The Rooted SCJ Median with Single Gene Duplications

Aniket C. Mane[1], Manuel Lafond[2], Pedro Feijão[3], and Cedric Chauve[1(✉)]

[1] Department of Mathematics, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada
cedric.chauve@sfu.ca
[2] Department of Computer Science, Université de Sherbrooke, 2500 Boul. de l'Université, Sherbrooke, Canada
[3] School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada

Abstract. The median problem is a classical problem in genome rearrangements. It aims to compute a gene order that minimizes the sum of the genomic distances to $k \geq 3$ given gene orders. This problem is intractable except in the related Single-Cut-or-Join and breakpoint rearrangement models. Here we consider the rooted median problem, where we assume one of the given genomes to be ancestral to the median, which is itself ancestral to the other genomes. We show that in the Single-Cut-or-Join model with single gene duplications, the rooted median problem is NP-hard. We also describe an Integer Linear Program for solving this problem, which we apply to simulated data, showing high accuracy of the reconstructed medians.

1 Introduction

Reconstructing the evolution of genomes at the level of large-scale genome rearrangements is an important problem in computational biology [17,19]. There are several computational problems related to rearrangements, ranging from the computation of pairwise distances in a given rearrangement model to the reconstruction of complete phylogenetic trees, often following a parsimony approach [12]. Among these problems, the reconstruction of ancestral gene orders given a species phylogeny has been considered in various frameworks, including the so-called Small Parsimony Problem (SPP), which aims at proposing gene orders at the internal nodes of the given species phylogeny while minimizing the sum of the genome rearrangement distances along its branches. The simplest instance of the SPP is the Median Problem, where the given phylogeny contains a single ancestral node whose gene order is to be reconstructed. In the present paper, we introduce novel results about the median problem, in a context where gene duplications are considered.

The median problem was introduced in 1996 [21], motivated by its application to iterative algorithms for solving the SPP [3]. Early results suggested that,

M. Blanchette and A. Ouangraoua (Eds.): RECOMB-CG 2018, LNBI 11183, pp. 28–48, 2018.
https://doi.org/10.1007/978-3-030-00834-5_2

even in the simple breakpoint distance model, computing a median gene order is intractable [20], and heuristics based on the Traveling Salesman Problem (TSP) were introduced to solve the breakpoint median problem [3,7]. However, in 2009, Tannier, Zheng and Sankoff proved that computing a median gene order that is allowed to contain an arbitrary mixture of linear and circular chromosomes was tractable in the breakpoint distance model, by using a reduction to the problem of computing a Maximum Weight Matching (MWM) [22]. This tractability result, the first of its kind in genome rearrangements, renewed the interest in gene order median problems, although most of the following work presented intractability results, even on variations of the breakpoint distance [5,9,14]. A notable exception was the Single-Cut-or-Join (SCJ) distance, introduced by Feijão and Meidanis [11], where it was shown that both the SCJ median problem and the SCJ SPP are tractable.

Gene duplication is another important evolutionary mechanism, ranging from single-gene duplication to whole-genome duplications (WGD) [13,15]. The first models of evolution by genome rearrangements considered the case of genomes with equal gene content, thus disregarding gene duplication and gene loss. When considered as a possible evolutionary event, gene duplication most often leads to intractability results, even for the simple pairwise gene order distance [1,4,6]. Notable exceptions include again variants of the SCJ distance. In [23] it was shown that in an evolutionary model including SCJ and whole-chromosome duplications, the pairwise distance problem is tractable. More recently, we introduced a variant of SCJ including single-gene duplications where the distance between an ancestral genome and a descendant genome can be computed, when orthology relations between the descendant and ancestral genes are provided [10]. We also showed that a directed median problem where the median is the ancestor of k given genomes is tractable, again by reduction to a MWM problem. These results raised the question of tractability boundaries towards the SPP in a rearrangement model, including gene duplication.

In the present work, we show that a different median problem, which involves an additional given ancestral genome, is intractable. More precisely, we introduce the rooted median problem, where we are provided with $k + 1 \geq 3$ genomes, $A, D_1, \ldots, D_k$, such that A is ancestral to $D_1, \ldots D_k$, and we are looking for a median M, whose gene content and orthology relation to the given genomes are provided, that minimizes the sum of the directed distances between A and M, and M and the D_is, in the distance model defined in [10]. In Sect. 3, we prove that this median problem is NP-hard even when $k = 2$. In Sect. 4, we describe a simple Integer Linear Program (ILP) for this problem, based on a reduction to a colored MWM problem. We provide in Sect. 5 experimental results on simulated data.

2 Preliminaries

Genes and Genomes. A genome consists of a set of chromosomes, each being a linear or circular ordered set of oriented genes. Following the usual encoding

of gene orders, we represent a genome by its *gene extremity adjacencies*. In this representation, a gene g is represented using a pair of gene extremities (g_t, g_h), g_t denotes the tail of the gene g and g_h denotes its head, and an *adjacency* is a pair of gene extremities that are adjacent in a genome. If a gene g_i is denoted with a subscript, we will denote the tail of g_i by $g_{i,t}$ and its head by $g_{i,h}$. A gene extremity is *free* if it does not belong to an adjacency.

We assume that a given gene g can have multiple copies in a genome, the number of copies being called its *copy number*. A genome in which every gene has copy number 1 is a *trivial genome*. A non-trivial genome sometimes cannot be represented unambiguously by its adjacencies, that can form a *multi-set*, unless we distinguish the copies of each gene, for example by denoting the copies of a gene g with copy number k by $g^1, \ldots, g^k$. Nevertheless, we identify a genome with its multi-set of gene extremity adjacencies, which we call adjacencies from now. A *chromosome* is a maximal contiguous sequence of genes; a chromosome with k genes can have either $k-1$ adjacencies, in which case it is a *linear* chromosome, or k adjacencies, in which case it is a *circular* chromosome.

Evolutionary Model. In this work, following [10], we consider a model of *directed evolution* in which, when comparing two genomes, we assume one, denoted by A, is a trivial genome and an ancestor of the other genome, denoted by D.

We now describe the evolutionary events defining our evolutionary model. Genome rearrangements are modeled by *Single-Cut-or-Join* (SCJ) operations, which either delete an adjacency from a genome (a cut) or join a pair of free gene extremities (a join), thus forming a new adjacency. For duplication events, we consider two types of duplications, both creating an extra copy of a single gene: *Tandem Duplications* (TD) and *Floating Duplications* (FD). A tandem duplication of an existing gene g introduces an extra copy of g, say g', by adding an adjacency $g_h g'_t$, and, if there was an adjacency $g_h x$ by replacing it by the adjacency $g'_h x$. A floating duplication introduces an extra copy g' of a gene g as a single-gene circular chromosome by adding the adjacency $g'_h g'_t$.

Given A and D, we denote by gene family all copies of a given gene observed in A and D. By definition, there is exactly one copy of the gene in A and there might be several, paralogous, copies of the gene in D. We assume here that every gene in A has at least one descendant gene in D and conversely, every gene in D has exactly one ancestral gene in A, so we do not consider gene gains or losses.

Problem Statements. In [10], Feijão *et al.* introduced the **directed SCJ-TD-FD (d-SCJ-TD-FD) distance problem** that asks to compute the minimum number of SCJ, TD and FD operations needed to transform A into D, denoted by $d_{\mathrm{DSCJ}}(A, D)$. They showed that this problem is tractable and that the distance can be computed using a simple set-theoretical formula, extending naturally the distance formula for the SCJ with no duplication model.

A first median problem was also introduced in [10], the **directed SCJ-TD-FD (d-SCJ-TD-FD) median problem**, defined as follows: given $D_1, \ldots, D_k$ ($k \geq 2$) (possibly) non-trivial genomes, such that no gene family is absent from any D_i, compute a trivial genome A on the same set of gene families, that

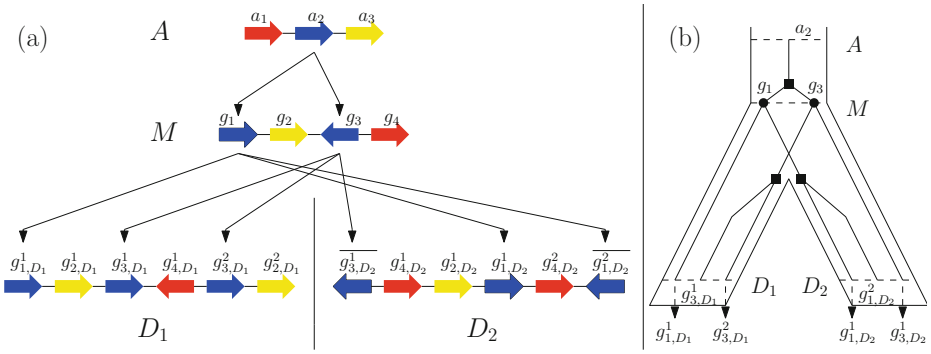

Fig. 1. In part (a), each color represents a gene family from A. Notice that each gene in D_1 and D_2 can be traced to a unique gene in M whereas a gene from A might have multiple daughters in M. Part (b) displays the gene tree of the gene family in blue (indicated by arrows in part (a)). Since the gene a_2 undergoes duplication (dark squares) to form g_1 and g_3 in M, M is not trivial w.r.t A. (Color figure online)

minimizes $\sum_{i=1}^{k} d_{\text{DSCJ}}(A, D_i)$. It was shown that this median problem is also tractable through a simple reduction to a MWM problem.

In the present work, we introduce the **rooted SCJ-TD-FD (r-SCJ-TD-FD) median problem**. We are given $k+1 \geq 3$ genomes, A, $D_1, \ldots, D_k$ such that A is a trivial genome, ancestor to the D_i's. The goal of the rooted median problem is to find a genome M which is a descendant of A and an ancestor of $D_1, \ldots, D_k$, minimizing the sum of its distance to A and to the $D_i's$. Following the approach introduced in [10], we assume we are given the *gene content* Γ of M and the *orthology relations* between A and M, as well as between M and the $D_i's$. This implies that every gene of M (resp. $D_1, \ldots, D_k$) has a unique ancestor in A (resp. in M), so M is a trivial genome compared to the $D_i's$ but might not be compared to A (see Fig. 1 for an illustration). To formally handle this difference, we assume that all copies of a gene g of A in M (i.e. the genes of M whose ancestor in A is gene g) are distinguishable (e.g. labeled, say $g_1, \ldots, g_k$) and, for a given gene g_i of M, we denote its ancestor in A by $a(g_i)$. Then for a given genome M on Γ, we denote by M_a the genome where every gene g is relabeled by $a(g)$. The goal of the rooted median problem is to find a genome M that minimizes the following function:

$$d_{\text{DSCJ}}(A, M_a) + \sum_{i=1}^{k} d_{\text{DSCJ}}(M, D_i). \tag{1}$$

Remark 1. If we assume there is no duplication from A to M, i.e. both have the same gene content, then the MWM algorithm introduced in [10] for the directed median problem applies to the rooted median problem and the problem is thus tractable. So the difficulty in solving the rooted median problem is to account for duplications from A to M.

The Pairwise Distance Formula. Given a gene $g \in \Gamma$, we call a *g-tandem array* a sequence of consecutive adjacencies $g_h g_t$; if this sequence forms a circular chromosome, it is called a *g-chromosome*. Given a genome X, we call an adjacency $g_h g_t$ an *observed duplication* if g has more than one copy in X. Observed duplications are part of a g-tandem array or a g-chromosome. Let $r(X)$ be the genome obtained from X by successively deleting an observed duplication from X, chosen arbitrarily, until there remains no observed duplication. Note that this corresponds to deleting every $g_h g_t$ adjacency, except that we keep one in the special case in which all copies of g are organized in g-chromosomes, as shown in Fig. 2. We call $r(X)$ the *reduced* genome of X. We define $t(X) = |X - r(X)|$, the number of adjacencies to delete to transform X into $r(X)$. Formally, the multi-set difference $X - Y$ between two multi-sets X and Y of adjacencies is the multi-set obtained as follows: it contains k copies of a given adjacency if and only if X contains exactly k more occurrences of this adjacency than Y (with $k = 0$ being possible).

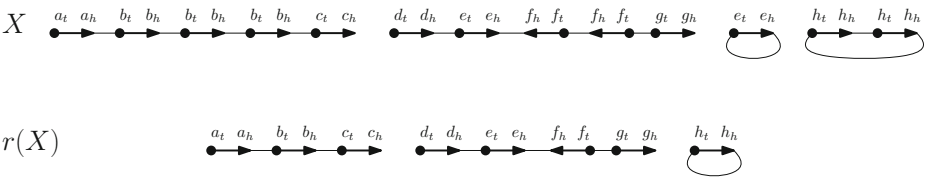

Fig. 2. An example of the reduced genome $r(X)$, of the genome X. Note that an instance of $h_h h_t$ is retained so that $r(X)$ contains at least one representative of gene family h. All observed duplications are removed in $r(X)$. Here, $t(X) = |X - r(X)| = 5$.

The directed SCJ-TD-FD distance between an ancestral genome A and a descendant genome D is given by [10]:

$$d_{\mathrm{DSCJ}}(A, D) = |A - r(D)| + |r(D) - A| + 2\delta(A, r(D)) + t(D) \tag{2}$$

where $\delta(A, r(D))$ is the difference between the number of genes of $r(D)$ and the number of genes of A (i.e. the number of duplications from A to $r(D)$). We introduce[1] now a slightly different formulation of d_{DSCJ} that will be useful in our hardness proof:

$$d_{\mathrm{DSCJ}}(A, D) = |A - r(D)| + |r(D) - A| + 2\delta(A, D) - t(D) \tag{3}$$

Remark 2. For $d_{\mathrm{DSCJ}}(M, D_i)$, the value of $t(D_i)$ does not depend on our choice of M, for $i = 1, \ldots, k$. We will therefore assume that the $D_i's$ are reduced (hence we may refer to $r(D_i)$ as simply D_i instead). However $t(M_a)$ has an impact on $d_{\mathrm{DSCJ}}(A, M_a)$, and so we will not assume that M is reduced.

[1] The proof is given in the Appendix.

3 The Rooted Median Problem Is NP-hard

We show that finding the optimal gene order for M is NP-hard even for $k = 2$, by reduction from the 2P2N-3SAT problem [2][2]. In 2P2N-3SAT, we are given n variables $x_1, \ldots, x_n$ and m clauses $C_1, \ldots, C_m$, each containing exactly 3 literals. Each x_i variable appears as a positive literal in exactly 2 clauses, and as a negative literal in exactly 2 clauses. Note that since each variable occurs in exactly 4 clauses and each clause has 3 literals, $m = 4n/3$. An example of a 2P2N-3SAT instance is shown in Fig. 3.

We now describe how we transform the x_i variables and C_j clauses into an instance of the rooted median. The genes of M are

$$\Gamma = \{g_1^+, \gamma_1^+, g_1^-, \gamma_1^-, \ldots, g_n^+, \gamma_n^+, g_n^-, \gamma_n^-, c_1, \ldots, c_m, \alpha_1, \ldots, \alpha_{2n-m}\}$$

The genes $g_i^+, \gamma_i^+, g_i^-, \gamma_i^-$ correspond to the x_i variable, and c_j to the clause C_j. The purpose of the $2n - m = 2n/3$ special α_i genes will become apparent later.

To simplify matters, every adjacency in our reduction is between the tails of two genes. Hence, the heads of each gene of A, D_1 and D_2 are telomeres (linear chromosomes extremities), so that all chromosomes are linear and have at most 2 genes. From now, we will omit the t subscript from the extremities for these adjacencies, with the understanding that every adjacency is between tails; for instance, we may write $g_i^+\gamma_i^+$ for the adjacency $g_{i,t}^+\gamma_{i,t}^+$.

We can now describe A, D_1 and D_2. The genes of A are $g'_1, \gamma'_1, \ldots, g'_n, \gamma'_n$, $c'_1, \ldots, c'_m, \alpha'_1, \ldots, \alpha'_{2n-m}$. The genes g_i^+ and g_i^- (resp. γ_i^+ and γ_i^-) are duplicates of g'_i (resp. γ'_i), and there are no other duplications in M compared to A. Formally, for each $i \in [n]$, put $a(g_i^+) = a(g_i^-) = g'_i$, $a(\gamma_i^+) = a(\gamma_i^-) = \gamma'_i$ and for each $j \in [m]$, put $a(c_j) = c'_j$. Finally, for each $i \in [2n - m]$, put $a(\alpha_i) = \alpha'_i$. The adjacencies of A are $\{g'_i\gamma'_i : i \in [n]\}$.

The genomes D_1 and D_2 are identical, i.e. they contain the same set of genes and of adjacencies. We simply describe the set of adjacencies of D_1 and D_2 with the understanding that if an extremity, say x, appears in two adjacencies xy and xz, then the two x are the tails of two distinct copies of the same gene on two distinct chromosomes. The adjacencies of D_1 and D_2 are described as follows.

- For each $i \in [n]$, add to D_1 and D_2 the adjacencies $g_i^+\gamma_i^+$ and $g_i^-\gamma_i^-$.
- For each $i \in [n]$, let C_{j_1}, C_{j_2} be the two clauses in which x_i occurs positively and let C_{k_1}, C_{k_2} be the two clauses in which x_i occurs negatively. Add to D_1 and D_2 the adjacencies $g_i^+c_{j_1}$ and $\gamma_i^+c_{j_2}$. Similarly, add to D_1 and D_2 the adjacencies $g_i^-c_{k_1}$ and $\gamma_i^-c_{k_2}$[3].
- Finally, for each $i \in [n]$ and each $j \in [2n - m]$, add to D_1 and D_2 the adjacencies $g_i^+\alpha_j, g_i^-\alpha_j, \gamma_i^+\alpha_j$ and $\gamma_i^-\alpha_j$.

[2] This problem is sometimes called the (3,B2)-SAT problem, where B2 indicates that the literals are *balanced* with two occurrences each.

[3] Intuitively, these adjacencies represent using a literal to satisfy a specific clause. For instance, the adjacency $g_i^+c_{j_1}$ represents "setting x_i to true and satisfying C_{j_1}".

This completes our construction. The intuition behind our hardness proof is that for each $i \in [n]$, we need to pick one of $g_i^+\gamma_i^+$ or $g_i^-\gamma_i^-$ in M, as we will show. Simultaneously, we would like to include as many adjacencies which are in both D_1 and D_2. It will possible to choose the positive and negative adjacencies *and* match all the c_j and α_j if and only if the 2P2N-3SAT instance is satisfiable.

It will be useful to think of D_1 (and D_2) as the set of adjacencies which are allowed to belong to M, as stated in the following.

Lemma 1. *Let a be an adjacency in M, such that $a \notin D_1$ (equivalently, $a \notin D_2$). Then $M - \{a\}$ achieves a smaller total distance to A, D_1 and D_2 than M.*

Proof. By cutting a, we increase the distance to A by at most 1, but decrease the distance to D_1 and D_2 by 1 each. This is because $|(M - \{a\}) - D_1| + |D_1 - (M - \{a\})| = |M - D_1| - 1 + |D_1 - M|$, the value of $\delta(M, D_1)$ is unchanged and $t(D_1) = 0$ by assumption (and the same holds for D_2). Therefore removing a from M yields a better median genome. □

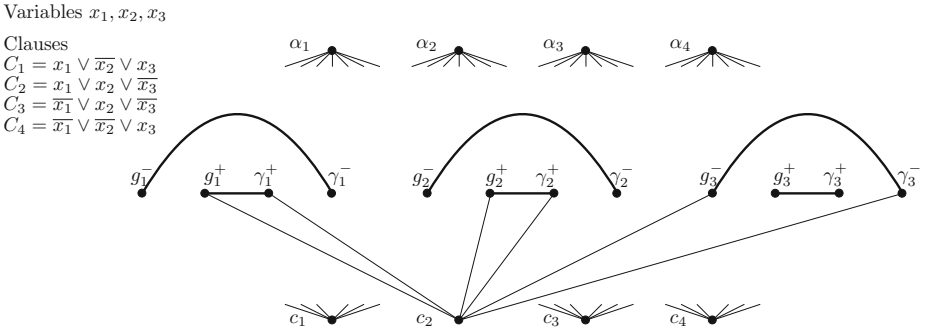

Fig. 3. An example of a 2P2N-3SAT instance, with an illustration of the genes of M (only the gene tails are shown) and the adjacencies that are allowed by D_1 and D_2. The fat edges represent pairs of adjacencies of which at least one must be present according to Lemma 2. Among the c_j extremities, only the adjacencies for c_2 are shown.

Therefore, we may assume that every adjacency of a median M belongs to D_1 and D_2. Note that this implies that M contains no observed duplications (with respect to A), as no such adjacency is in D_1 and D_2. Thus we will ignore the $t(M_a) = 0$ term in $d_{\mathrm{DSCJ}}(A, M_a)$ (Eq. (3)), and we will not make a distinction between M_a and $r(M_a)$, as these are equal.

Another property of M is that it must contain at least one "positive" or one "negative" adjacency for each $i \in [n]$.

Lemma 2. *For $i \in [n]$, M contains at least one of $g_i^+\gamma_i^+$ and $g_i^-\gamma_i^-$.*

The proof of this lemma is provided in the Appendix.

We now formally prove the hardness of computing the SCJTDFD median.

Theorem 1. *The rooted SCJ-TD-FD median problem is NP-hard.*

Proof. Let $x_1, \ldots, x_n$ and $C_1, \ldots, C_m$ be a 2P2N-3SAT-instance, and let A, D_1, D_2 and the genes Γ of M be the corresponding instance of the r-SCJ-TD-FD median genome problem. We will show that the given 2P2N-3SAT instance is satisfiable if and only if there exists a median genome M satisfying

$$d_{\text{DSCJ}}(A, M_a) + d_{\text{DSCJ}}(M, D_1) + d_{\text{DSCJ}}(M, D_2) \leq 2|D_1| - 2n + 4\delta(M, D_1)$$

($\Rightarrow$) Suppose that the 2P2N-3SAT can be satisfied by an assignment of the x_i variables to true or false. Construct a median genome using the following steps.

1. For each $i \in [n]$, if x_i is set to true, then add $g_i^- \gamma_i^-$ to M, and if instead x_i is set to false, add $g_i^+ \gamma_i^+$ to M.
2. Then, add to M these adjacencies in an algorithmic fashion: for each $j = 1, 2, \ldots, m$, consider clause C_j and let x_i be any variable satisfying C_j.
   - If x_i is set to true, then note that g_i^+ and γ_i^+ have not been matched in Step 1. Add $g_i^+ c_j$ to M if g_i^+ is not part of an adjacency of M yet, or add $\gamma_i^+ c_j$ to M otherwise.
   - If instead x_i is set to false, then g_i^- and γ_i^- have not been matched in Step 1. Add $g_i^- c_j$ if g_i^- is not part of an adjacency in M yet, or add $\gamma_i^- c_j$ to M otherwise.

   Note that since each x_i can satisfy at most two clauses, it will always be possible to find an extremity to match c_j with.
3. Finally, observe that so far each of the g_i^+, g_i^-, γ_i^+ and γ_i^- extremities are in an adjacency M, except $4n - 2n - m = 2n - m$ of them. Associate each such extremity g with a distinct α_j extremity arbitrarily, and add each $g\alpha_j$ to M, noting that there are just enough α_j genes to do so.

Note that M contains $n + m + 2n - m = 3n$ adjacencies in total, exactly n of which correspond to an adjacency of A (those included in Step 1). Also, every adjacency of M occurs in both D_1 and D_2. We have

$$\begin{aligned} d_{\text{DSCJ}}(A, M_a) &= |A - M_a| + |M_a - A| + 2\delta(A, M_a) - t(M_a) \\ &= 0 + 2n + 2n - 0 = 4n \end{aligned}$$

As for D_1 and D_2,

$$\begin{aligned} d_{\text{DSCJ}}(M, D_1) = d_{\text{DSCJ}}(M, D_2) &= |D_1 - M| + |M - D_1| + 2\delta(M, D_1) \\ &= |D_1| - 3n + 0 + 2\delta(M, D_1) \end{aligned}$$

Therefore the total distance is $4n + 2(|D_1| - 3n + 2\delta(M, D_1)) = 2|D_1| - 2n + 4\delta(M, D_1)$, as we predicted.

($\Leftarrow$) Suppose that there exists a median genome M of total distance at most $2|D_1| - 2n + 4\delta(M, D_1)$. By Lemma 1, we may assume that every adjacency of M is present in both D_1 and D_2.

With the next two claims, we will prove that M has exactly $3n$ adjacencies, of which exactly n are adjacencies corresponding to those in A.

Claim 1. *$|M| \leq 3n$, and $|M| = 3n$ only if every c_j and α_j extremity is in some adjacency of M.*

For the rest of the proof, denote by q the number of distinct adjacencies $ab \in A$ for which there exists $xy \in M$ such that $a(x)a(y) = ab$.

Claim 2. *$|M| = 3n$ and $q = n$.*

The proofs of both claims will be discussed in detail in the Appendix.

Because $q = n$, Claim 2 implies that for each $i \in [n]$, (at least) one of $g_i^+\gamma_i^+$ and $g_i^-\gamma_i^-$ is in M. This lets us define as assignment for our 2P2N-3SAT instance: for each $i \in [n]$, set x_i to *true* if $g_i^-\gamma_i^-$ is in M, and otherwise set x_i to *false*. We claim this this assignment satisfies every clause.

To see this, let C_j be a clause and let c_j be its corresponding extremity in M. By Claim 2, every extremity that is part of some adjacency in D_1 must be part of an adjacency in M, including c_j. Thus there is some e such that $c_je \in M$. By Lemma 1, the adjacency c_je must also be in D_1, and by construction either (1) $e \in \{g_i^+, \gamma_i^+\}$ for some x_i that occurs positively in C_j, or (2) $e \in \{g_i^-, \gamma_i^-\}$ for some x_i that occurs negatively in C_j. Suppose that case (1) applies. Then $c_jg_i^+$ or $c_j\gamma_i^+$ being in M means that $g_i^+\gamma_i^+ \notin M$, implying in turn that $g_i^-\gamma_i^-$ is in M. In this situation, we have set x_i to *true* and we satisfy C_j. Suppose instead that case (2) applies. Then $g_i^-\gamma_i^- \notin M$, in which case we have set x_i to false and satisfy C_j. As the argument applies to any clause C_j, this concludes the proof. □

Remark 3. In the reduction above, none of the considered genomes contain a g-tandem array or a g-chromosome. So our result also implies the hardness of the rooted median problem where the distance between two genomes A and D, where A is an ancestor of D, is computed in a simpler way as $|A-D|+|D-A|+2\delta(A, D)$, i.e. does not contain a term related to reducing the descendant genome.

4 An Integer Linear Program

We now describe a simple Integer Linear Program (ILP) to solve the rooted median problem. The key idea, already used in previous median problems [10,22] is to convert the rooted median problem into an instance of a MWM problem, albeit with certain additional constraints. More precisely, in this approach we define a complete graph G on the extremities g_h and g_t of every gene g in Γ. A pair of distinct extremities defines an edge and thus a potential adjacency in M, which is thus defined by a matching in G. Each edge is assigned a weight that reflects the number of descendant genomes which contain the corresponding adjacency. Further, each edge is assigned a color that reflects its corresponding adjacency in the ancestral genome, if any, and the number of colors of the selected edges also contributes to the weight of the matching defining the median M.

An Alternative Formulation for the Distance. We first introduce an alternative formula to compute the directed distance, denoted by $d_{\text{DSCJ}}(u, v)$, from an ancestor u to a descendant v. For the rooted median problem, the pair (u, v) can represent either the pair (A, M_a) or any pair (M, D_i). The new formulation is easier to handle in an ILP framework than Eq. (3). We denote by $n_v(g)$ the number of copies of gene g in v, by $n_v(g_h g_t)$ the number of occurrences of adjacency $g_h g_t$ in v, and by $t_v(g)$ the number of observed duplications of gene g in v. Note that $t_v(g) \in \{n_v(g_h g_t) - 1, n_v(g_h g_t)\}$, the case $t_v(g) = n_v(g_h g_t) - 1$ occurring when adjacencies $g_h g_t$ form only g-chromosomes. Further, let $t(v) = \sum_{g \in \Gamma_u} t_v(g)$ denote the total number of observed duplications in v, where Γ_u is the set of genes of u and also the alphabet of genes of v.

To rewrite $d_{\text{DSCJ}}(u, v)$, we introduce an indicator variable $\alpha_{g,uv}$, where $\alpha_{g,uv} = 1$ if $g_h g_t$ is common to both u and v, but all occurrences were removed while reducing v. Formally, $\alpha_{g,uv} = 1$ if $g_h g_t \in u \cap v$ and $g_h g_t \notin r(v)$; otherwise $\alpha_{g,uv} = 0$. It is then relatively straightforward to show[4] that

$$d_{\text{DSCJ}}(u, v) = |u - v| + |v - u| + 2\delta(u, v) - 2t(v) + 2 \sum_{g \in \Gamma_u} \alpha_{g,uv} \tag{4}$$

This formulation is interesting due to the fact it does not rely on the notion of a reduced genome. We will discuss later how variables $\alpha_{g,uv}$ and $t_v(g)$ can be handled simply in an ILP framework.

Reformulating the Objective Function. We now use Eq. (4) to reformulate the objective function of the rooted median problem[5].

Claim 3. *Minimizing the function Eq. (1) defining the evolutionary cost of a median M is equivalent to maximizing the following expression:*

$$\sum_{i=1}^{k} \left(2|M \cap D_i| - 2 \sum_{g \in \Gamma_M} \alpha_{g,MD_i} \right) + 2|A \cap M_a| + 2t(M_a) - 2 \sum_{g \in \Gamma_A} \alpha_{g,AM_a} - (k+1)|M| \tag{5}$$

where Γ_A and Γ_M are the set of genes of A and M, respectively, and so also the gene alphabets for M and the D_is, and variables α_{g,AM_a} and α_{g,MD_i} are defined as $\alpha_{g,uv}$ above.

Such a reformulation of the objective function is inspired by [10]. This revision enables us to translate the problem as an instance of a *colored MWM problem*, as will be made clear in the subsequent paragraphs.

An Interpretation as a Colored MWM Problem. The terms $\alpha_{g,uv}$ and $t(M_a)$ in Eq. (5) account for the presence of observed duplications. In the absence of observed duplications however, solving the rooted median problem requires finding a matching in G that maximizes the sum of the weight of the selected edges

[4] A proof is provided in the Appendix.

[5] The proof of this claim is discussed in the Appendix.

and of the number of colors represented by the matching edges. The matching edges weight is partly accounted for by the term $|M \cap D_i|$, while on the other hand, $|A \cap M_a|$ determines the number of colors used in the matching. Using the intersection terms in the objective function, we now interpret the notion of *weight* and *color* of an edge in terms of decision variables of an ILP.

In order to compute $|M \cap D_i|$, we introduce the variable $\gamma_i(e)$ denoting the existence of a potential adjacency e of M in a genome D_i: we put $\gamma_i(e) = |e \cap D_i|$, i.e. $\gamma_i(e) = 1$ if $e \in D_i$ and 0, otherwise. For each adjacency e in the graph G, the weight $w(e)$ of e is determined using the weight function $w : E(G) \to \mathbb{N}$:

$$w(e) = 2\left(\sum_{i=1}^{k} \gamma_i(e)\right) - (k+1)$$

Since M is trivial w.r.t. every D_i, the weights for edges $e \in M$ will account for the term $\sum_{i=1}^{k} 2|M \cap D_i| - (k+1)|M|$ in Eq. (5). However, this principle does not work with A. Indeed, it is possible that $x_1y_1 \in M$ and $x_2y_2 \in M$ such that $a(x_1)a(y_1) = a(x_2)a(y_2) \in A$. In this situation, only one of x_1y_1 or x_2y_2 can contribute to $|A \cap M_a|$, but both $|x_1y_1 \cap A|$ and $|x_2y_2 \cap A|$ equal to 1. In other words, we cannot simply sum the adjacencies of M_a which are in A.

To address this issue, we introduce the notion of a *color family*. Let m_A be the number of adjacencies in A. Each number from the set $\{1, 2, ..., m_A\}$ represents a distinct color. We arbitrarily assign a distinct color from this set to each adjacency in A. If $E(G)$ is the edge set of G, representing all possible adjacencies in M, then every adjacency in $E(G)$ is assigned a color from $\{1, 2, ..., m_A\} \cup \{0\}$, consistent with the orthology relations: the adjacency $xy \in M$ receives color $i \neq 0$ if the adjacency $a(x)a(y)$ is present in A and was assigned color i, and color 0 if $a(x)a(y)$ is not present in A. The set of adjacencies having the same color i form a color family, represented by E_i. We denote by C the coloring function $E(G) \to \{0, 1, ..., m_A\}$ defined as described above. Notice that a color i contributes exactly once to the term $|A \cap M_a|$ if there exists at least one adjacency in M that belongs to the color family i.

Reducing the Size of the ILP. The size of the ILP we are about to describe is polynomial in the sum of the considered genomes. As the total number of adjacencies is quadratic in the number of genes in M, it can reach large values when dealing with large genomes, thus making the ILP challenging to solve in practice. We show that the set of decision variables can be restricted to specific adjacencies, which we call *candidate adjacencies*.An adjacency xy is a *candidate adjacency* for the median if at least $\lfloor \frac{k+1}{2} \rfloor + 1$ genomes from the set $\{A, D_1, D_2, ..., D_k\}$ contain xy (where here A contains xy if $a(x)a(y) \in A$). Lemma 3, proved in the Appendix, shows that the number of adjacencies to consider in an ILP is linear in the sum of the sizes of the input genomes.

Lemma 3. *There exists an optimal median consisting of only candidate adjacencies. Furthermore, when k is even, an adjacency which is not a candidate adjacency can not be a part of any optimal median.*

Remark 4. The difficulty of the rooted median problem stems from the fact that duplication from M to the D_is can create conflicting adjacencies, where a median gene extremity belongs to several candidate adjacencies. It is interesting to observe that this can happen only due to convergent evolution, i.e. the fact that the same adjacency is created independently in several D_is. This suggests that in the practical context of a limited level of convergent evolution, the rooted median problem is easy to solve.

The ILP for the Rooted Median Problem. We can now provide the complete ILP formulation to solve the rooted SCJ-TD-FD median problem. Let $x(e)$ be a binary decision variable denoting the inclusion of edge (candidate adjacency) $e \in E(G)$ in M. Also, let c_i be a binary decision variable indicating if at least one edge with color i belongs to M. From the previous paragraph, one can write the objective function as

Maximize:

$$\sum_{e \in E(G)} w(e)x(e) + 2\sum_{i=1}^{m_A} c_i + 2t(M_a) - 2\sum_{g \in \Gamma_A} \alpha_{g,AM_a} - 2\sum_{i=1}^{k} \sum_{g \in \Gamma_M} \alpha_{g,MD_i}$$

We now describe the constraints of the ILP. The first set of constraints concern the *consistency* of the set of chosen adjacencies, that ensures that each gene extremity in M belongs to at most one adjacency, or in other words that M is a matching for the graph G (these are the first two sets of constraints below). Next, we use an additional set of constraints to determine the values of c_i, $i = \{1, 2, ..., m_A\}$. If at least one adjacency of color i is present in the median, $c_i = 1$, otherwise $c_i = 0$. The following inequalities define these color constraints:

$$\sum_{e=(y_h,z)} x(e) \leq 1 \qquad \forall y \in \Gamma_M \tag{6}$$

$$\sum_{e=(y_t,z)} x(e) \leq 1 \qquad \forall y \in \Gamma_M \tag{7}$$

$$c_i = \left\lceil \frac{\sum_{C(e)=i} x(e)}{|E_i|} \right\rceil \qquad \forall i \in \{1, 2, ..., m_A\} \tag{8}$$

Note that for c_i above, the constraints of the type $x = \lceil y \rceil$ are not linear, but if x is restricted to be in $\{0, 1\}$, it can be replaced by the constraint $y \leq x \leq y+\epsilon$, where ϵ is very close to 1, say 0.999. A similar trick can be used for floor functions.

In order to compute $\alpha_{g,uv}$ for every pair (u, v) – where either $u = A,\ v = M_a$ or $u = M, v = D_i$ for some i – and every gene $g \in \Gamma_u$, we use some additional constraints. Let $p_v(e)$ be the binary variable denoting if the adjacency e exists in v. We use an indicator variable $\lambda_{g,uv}$ such that $\lambda_{g,uv} = 1$ if and only if all copies of g are involved in $g_h g_t$ adjacencies. Consequently, $\lambda_{g,uv} = 1$ ensures the existence of the $g_h g_t$ adjacency in $r(v)$. Thus, $\lambda_{g,uv} = \left\lfloor \frac{n_v(g_h g_t)}{n_v(g)} \right\rfloor$. Further, we use $\Lambda_{g,uv}$ to indicate if at least one instance of $g_h g_t$ has been observed in v. Thus, we

can represent $\Lambda_{g,uv}$ as $\left\lceil \frac{n_v(g_h g_t)}{n_v(g)} \right\rceil$. Since we already know the gene orders of A and each D_i, the values of $p_A(e)$ and $p_{D_i}(e)$ are known. Further, $p_M(e) = x(e)$. Thus, we obtain the following constraints for every gene g and branch (u, v):

$$\lambda_{g,uv} = \left\lfloor \frac{n_v(g_h g_t)}{n_v(g)} \right\rfloor \tag{9}$$

$$\Lambda_{g,uv} = \left\lceil \frac{n_v(g_h g_t)}{n_v(g)} \right\rceil \tag{10}$$

$$\alpha_{g,uv} = \min(p_u(g_h g_t), \Lambda_{g,uv} - \lambda_{g,uv}) \tag{11}$$

$$t_v(g) = n_v(g_h g_t) - \lambda_{g,uv} \tag{12}$$

We use the fact that if $g_h g_t \notin v$ for some g then $g_h g_t \notin r(v)$. Thus, if $g_h g_t \notin v$, $\lambda_{g,uv} = 0$ thereby ensuring the correctness of constraints to find $\alpha_{g,uv}$. Again, note that the min function is not linear, but that a constraint $x = \min(y, z)$ can be replaced by $x \geq y$ and $x \geq z$, assuming that $x, y, z \in \{0, 1\}$.

5 Experimental Results

We ran experiments on simulated data in order to evaluate the ability of the ILP to correctly predict the gene order of the median genome. The input for the program, including gene orders for the ancestor genome A and the descendant genomes D_i, along with the orthology relations, generated using the ZOMBI genome simulator [8]. The ILP was solved using the Gurobi solver.

Simulations Parameters. Our input genomes consisted of one ancestor A and two descendants D_1 and D_2. We started with the ancestral genome A as a single circular chromosome consisting of 1000 genes, belonging to different gene families (so without duplicate genes). The genome A evolved into the median genome M using duplications, inversions and translocations. The genome M was further evolved along two independent branches to yield the descendant genomes, D_1 and D_2. The total number of rearrangements (inversions + translocations) from A to M and from M to D_i was varied from 100 to 500, in steps of 100. The parameter for duplication events was kept constant throughout the experiments. The average number of duplicated genes, over all three branches collectively, was found to be 362.8 with a standard deviation of 82 genes. Considering the number of duplication events, the mean and standard deviation of segmental duplications over the three branches was 72.6 and 15.8 respectively. The lengths of segmental duplications, inversions and translocations were controlled using specific extension rates. These extension rates (all between 0 and 1) are the parameters of a geometric distribution dictating the respective lengths. Thus, the length of the segment being acted upon would be 1 if the extension rate parameter is set to 1 and would increase as the parameter value reduces. In our experiments, the inversion, translocation and duplication extension rates were 0.05, 0.3 and 0.2 respectively. For each setting (number of rearrangements) we ran 40 simulations.

Results. For each simulation, we compared the optimal median according to the ILP to the actual median generated by the simulator. For each group, we measured the average precision and recall statistics. The ILP predicts the median genome in the form of its adjacency set. Thus, in this context, precision refers to the ratio of number of correctly predicted adjacencies to the total number of adjacencies in the computed optimal median. On the other hand, recall represents the ratio of the correctly predicted adjacencies to the total number of adjacencies in the actual median. For each instance, we measured the number of candidate adjacencies used in the ILP. Additionally, to evaluate the effectiveness of our approach, we also measured the number of adjacencies in the solution which were common to all genomes (A, D_1 and D_2) and those common to only two of the three.

An overview of the results is given in Table 1. The ILP rarely predicts an erroneous adjacency to be a part of the optimal median, with a near-perfect precision. This property is observed throughout the experiments irrespective of the number of rearrangement events. On the other hand, the ILP predicts more than 90% of the median for lower rates of rearrangement and a decreasing trend is observed as the number of rearrangement events increase. This can be partly attributed to the decrease in the number of candidate adjacencies. In general, the number of candidate adjacencies is lower than the true number of adjacencies in the median, as including other adjacencies may result in a non-optimal median. This, however, emphasizes the practicality of Lemma 3, as the number of adjacency variables is significantly reduced. It can also be observed that the number of adjacencies common to all genomes decreases with increase in rearrangements. These adjacencies will be preferred by the ILP on account of higher weight.

Table 1. Statistics of the ILP median experiment on simulated data.

Events	Adj. in true median	Cand. adj.	Adj. in ILP median	Precision	Recall	% Adj. common to all genomes	% Adj. common to two genomes	No. of optimal solutions	Avg. time per run (in sec)
100	1514	1503	1493	0.9998	0.9859	86.43	13.57	2.3	53
200	1107	1062	1044	0.9991	0.9428	69.49	30.51	15.8	29
300	1312	1192	1155	0.9985	0.8758	52.94	47.06	40.3	38
400	1151	985	961	0.9981	0.8329	49.44	50.56	393.7	51
500	1430	1174	1132	0.9972	0.7897	46.68	53.32	3682.6	84

Another notable observation is the increase in the number of optimal solutions with larger rates of rearrangement. This correlates naturally with the decrease in the number of adjacencies which are common to all genomes. For only 100 rearrangements, the ILP outputs a unique optimal median in most runs, with an overall average of 2.3 solutions. However, the average number of

optimal solutions exceeded 3000 in case of 500 rearrangements. Despite a pool of optimal solutions, the SCJ distance between the actual median and an optimal median does not vary by much. If the SCJ distance between the actual median and a randomly chosen optimal median is D, then the distance between the actual median and any other optimal median was observed to stay within the range $(D-2, D+2)$. For most of our simulations, the ILP output an optimal median in under a minute, with the exception of the case with 500 rearrangement events.

6 Conclusion

In this chapter, we introduced the directed and rooted median problems and studied them under the SCJ-TD-FD model. We proved that computing the median with the most parsimonious directed distance for an ancestor A and descendants D_i, $i = 1$ to k is NP-hard by reduction from the 2P2N-3SAT problem. This contrasts with the directed median problem which does not involve an ancestral genome A. An interesting feature of our hardness proof is that it relies on two identical descendant genomes, showing a sharp tractability boundary between the directed pairwise distance problem and the rooted median of three genomes problem. Similarly to other SCJ-related median problems, our rooted median problem aims at selecting adjacencies among candidate adjacencies which are seen in a majority of the given input genomes; nevertheless the possibility of conflicting median adjacencies due to convergent evolution is at the heart of the intractability of the problem (Remark 4). To address this intractability, we provide a simple Integer Linear Program that computes an optimal median. Without surprise, we observe that our ILP outputs a more reliable estimate of the median in case of lower rates of rearrangements. Moreover, we observe that despite having many more optimal solutions for higher rates of rearrangement, the distance of a random solution from the actual median does not deviate by much.

Our work can be commented with regard to the Small Parsimony Problem under the directed SCJ-TD-FD model. The hardness result of the rooted median problem likely implies the corresponding SPP problem is also NP-hard. This motivates our current work about extending the rooted median ILP toward the SPP. It is worth noting that our median ILP can also be used to solve the SPP by iterative application from an initial assignment of ancestral gene orders, similarly to the early SPP solvers for genome rearrangements such as GRAPPA [18]. Considering the multiplicity of the solutions, it also remains to be investigated if the sampling and subsequent analysis of co-optimal evolutionary scenarios, in a similar manner as [16], is possible within this framework.

Acknowledgments. CC is supported by Natural Science and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2017-03986. CC and PF are supported by CIHR/Genome Canada Bioinformatics and Computational Biology grant B/CB 11106. Most computations were done on the Cedar system of Compute Canada through a resource allocation to CC.

Appendix

Proof of Eq. (3). We remind that the original pairwise distance formula (Eq. (2)) is

$$d_{\mathrm{DSCJ}}(A, D) = |A - r(D)| + |r(D) - A| + 2\delta(A, r(D)) + t(D)$$

and we want to prove it is equivalent to

$$d_{\mathrm{DSCJ}}(A, D) = |A - r(D)| + |r(D) - A| + 2\delta(A, D) - t(D).$$

Notice that the $2\delta(A, r(D))$ term from the original formula was switched for the $2\delta(A, D)$ term. Consider the difference in the number of genes from D to $r(D)$. Each time we remove a $g_h g_t$ observed duplication from D while reducing it, it corresponds to removing a copy of g from D. Thus D has $t(D)$ more genes than $r(D)$, so that $2\delta(A, D) = 2\delta(A, r(D)) + 2t(D)$. This implies $2\delta(A, D) - t(D) = 2\delta(A, r(D)) + t(D)$. □

Proof of Lemma 2. Suppose that for some i, M contains none of $g_i^+\gamma_i^+$ or $g_i^-\gamma_i^-$. Note that M does not contain $g_i^+\gamma_i^-$ nor $g_i^-\gamma_i^+$, by Lemma 1. This implies that $g_i'\gamma_i' \notin M_a$, as we have excluded all the four possibilities of having this adjacency in M_a.

Consider the median M' obtained from M by adding $g_i^+\gamma_i^+$, cutting the adjacencies that g_i^+ and γ_i^+ were contained in, if needed. If g_i^+ and γ_i^+ are both telomeres in M, then it is easy to check that $M' = M + g_i^+\gamma_i^+$ (M augmented by the adjacency $g_i^+\gamma_i^+$) attains a better distance than M since $g_i^+\gamma_i^+ \in D_1, D_2$ and $a(g_i^+)a(\gamma_i^+) = g_i'\gamma_i' \in A$ (this decreases the distance by 3).

Suppose that $g_i^+ x \in M$ for some x, and that γ_i^+ is a telomere in M. By Lemma 1, $g_i^+ x$ is in both D_1 and D_2, which implies that $x = c_j$ or $x = \alpha_j$ for some j. This implies in turn that $a(g_i^+)a(x) \notin A$. We can argue that $M' = M - g_i^+ x + g_i^+\gamma_i^+$ is better. To see this, observe that $|M' - D_1| = |M - D_1|$ and $|D_1 - M'| = |D_1 - M|$ (and the same with D_2). On the other hand, recalling that $g_i'\gamma_i' \notin M_a$, we have $|M_a' - A| = |M_a - A| - 1$ (because $a(g_i^+)a(x) \notin A$ and $a(g_i^+)a(\gamma_i^+) \in A$) and $|A - M_a'| = |A - M_a| - 1$ (because $a(g_i^+)a(\gamma_i^+) \in A$). We have thus decreased the distance by 2. The same argument applies if g_i^+ is a telomere but γ_i^+ is not.

Finally, suppose that $g_i^+ x$ and $\gamma_i^+ y$ are adjacencies of M. As we argued above, $a(g_i^+)a(x) \notin A$ and $a(\gamma_i^+)a(y) \notin A$. Letting $M' = M - g_i^+ x - \gamma_i^+ y + g_i^+\gamma_i^+$, we find that $|M' - D_1| = |M - D_1|$ and $|D_1 - M'| = |D_1 - M| + 1$. As the same holds with D_2, we have increased the distance to D_1 and D_2 by 2. On the other hand, $|A - M_a'| = |A - M_a| - 1$ and $|M_a' - A| = |M_a - A| - 2$. To sum up, the total distance decreases by 1. □

Proof of Claim (1). Call an extremity e of a gene in Γ *matchable* if there exists an adjacency of D_1 that contains e. By Lemma 1, the adjacencies of M only contain matchable extremities. The g_i^+, g_i^-, γ_i^+ and γ_i^- extremities account for $4n$ matchable extremities. The c_j genes account for m matchable extremities and the α_j genes for $2n - m$ matchable extremities. Thus there are $4n + m + 2n - m = 6n$

matchable extremities. Because an adjacency contains 2 extremities, there can be at most $3n$ adjacencies in M. The second part of the claim follows from the fact that we have to assume that every c_j and α_j is matched to attain this bound. □

Proof of Claim (2). By the definition of q, we have $|A - M_a| = n - q$ and $|M_a - A| = |M| - q$. It follows that

$$\begin{aligned} d_{\mathrm{DSCJ}}(A, M_a) &= |A - M_a| + |M_a - A| + 2\delta(A, M_a) - t(M_a) \\ &= n - q + |M| - q + 2n - 0 \\ &= |M| + 3n - 2q \end{aligned}$$

Using Lemma 1, we also have $d_{\mathrm{DSCJ}}(M, D_1) = |M - D_1| + |D_1 - M| + 2\delta(M, D_1) = 0 + |D_1| - |M| + 2\delta(M, D_1)$. Thus the sum of the 3 distances is

$$|M| + 3n - 2q + 2|D_1| - 2|M| + 4\delta(M, D_1) \leq 2|D_1| - 2n + 4\delta(M, D_1)$$

(this inequality is due to our initial assumption on the total distance of M). After simplifying, this gives $5n \leq |M| + 2q$. By Claim 1, $|M| \leq 3n$ and because A has n adjacencies, $q \leq n$. Hence, this inequality is only possible if $|M| = 3n$ and $q = n$. □

Proof of Eq. (4). From Eq. (3), we have $d_{DSCJ}(u, v) = |u - r(v)| + |r(v) - u| + 2\delta(u, v) - t(v)$. However, it is easier to express the distance without the reduced genome terms. Hence, we eliminate the need for computing the reduced genomes by replacing $|u - r(v)|$ and $|r(v) - u|$ by suitable expressions as follows. We show that (1) $|u - r(v)| = |u - v| + \sum_{g \in \Gamma_u} \alpha_g$, and (2) $|r(v) - u| = |v - u| - t(v) + \sum_{g \in \Gamma_u} \alpha_g$. Substituting the terms in Eq. (3) yield Eq. (4).

(1) Consider first the difference between $u - r(v)$ and $u - v$. Suppose that $xy \in u - v$ but $xy \notin u - r(v)$. Then $xy \in r(v)$ but $xy \notin v$, which is not possible. Thus the difference can only be due to some $xy \in u - r(v)$ such that $xy \notin u - v$. This means that $xy \notin r(v)$ and $xy \in v$, which only happens when $xy = g_h g_t$ for some gene g. As we have $xy = g_h g_t \in u \cap v$ and $g_h g_t \notin r(v)$, we also have $\alpha_g = 1$, by definition. Since only one such adjacency is possible for each gene g (because u is trivial), $u - r(v)$ and $u - v$ differ only by adjacencies on genes for which $\alpha_g = 1$. We have shown that $|u - r(v)| = |u - v| + \sum_{g \in \Gamma_u} \alpha_g$.

(2) Now consider the difference between $r(v) - u$ and $v - u$. Note that there are $t(v)$ adjacencies in v not in $r(v)$, all observed duplications of the type $g_h g_t$. Let $g \in \Gamma_u$. If $g_h g_t \notin u$, then all of the $t(g)$ observed duplications in g are counted in $v - u$ but not in $r(v) - u$. This is also true when $g_h g_t \in u$ and $g_h g_t \in r(v)$. In these cases, $\alpha_g = 0$. However when $g_h g_t \in u \cap v$ but $g_h g_t \notin r(v)$, there are $t(g) - 1$ of the $g_h g_t$ adjacencies counted in $v - u$ not counted in $r(v) - u$ (this is because exactly one $g_h g_t$ adjacency of v can be matched with the $g_h g_t$ adjacency in u, and $r(v)$ has no such adjacency). This case occurs precisely when $\alpha_g = 1$. This shows that $|r(v) - u| = |v - u| - \sum_{g \in \Gamma_u} (t(g) - \alpha_g) = |v - u| - t(v) + \sum_{g \in \Gamma_u} \alpha_g$. □

Proof of Claim (3). By Eq. (4), we know that

$$d_{\mathrm{DSCJ}}(A, M_a) = |A - M_a| + |M_a - A| + 2\delta(A, M_a) - 2t(M_a) + 2\sum_{g \in \Gamma_A} \alpha_{g,AM_a}$$

$$d_{\mathrm{DSCJ}}(M, D_i) = |M - D_i| + |D_i - M| + 2\delta(M, D_i) - 2t(D_i) + 2\sum_{g \in \Gamma_M} \alpha_{g,MD_i}$$

where Γ_A and Γ_M are the set of genes in the gene orders of A and M, respectively, and so also the genes alphabets for M and the D_is. Variables α_{g,AM_a} and α_{g,MD_i} are defined as $\alpha_{g,uv}$ above.

For any two adjacency sets X and Y, we use the identity $|X-Y|+|Y-X| = |X| + |Y| - 2|X \cap Y|$ to obtain

$$d_{\mathrm{DSCJ}}(A, M_a) = |A| + |M_a| - 2|A \cap M_a| + 2\delta(A, M_a) - 2t(M_a) + 2\sum_{g \in \Gamma_A} \alpha_{g,AM_a},$$

$$d_{\mathrm{DSCJ}}(M, D_i) = |M| + |D_i| - 2|M \cap D_i| + 2\delta(M, D_i) - 2t(D_i) + 2\sum_{g \in \Gamma_M} \alpha_{g,MD_i}.$$

This eliminates the need to count the actual number of cut and join events along every branch. Instead, it suffices to compute the common adjacencies in the parent and child genomes (using the terms $|A \cap M_a|$ and $|M \cap D_i|$) for each branch (A, M_a) and (M, D_i).

For a median M, let $s(M) = d_{\mathrm{DSCJ}}(A, M_a) + \sum_{i=1}^{k} d_{\mathrm{DSCJ}}(M, D_i)$ be the *score* of M. It follows easily from above that

$$\begin{aligned} s(M) = &\left[|A| + 2\delta(A, M_a) + \sum_{i=1}^{k} \left(|D_i| + 2\delta(M, D_i)\right)\right] \\ &- \left[\sum_{i=1}^{k} \left(2|M \cap D_i| + 2t(D_i) - 2\sum_{g \in \Gamma_M} \alpha_{g,MD_i}\right)\right. \\ &\left. + 2|A \cap M_a| + 2t(M_a) - 2\sum_{g \in \Gamma_A} \alpha_{g,AM_a} - (k+1)|M|\right] \end{aligned}$$

Let $N = |A| + 2\delta(A, M_a) + \sum_{i=1}^{k} \big(|D_i| + 2\delta(M, D_i) + 2t(D_i)\big)$. Given that N depends only on A and D_i and not on M, it is constant (note that $\delta(A, M_a)$ and $\delta(M, D_i)$ are constant as the gene content of M is an input to the problem). Thus in order to minimize the score $s(M)$, we only need to maximize the term:

$$\sum_{i=1}^{k} \left(2|M \cap D_i| - 2\sum_{g \in \Gamma_M} \alpha_{g,MD_i}\right) + 2|A \cap M_a| + 2t(M_a) - 2\sum_{g \in \Gamma_A} \alpha_{g,AM_a} - (k+1)|M|$$

which is negated in $s(M)$, as required in Eq. (5). □

Proof of Lemma (3). To prove this lemma, we start with a median containing a non-candidate adjacency. For odd values of k, we prove that removing the non-candidate adjacency results in another median of the same cost whereas for even k, it is shown that the resultant median (on removing the non-candidate adjacency) is better. We temporarily ignore the influence of reduced genomes for this proof.

Consider an adjacency xy that is not a candidate. Recall that since xy is not a candidate it is present in at most $\lfloor \frac{k+1}{2} \rfloor$ genomes from $\{A, D_1, ..., D_k\}$. Assume that M is a median genome and xy is present in M. Further, assume that M is optimal. Thus, the sum of the distances $d_{\text{DSCJ}}(A, M_a) + \sum_{i=1}^{k} d_{\text{DSCJ}}(M, D_i)$ should be the least over all medians. Let M' be the genome obtained by removing xy from M.

Let $D_{xy} \subseteq \{D_1, ..., D_k\}$ be the set of descendant genomes that contain xy, and let $\overline{D_{xy}}$ be the set of those that do not. For any $D_i \in D_{xy}$, the adjacency need not be cut along (M, D_i), however it has to be added along (M', D_i), introducing an extra cost of 1 to the total distance. Thus, $d_{\text{DSCJ}}(M, D_i) = d_{\text{DSCJ}}(M', D_i) - 1$, for all $D_i \in D_{xy}$. On the other hand, if $D_i \notin D_{xy}$, then it does not contain xy. Consequently, for all such D_i, the adjacency has to be cut along (M, D_i) but not along (M', D_i) (since M' does not contain it in the first place). Thus, for all $D_i \notin D_{xy}$, $d_{\text{DSCJ}}(M, D_i) = d_{\text{DSCJ}}(M', D_i) + 1$.

Further if A contains $a(x)a(y)$, it need not be cut along (A, M_a) but may need to be cut along (A, M'_a) thereby introducing a possible extra cost of 1 (note here the possibility that some $x^*y^* \in M$ distinct from xy such that $a(x^*)a(y^*) = a(x)a(y)$). Thus, $d_{\text{DSCJ}}(A, M_a) \geq d_{\text{DSCJ}}(A, M'_a) - 1$. If instead, A does not contain xy then it has to be joined along (A, M_a) and not along (A, M'_a). Unlike the previous case, the cost of the join is unavoidable. Hence, $d_{\text{DSCJ}}(A, M_a) = d_{\text{DSCJ}}(A, M'_a) + 1$.

Case 1: A contains xy. Then $|D_{xy}| \leq \lfloor \frac{k+1}{2} \rfloor - 1$.

$$
\begin{aligned}
&d_{\text{DSCJ}}(A, M_a) \geq d_{\text{DSCJ}}(A, M'_a) - 1 && \\
&d_{\text{DSCJ}}(M, D_i) = d_{\text{DSCJ}}(M', D_i) - 1 && \forall D_i \in D_{xy} \\
&d_{\text{DSCJ}}(M, D_i) = d_{\text{DSCJ}}(M', D_i) + 1 && \forall D_i \notin D_{xy}
\end{aligned}
$$

Summing over all the input genomes, we get

$$
\begin{aligned}
d_{\text{DSCJ}}(A, M_a) + \sum_{D_i \in D_{xy}} d_{\text{DSCJ}}(M, D_i) \geq d_{\text{DSCJ}}(A, M'_a) + \sum_{D_i \in D_{xy}} d_{\text{DSCJ}}(M', D_i) \\
+ |\overline{D_{xy}}| - (|D_{xy}| + 1)
\end{aligned}
$$

We know that $|D_{xy}| + 1 \leq \lfloor \frac{k+1}{2} \rfloor$. If k is even, $|\overline{D_{xy}}| > |D_{xy}| + 1$. Hence,

$$
d_{\text{DSCJ}}(A, M_a) + \sum_{D_i \in D_{xy}} d_{\text{DSCJ}}(M, D_i) > d_{\text{DSCJ}}(A, M'_a) + \sum_{D_i \in D_{xy}} d_{\text{DSCJ}}(M', D_i)
$$

Thus, the cost of M' is better than that of the optimal median M and we have a contradiction. If k is odd, then $|\overline{D_{xy}}| = |D_{xy}| + 1$ and hence both M and M' incur the same overall cost. In other words, the removal of a non-candidate adjacency does not increase the cost of the optimal median. Thus, iteratively removing all such adjacencies will yield an optimal median that consists solely of candidate adjacencies.

Case 2: A does not contain xy. Then $|D_{xy}| \leq \lfloor \frac{k+1}{2} \rfloor$.

$$\begin{aligned} d_{\text{DSCJ}}(A, M_a) &= d_{\text{DSCJ}}(A, M') + 1 \\ d_{\text{DSCJ}}(M, D_i) &= d_{\text{DSCJ}}(M', D_i) - 1 && \forall D_i \in D_{xy} \\ d_{\text{DSCJ}}(M, D_i) &= d_{\text{DSCJ}}(M', D_i) + 1 && \forall D_i \notin D_{xy} \end{aligned}$$

The analysis in this case is similar to Case 1. On adding all the equations and using $|D_{xy}| \leq \lfloor \frac{k+1}{2} \rfloor$, once again we reach a contradiction when k is even. When k is odd, both M and M' yield the same overall distance. Thus, we can still obtain the optimal median by iteratively removing non-candidate adjacencies.

Thus, when k is odd, there exists at least one optimal median consisting only of candidate adjacencies. However, when k is even, the optimal median must consist only of candidate adjacencies. □

References

1. Angibaud, S., Fertin, G., Rusu, I., Thévenin, A., Vialette, S.: On the approximability of comparing genomes with duplicates. J. Graph Algorithms Appl. **13**(1), 19–53 (2009)
2. Berman, P., Karpinski, M., Scott, A.D.: Approximation hardness of short symmetric instances of MAX-3SAT. Technical report TR03-049, Electronic Colloquium on Computational Complexity (ECCC) (2003)
3. Blanchette, M., Bourque, G., Sankoff, D.: Breakpoint phylogenies. Genome Inform. **8**, 25–34 (1997)
4. Blin, G., Chauve, C., Fertin, G., Rizzi, R., Vialette, S.: Comparing genomes with duplications: a computational complexity point of view. IEEE/ACM Trans. Comput. Biol. Bioinform. **4**(4), 523–534 (2007)
5. Boyd, S.C., Haghighi, M.: Mixed and circular multichromosomal genomic median problem. SIAM J. Discret. Math. **27**(1), 63–74 (2013)
6. Bryant, D.: The complexity of calculating exemplar distances. In: Sankoff, D., Nadeau, J.H. (eds.) Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families, pp. 207–211. Springer, Dordrecht (2000). https://doi.org/10.1007/978-94-011-4309-7
7. Bryant, D.: A lower bound for the breakpoint phylogeny problem. J. Discret. Algorithms **2**(2), 229–255 (2004)
8. Davin, A.A., Tricou, T., Tannier, E., de Vienne, D.M., Szollosi, G.J.: Zombi: a simulator of species, genes and genomes that accounts for extinct lineages. bioRxiv (2018). https://doi.org/10.1101/339473
9. Doerr, D., Balaban, M., Feijão, P., Chauve, C.: The gene family-free median of three. Algorithms Mol. Biol. **12**(1), 14:1–14:14 (2017)

10. Feijão, P., Mane, A.C., Chauve, C.: A tractable variant of the single cut or join distance with duplicated genes. In: Meidanis, J., Nakhleh, L. (eds.) RECOMB CG 2017. LNCS, vol. 10562, pp. 14–30. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67979-2_2
11. Feijão, P., Meidanis, J.: SCJ: a breakpoint-like distance that simplifies several rearrangement problems. IEEE/ACM Trans. Comput. Biol. Bioinform. **8**(5), 1318–1329 (2011)
12. Fertin, G., Labarre, A., Rusu, I., Tannier, E., Vialette, S.: Combinatorics of Genome Rearrangements. Computational Molecular Biology. MIT Press, Cambridge (2009)
13. Kondrashov, F.A.: Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc. R. Soc. Lond. B Biol. Sci. **279**(1749), 5048–5057 (2012)
14. Kovác, J.: On the complexity of rearrangement problems under the breakpoint distance. J. Comput. Biol. **21**(1), 1–15 (2014)
15. Levasseur, A., Pontarotti, P.: The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. Biol. Direct **6**(1), 11 (2011)
16. Luhmann, N., Lafond, M., Thèvenin, A., Ouangraoua, A., Wittler, R., Chauve, C.: The SCJ small parsimony problem for weighted gene adjacencies. IEEE/ACM Trans. Comput. Biol. Bioinform. (2017). https://doi.org/10.1109/TCBB.2017.2661761
17. Ming, R., VanBuren, R., Wai, C.M., et al.: The pineapple genome and the evolution of CAM photosynthesis. Nat. Genet. **47**(12), 1435–1442 (2015)
18. Moret, B.M.E., Wyman, S.K., Bader, D.A., Warnow, T.J., Yan, M.: A new implementation and detailed study of breakpoint analysis. In: Pacific Symposium on Biocomputing, pp. 583–594 (2001)
19. Neafsey, D., Waterhouse, R., Abai, M., et al.: Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. Science **347**(6217), 1258522 (2015)
20. Pe'er, I., Shamir, R.: The median problems for breakpoints are np-complete. Technical report TR98-071, Electronic Colloquium on Computational Complexity (ECCC) (1998)
21. Sankoff, D., Sundaram, G., Kececioglu, J.D.: Steiner points in the space of genome rearrangements. Int. J. Found. Comput. Sci. **7**(1), 1–9 (1996)
22. Tannier, E., Zheng, C., Sankoff, D.: Multichromosomal median and halving problems under different genomic distances. BMC Bioinform. **10**, 120 (2009)
23. Zeira, R., Shamir, R.: Sorting by cuts, joins, and whole chromosome duplications. J. Comput. Biol. **24**(2), 127–137 (2017)