, M., Birren, B.W., Lander, E.S.: Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. Nature **428**(6983) (2004) 617–624
[24] Alekseyev, M.A., Pevzner, P.A.: Colored de Bruijn graphs and the genome halving problem. IEEE/ACM Trans Comput Biol Bioinform **4**(1) (2007) 98–107
[25] Mixtacki, J.: Genome halving under DCJ revisited. In: Proc. COCOON 2008. Volume 5092 of LNCS. (2008) 276–286
[26] Warren, R., Sankoff, D.: Genome aliquoting with double cut and join. BMC Bioinformatics **10 Suppl 1** (2009) S2
[27] Zheng, C., Zhu, Q., Adam, Z., Sankoff, D.: Guided genome halving: hardness, heuristics and the history of the hemiascomycetes. Bioinformatics **24**(13) (2008) i96–104