The Distance and Median Problems in the Single-Cut-or-Join model with Single-Gene Duplications

Aniket Mane (amane@sfu.ca)
Simon Fraser University - Burnaby https://orcid.org/0000-0002-0130-7149
Manuel Lafond
Universite de Sherbrooke
Pedro Feijao
Simon Fraser University
Cedric Chauve
Simon Fraser University

Research

Keywords: Genomes Rearrangement, Gene Duplication, Genomic Distance, Genome Median

Posted Date: February 13th, 2020

DOI: https://doi.org/10.21203/rs.2.23403/v1

Version of Record: A version of this preprint was published at Algorithms for Molecular Biology on May 4th, 2020. See the published version at https://doi.org/10.1186/s13015-020-00169-y.

RESEARCH

The Distance and Median Problems in the Single-Cut-or-Join model with Single-Gene Duplications

Aniket C Mane[1], Manuel Lafond[2], Pedro C Feijao[3] and Cedric Chauve[1*]

*Correspondence:
cedric.chauve@sfu.ca
[1]Department of Mathematics, Simon Fraser University, 8888 University Drive, V5A 1S6 Burnaby, Canada
Full list of author information is available at the end of the article

Abstract

Background. In the field of genome rearrangement algorithms, models accounting for gene duplication lead often to hard problems. For example, while computing the pairwise distance is tractable in most duplication-free models, is NP-complete for most extensions of these models accounting for duplicated genes. Moreover, problems involving more than two genomes, such as the genome median and the Small Parsimony problem, are intractable for most duplication-free models, with some exceptions, for example the Single-Cut-or-Join (SCJ) model.
Results. We introduce a variant of the SCJ distance that accounts for duplicated genes, in the context of directed evolution from an ancestral genome to a descendant genome where orthology relations between ancestral genes and their descendant are known. Our model includes two duplication mechanisms: single-gene tandem duplication and the creation of single-gene circular chromosomes. We prove that in this model, computing the directed distance and a parsimonious evolutionary scenario in terms of SCJ and single-gene duplication events can be done in linear time. We also show that the directed median problem is tractable for this distance, while the rooted median problem, where we assume that one of the given genomes is ancestral to the median, is NP-complete. We also describe an Integer Linear Program for solving this problem. We evaluate the directed distance and rooted median algorithms on simulated data.
Conclusion. Our results provide a simple genome rearrangement model, extending the SCJ model to account for single-gene duplications, for which we prove a mix of tractability and hardness results. For the NP-complete rooted median problem, we design a simple Integer Linear Program. Our publicly available implementation of these algorithms for the directed distance and median problems allow to solve efficiently these problems on large instances.
Availability. https://github.com/cchauve/SCJ-with-SGD

Keywords: Genomes Rearrangement; Gene Duplication; Genomic Distance; Genome Median

Background

Reconstructing the evolution of genomes at the level of large-scale genome rearrangements is an important problem in computational biology; e.g. [1, 2]. For a given genome rearrangement model, there are several computational problems that can be defined, from the computation of pairwise distances to the reconstruction of complete phylogenetic trees, often following a parsimony approach [3]. Among these problems, the reconstruction of ancestral gene orders given a species phy-

logeny has been considered in various frameworks, including the Small Parsimony Problem (SPP), which aims at computing 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 Genome Median Problem, where the given species phylogeny contains a single ancestral node.

For most genome rearrangement models that do not consider gene duplication, computing the pairwise distance is tractable [3]. This contrasts with the median problem, that has been shown to be intractable in most models. The median problem was introduced in 1996 [4], motivated by its application in heuristics for the SPP [5]. Early results suggested that, even in the simple breakpoint distance model, computing a median gene order is intractable [6], and heuristics based on the Traveling Salesman Problem were introduced to solve the breakpoint median problem [5, 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 fragments is tractable in the breakpoint distance model, by using a reduction to a Maximum Weight Matching (MWM) problem [8]. This tractability result, the first of its kind in genome rearrangement algorithms, renewed the interest in gene order median problems, although most of the following work presented intractability results, even on variations of the breakpoint distance [9, 10, 11]. A notable exception was the Single-Cut-or-Join (SCJ) distance, introduced by Feijão and Meidanis [12], where it was shown that both the median problem and the SPP are tractable.

Gene duplication is another important evolutionary mechanism; genes can be duplicated through different kinds of evolutionary events, from single-gene duplication to whole-genome duplications (WGD) [13, 14]. The first models of evolution by genome rearrangements considered the case of genomes with equal gene content, thus disregarding gene duplication and gene loss. However, for most models that account for gene duplication, the pairwise distance problem is intractable. For example, whereas the distance between two genomes can be computed in linear time for genomes without duplicated genes under the Double-Cut and Join (DCJ) model, it becomes NP-complete to compute the distance when duplicated genes are considered [15, 16], although it can be approximated when the gene content in both genomes is balanced [17]. So far, even in simpler genome rearrangement models, the general problem of computing a distance with duplicated genes is difficult [18, 19], with the exception of polynomial time algorithms for two extensions of the SCJ model that include large-scale duplications: the SCJ double distance [12], where duplicated genes occur through a WGD, and the SCJ and whole chromosome duplication (WCD) problem, motivated by cancer genomics [20].

In the present paper, we introduce novel results about the pairwise distance and median problems, in a model accounting for gene duplications. Our evolutionary model is an extension of the SCJ model that includes single-gene duplications of two different types, Tandem Duplications (TD) or Floating Duplication (FD) in which a new copy is introduced as a circular chromosome. We call this genome rearrangement model the SCJ-TD-FD model. The the pairwise distance problem we consider, is a *directed distance* problem, where we assume that one genome, say A, is duplication-free, while the other one, denoted by D, can contain duplicated genes. This setting is motivated by (1) the SPP where distances are considered

along the branches of a given species phylogeny, so between an ancestral genome A and a descendant genome D and (2) the fact that developments in phylogenomics methods – especially gene trees / species tree reconciliation algorithms – make it realistic to assume that orthology relations between genes of an ancestral gene order and genes of a descendant gene order are known, allowing to see the ancestral gene order as duplication-free with regard to its descendant. This general framework was introduced by Sankoff and El-Mabrouk in [21] (see also [22]) and was later implemented in the DeCo* family of algorithms [23] to reconstruct ancestral gene adjacencies in a duplication-aware evolutionary model from data including extant gene orders and reconciled gene trees. We show that in the SCJ-TD-FD model the directed pairwise distance problem is tractable, and that a parsimonious scenario, can be computed in linear time. We also introduce two genome median problems, the directed median problem and the rooted median problem. In the directed median, we aim to reconstruct a parsimonious duplication-free ancestral gene order from the gene orders of $k \geq 2$ descendant genomes, that minimizes the sum of the directed distances to the k descendants gene orders. In the rooted median problem, we aim to reconstruct a parsimonious median genome between an ancestral genome A and $k \geq 2$ descendant genomes, where we assume that the gene content of the median is given and that unambiguous orthology relations between the median genes and the given gene orders are provided. We prove that the directed median problem is tractable, while the rooted median problem is NP-complete, and we provide a simple simple Integer Linear Program (ILP) for this problem, based on a reduction to a colored MWM problem. We evaluate our algorithms on simulated data and observe that they generate efficiently very accurate results.

Preliminaries

Genes, adjacencies 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*, which we call adjacencies from now. 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*, also called a *telomere*, if it does not belong to an adjacency. 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.

In our work, we consider that genes can be duplicated. This implies that a given gene g can have multiple copies in a genome, the number of copies being called its *copy number*. Given a set of genomes, we call a *gene family* all copies of a given gene observed in the genomes. A set of genomes is said to have *equal gene family content* if every genome contains at least one gene from every gene family. A genome in which every gene has copy number 1 (i.e. is duplication-free) is called a *trivial genome*. A gene family is said to be trivial if each genome contains a single gene from this family.

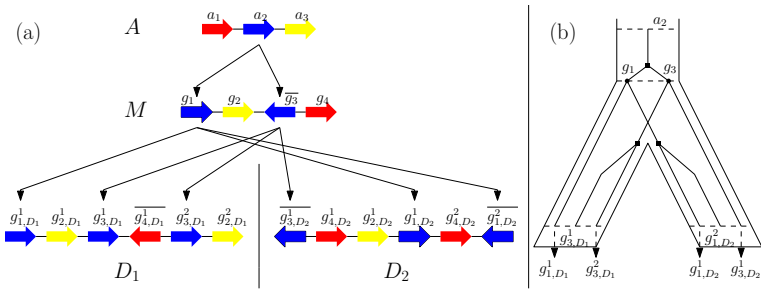

Figure 1 In part (a), each color represents a gene family from A. Notice that each gene in D_1 and D_2 can be traced to a unique gene in M whereas a gene from A might have multiple daughters in M. Part (b) displays the gene tree of the gene family in blue (indicated by arrows in part (a)). Since the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

It is important to note that a non-trivial genome can not always 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 $g^1, \ldots, g^k$. Generally a multi-set of gene adjacencies can have several realizations as a gene order with duplicated genes. Nevertheless, in our work we identify a genome with its multi-set of adjacencies, as we will show this is sufficient in order to solve the directed pairwise distance problem we introduce.

For a given gene order X, we denote by Γ_X its gene content, which is a set if X is trivial and a multi-set otherwise. We call the set induced by Γ_X (which is exactly Γ_X only if X is trivial) its *gene alphabet*; it follows that two gene orders X and Y have equal gene family content if and only if they have the same gene alphabet. A key assumption in our work is that when we compare a pair of gene orders, we assume that one, say A is an ancestor of the second one (say D). In that context, gene families define complete bipartite graphs (bi-cliques) between the two multi-sets of genes of A and D, that define *orthology relations*; we say that these orthology relations are *unambiguous* if all such bi-cliques contain a single gene of A, which is equivalent to state that all members of a gene family in D evolved, by gene duplications, from a unique gene in A. As a consequence, if a genome A is not trivial but is compared to a descendant genome D such that orthology relations between A and D are unambiguous, we say that A is *trivial with respect to* D. This is illustrated in Fig. 1. In a practical context, for a given set of extant gene orders and a species phylogeny for these gene orders, unambiguous orthology relations can be obtained, among other methods, by computing a gene tree per gene family and reconciling the gene trees with the species phylogeny; we refer to [22] for a discussion on the use of reconciled gene trees for the study of gene orders with duplicated genes.

Given two multi-sets X and Y of adjacencies, we define $X - Y$ as 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; so if Y contains more copies of an adjacency than X, $X - Y$ contains no copy of this adjacency. Note that this operator is not symmetric as $X - Y$ can be different from $Y - X$.

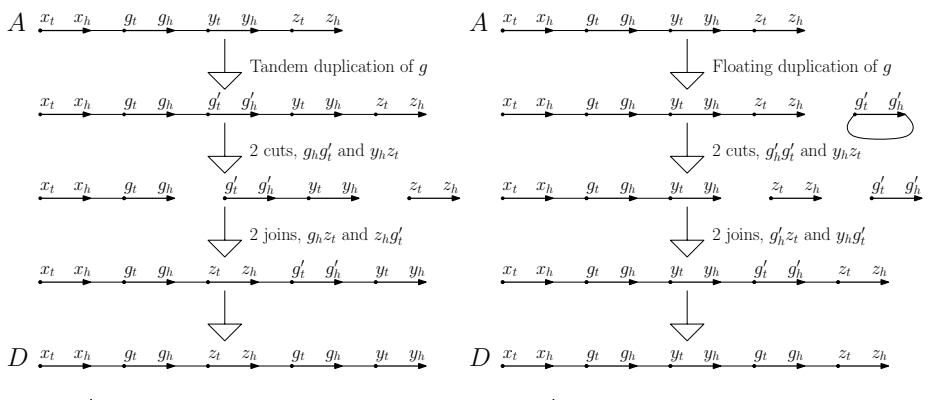

Figure 2 (Left) A tandem duplicate of gene g is introduced. The adjacency $g_h y_t$ has been replaced by $g'_h y_t$ and an adjacency $g_h g'_t$ has been introduced. In this case the total number of operations to obtain D from A is 5. Note that the number of cut and join operations is dependent on the adjacencies of the gene g in A and D. (Right) A floating duplicate of gene g is introduced. An adjacency $g'_h g'_t$ has been added. In this case the total number of operations to obtain D from A is 5.

Evolutionary model: the SCJF-TD-FD model

In the SCJ-TD-FD 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 $g'_h g'_t$. We illustrate the two kinds of duplications in Fig. 2.

The motivation for considering floating duplications is that gene insertions and gene deletions have been modeled with artificial circular chromosomes before, greatly simplifying how to deal with such type of operations. For instance, in DCJ model, a deletion of a gene can be seen as a DCJ operation that applies two cuts to remove the given gene from a chromosome, followed by two joins to "repair" the broken chromosome and to circularize the deleted gene. A gene insertion is the inverse of this operation. This idea was effectively used in the DCJ-indel model by Compeau [24].

Note also that our model does not include gene loss. The extension to include this evolutionary mechanism will be developed in a further work.

Problem statements

The first computational problem we consider is the **directed SCJ-TD-FD (d-SCJ-TD-FD) distance problem**. We consider a model of *directed evolution* in which, when comparing two genomes, we assume one, denoted by A, is an ancestor of the other genome, denoted by D; as the SCJ-TD-FD model does not consider gene losses, both A and D have equal gene family content. Moreover, we require that the genome A is trivial with respect to D. The d-SCJ-TD-FD distance problem

asks to compute the minimum number of SCJ, TD and FD operations needed to transform the ancestral gene order A into the descendant gene order D, denoted by $d_{\text{DSCJ}}(A, D)$. Note that formally this way to measure the dissimilarity between genomes is not a distance as it is not symmetric, due to the assumption of A being trivial with respect to D; it is easy to see that symmetry is the only property of a distance which is not satisfied, so we are dealing actually with a *quasimetric*, although we call it a distance for the sake of consistency with the terminology used in standard genome rearrangement problems.

The second problem we consider is a genome median problem, the **directed SCJ-TD-FD (d-SCJ-TD-FD) median problem**. It is defined as follows: given $D_1, \ldots, D_k$ ($k \geq 2$) (possibly) non-trivial genomes having equal gene family content, we want to compute a trivial genome M on the same set of gene families, that minimizes

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

Note that, while generally in genome rearrangement models the median of two problem is trivial, as any gene order along a parsimonious scenario between D_1 and D_2 is a median, it is different in the context of a directed distance, which motivates this problem.

Last, we consider another genome median problem, motivated by the SPP on rooted phylogenies, the **rooted SCJ-TD-FD (r-SCJ-TD-FD) median problem**. It takes as input (1) $k+1 \geq 3$ genomes, A, $D_1, \ldots, D_k$ with equal gene family content, such that A is a trivial genome, ancestor to the D_is, (2) the gene content Γ_M of a median genome M which is a descendant of A and an ancestor of the genomes $D_1, \ldots, D_k$, and (3) unambiguous orthology relations between Γ_A and Γ_M, and between Γ_M and the genes of each D_i (so one set of orthology relations for each D_i). The goal of the rooted median problem is to compute a gene order for M (i.e. a set of adjacencies on Γ_M) minimizing

$$d_{\text{DSCJ}}(A, M) + \sum_{i=1}^{k} d_{\text{DSCJ}}(M, D_i). \tag{2}$$

Note that in the objective function above, when M is compared to A, genes from the same family are indistinguishable, while they are distinguished when M is compared to each D_i as the provided orthology relations between M and each D_i are unambiguous. So Γ_M can be considered as a multi-set in $d_{\text{DSCJ}}(A, M)$ and a set in the terms $d_{\text{DSCJ}}(M, D_i)$.

The pairwise distance problem

In this section, we show that the d-SCJ-TD-FD distance between A and D can be calculated with the symmetric difference between the adjacency (multi)sets of the input genomes, with extra terms accounting for the observed TD and FD in D. We then provide an alternative formula for the distance, that is more amenable to being implemented in a Integer Linear Program (ILP) in the rooted genome median problem. Last, we describe a linear time algorithm to compute a parsimonious SCJ-TD-FD scenario.

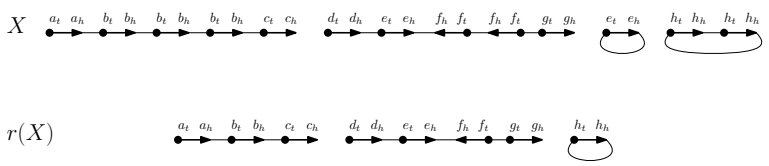

Figure 3 An example of the reduced genome $r(X)$, of the genome X. Note that an instance of $h_h h_t$ is retained so that $r(X)$ contains at least one representative of gene family h. All observed duplications are removed in $r(X)$. Here, $t(X) = |X - r(X)| = 5$.

The directed SCJ-TD-FD distance

Given a gene g, we call a *g-linear array* a sequence of consecutive adjacencies $g_h g_t$; if this sequence forms a circular chromosome, it is called a *g-circular array*. 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-linear array or a g-circular array. 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. 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-circular arrays, as shown in Fig. 3. We call $r(X)$ the *reduced* genome of X.

It is immediate to see that the order in which observed duplications are removed does not matter on the result, i.e. $r(X)$ does not depend on this order. We define $t(X) = |X - r(X)|$, the number of adjacencies to delete to transform X into $r(X)$. Last, for two genomes X and Y, we denote by $\delta(X, Y)$ the absolute difference between the number of genes of X and the number of genes of Y. Our first main result is given in Theorem 1, which generalizes the SCJ distance formula introduced in [12] in the case where A and D are both trivial and with equal gene content.

Theorem 1 *Let A and D be two gene orders with equal gene family content, such that A is trivial with respect to D.*

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

We provide in Additional File 1 a complete proof that we outline here. We first need to define the notion of *context conservation* of a gene between A and D. Assume that a gene g in A is not a telomere (and so there are two adjacencies involving g in A, say $g_t x$ and $g_h y$) and there is a copy of g in D whose extremities form also adjacencies $g_t x$ and $g_h y$. We say that the context of g is *strongly conserved* between A and D. Note that x and y do not need to belong to trivial gene families and there might be several copies of x, y, g in D that conserve the context of g in A. Assume now that the context of g is not strongly conserved between A and D, but both adjacencies involving g, $g_t x$ and $g_h y$, are present in D *using different copies* of g. We say that the context of g is *weakly conserved* between A and D. Again x and y need not to be trivial gene families and there might be several occurrences of adjacencies $g_t x$ and $g_h y$ in D. Last, if the context of g in A is neither strongly nor weakly conserved, and so at most one adjacency involving g in A is also present in D, then we say the context of g is *not conserved*. The principle of the proof is to

proceed by induction on the number of duplicate copies in $r(D)$ (which is equal to $\delta(A, r(D))$) and to pick an arbitrary gene from a non-trivial gene family for which one duplicate is introduced using an FD if the context is strongly conserved, a TD if it is weakly conserved and either an FD or a TD (both can be chosen arbitrarily) if the context is not conserved.

We now introduce an alternative formula to compute the directed distance, easier to implement in an ILP than the formula provided in Theorem 1 as it does not require to consider the reduction $r(D)$ of D. To rewrite $d_{\mathrm{DSCJ}}(A, D)$, we introduce an indicator variable $\alpha_{g,AD}$, where $\alpha_{g,AD} = 1$ if $g_h g_t$ is an adjacency present in both A and D, but all its occurrences in D were removed while reducing D. Formally, $\alpha_{g,AD} = 1$ if $g_h g_t \in A \cap D$ and $g_h g_t \notin r(D)$; otherwise $\alpha_{g,AD} = 0$. We then obtain the following result, whose proof is also provided in Additional File 1.

Corollary 2 *Let A and D be two gene orders with equal gene family content, such that A is trivial with respect to D.*

$$d_{DSCJ}(A, B) = |A - D| + |D - A| + 2\delta(A, D) - 2t(D) + 2 \sum_{g \in \Gamma_A} \alpha_{g,AD} \qquad (4)$$

Computing a parsimonious scenario

It follows from Theorem 1 that computing the d-SCJ-TD-FD distance can be done in linear time in the size of the considered genomes A and D. However, unlike in the case where both A and D are trivial, this does not imply in a straightforward way an algorithm to transform A into D, due to the fact that an adjacency multi-set can have several realizations. Nevertheless, we present a simple algorithm that computes a parsimonious scenario in terms of duplications, cuts and joins from A to D, Algorithm 1 below.

Algorithm 1 Compute an SCJ-TD-FD parsimonious scenario from gene order A to gene order D

```
Reduce D into r(D)
Let A' = A, D' = D and i = 1
while D' is non trivial do
    Let g be an arbitrary gene from A' having more than one descendant gene in D'.
    Relabel g by g^i.
    if the context of g is strongly conserved then
        Relabel the corresponding copy of g in D' by g^i.
        Add to A' a single-gene circular chromosome g.
    else if the context of g is weakly conserved then
        Create an extra copy of g^i with a TD.
        Relabel a copy of g involved in adjacency g_t x in D' by g^i.
    else if one adjacency of g is conserved in D' then
        Relabel the corresponding copy of g in D' by g^i.
        Add to A' a single-gene circular chromosome g^i.
    else
        Relabel an arbitrary copy of g in D' by g^i
        Add to A' a single-gene circular chromosome g^i.
    i = i + 1
Compute an SCJ scenario from A' to D'.
Recreate in D', the linear and circular arrays removed when reducing D into r(D).
```

Theorem 3 *Let A and D be two gene orders with equal gene family content, such that A is trivial with respect to D and D has n_D genes. Algorithm 1 computes a parsimonious SCJ-TD-FD scenario that transforms A into D and can be implemented to run in time and space $O(n_D)$.*

The correctness of the algorithm follows immediately from the fact that it implements exactly the rules described to compute the SCJ-TD-FD distance (proof of Theorem 1). The linear time and space complexity follows from the fact that these rules are purely local and require only to check for the conservation of adjacencies in both considered genomes. Every iteration of the while loop in Algorithm 1 takes place only if there is a non-trivial gene family left in D'. The maximum number of iterations is the number of duplicates genes, $n_d = n_D - n_A$ which is $O(n_D)$ when $n_D \geq n_A$. In each iteration, we check if the context of the chosen gene g is strongly conserved, weakly conserved or not conserved. This involves trying to match the adjacencies of g in A with those in the adjacency set of D' that involve a copy of g. This can be done in constant time, with a linear time preprocessing of the data. Hence, the worst-case time complexity is $O(n_D)$.

The directed and rooted median problems

The Directed Median Problem

Let us remind that the *directed median problem* asks, given k non-trivial genomes $D_1, \ldots, D_k$, $k \geq 2$, with equal gene family content, to find a trivial genome M, such that $\sum_{i=1}^{k} d_{\mathrm{DSCJ}}(M, D_i)$ is minimized. We denote by Γ_M the gene content of M, that is the set induced by the multi-sets Γ_{D_i}. We denote by n the total number of adjacencies in the D_is, $n = \sum_{i=1}^{k} |D_i|$ and by m the total number of gene families, i.e. $m = |\Gamma_M|$.

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

$$s(M) = \sum_{i=1}^{k} d_{\mathrm{DSCJ}}(M, D_i) = \sum_{i=1}^{k} (|M - D_i| + |D_i - M| + 2\delta(M, D_i))$$

Using the identity $|M - D| + |D - M| = |M| + |D| - 2|M \cap D|$, that holds even if the D is a multi-set, and denoting $N_d = \sum_{i=1}^{k} (2\delta(M, D_i) + |D_i|)$, we derive

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

Therefore, minimizing $s(M)$ is equivalent to maximizing $2\sum_{i=1}^{k} |M \cap D_i| - k|M|$.

For a given adjacency a, let $\gamma_i(a)$ be 1 if $a \in D_i$, and 0 otherwise. The score of a genome with a single adjacency a is $s(\{a\}) = N_d - \left(2\sum_{i=1}^{k} \gamma_i(a) - k\right)$. This motivates the following approach, similar to the breakpoint median algorithm of [8]. We build a graph G where the vertices are the extremities (head and tail) of the genes of Γ_M, and weighted edges are defined as follows: for any edge $e = (x, y)$ the weight of e is $w(e) = 2\sum_{i=1}^{k} \gamma_i(e) - k$. So any edge that does not appear in at least half of the descendant genomes has a negative weight. Any matching on G defines a

trivial genome, having the adjacencies corresponding to the edges in the matching; so from now we identify matchings in G and trivial genomes on Γ_M. We denote the weight of M as $w(M) = \sum_{e \in M} w(e)$. It follows that

$$s(M) = N_d - \sum_{e \in M} \left(2 \sum_{i=1}^{k} \gamma_i(e) - k \right) \quad = N_d - w(M).$$

Therefore, solving the MWM problem on G solves the d-SCJ-TD-FD median problem.

To handle the case where some D_i are not reduced, we rely on the fact that the genomes can be reduced first without impacting the optimality of a trivial genome obtained by a MWM. Combined with the tractability of computing a MWM [25], and the fact that any edge that does not correspond to an adjacency observed in a D_i has negative weight and thus does not contribute to a MWM, we obtain the following result.

Theorem 4 *Let $k \geq 2$ and $D_1, \ldots, D_k$ be k genomes having equal gene family content. Let m be the number of gene families and n the number of adjacencies the D_is. The directed SCJ-TD-FD median problem with input $D_1, \ldots, D_k$ can be solved in time and space $O(mn \log_2(m))$.*

The Rooted Median Problem is intractable

We now describe two results for the rooted SCJ-TD-FD median problem. We remind that this problem asks, given $k+1$ non-trivial genomes $A, D_1, \ldots, D_k$, $k \geq 2$, with equal gene family content, the gene content Γ_M of a median genome and unambiguous orthology relations between Γ_A and Γ_M and Γ_M and the Γ_{D_i}s, to find a gene order M on Γ_M, such that $d_{\text{DSCJ}}(A, M) + \sum_{i=1}^{k} d_{\text{DSCJ}}(M, D_i)$ is minimized. As M might be non-trivial with respect to A but is trivial with respect to the D_is, we denote its adjacencies by M_a in the former case (M_a might be a multi-set of adjacencies) and M in the later (a set of adjacencies); so M is induced by distinguishing the copies of a same gene family in M_a. For a given adjacency xy on Γ_M, we denote by $a(x)a(y)$ the adjacency on Γ_A obtained by replacing x and y by their respective orthologs in A, denoted by $a(x)$ and $a(y)$.

Our first result is that, unlike the directed SCJ-TD-FD median problem, the rooted SCJ-TD-FD median problem is NP-complete. Our second result is a simple ILP to solve the rooted SCJ-TD-FD median problem.

Theorem 5 *The rooted SCJ-TD-FD median problem is NP-complete.*

The full proof of this result is given in Additional File 1. We provide here an outline of the proof, together with some comments on the specific instances for which the rooted median problem is shown to be intractable. We show that finding the optimal gene order for M is NP-complete even for $k = 2$, by reduction from the 2P2N-3SAT problem [26] (This problem is sometimes called the (3,B2)-SAT problem, where B2 indicates that the literals are *balanced* with two occurrences each). 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 variable x_i 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, $m = 4n/3$. Then we show that from a given instance of 2P2N-3SAT, we can design a polynomial size instance of the r-SCJ-TD-FD median such that the initial 2P2N-3SAT instance is satisfiable if and only if there exists a median genome M satisfying

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

We can make two interesting observations about our hardness proof:

- In our reduction from 2P2N-3SAT, 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 $r(D)$.
- The reduction we provide involves $k = 2$ and two descendant genomes $D_1,\ D_2$ such that $D_1 = D_2$. It is somewhat striking to remark that computing the distance between A and D is tractable, while computing the distance between A and two identical copies of D, constrained by the gene content and orthology relations of an intermediate genome is hard. However our hardness proof does not imply that computing a median between A and D_1 (with given gene content and unambiguous orhtology relations with A and D_1), and we even conjecture it is tractable.

An Integer Linear Program for the Rooted Median Problem

We now describe a simple Integer Linear Program (ILP) to solve the rooted median problem. The key idea is again 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_M. 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 A, if any, and the number of colors of the selected edges also contributes to the weight of the matching defining the median M.

We first use Eq. (4) to reformulate the objective function of the rooted median problem. The claim below is formally proved in Additional File 1.

Claim 1 *Minimizing the function Eq. (2) defining the evolutionary cost of a median M is equivalent to maximizing the following expression:*

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

An interpretation as a colored MWM problem. In order to compute $|M \cap D_i|$, we use again variables $\gamma_i(e)$ as defined for the directed median algorithm, define a graph G over the vertex set Γ_M and weighted edges, with a weight defined as $w(e) = 2\sum_{i=1}^{k} \gamma_i(e) - (k+1)$.

Since M is trivial with respect to every D_i, the weights for edges $e \in M$ in the graph G defined as in the directed median problem 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 $|x_2y_2 \cap A|$ are equal to 1. In other words, we cannot simply sum the adjacencies of M_a which are present in A.

To address this issue, we introduce the notion of a *color family* (see Fig. 4). 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 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 is 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. Note 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 E_i.

Candidate adjacencies. With the aim to reduce the number of edges in the graph to be considered, we show that we only need to consider specific adjacencies, which we call *candidate adjacencies*. An adjacency xy is a *candidate adjacency* for the median M 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 6 below is proved in the Additional File 1 and implies that the number of adjacencies to consider in our ILP is linear in the sum of the sizes of the input genomes.

Lemma 6 *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.*

Let us remark that, as a consequence, the hardness 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 actually easy to solve with the need to rely on an ILP.

An 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 the binary

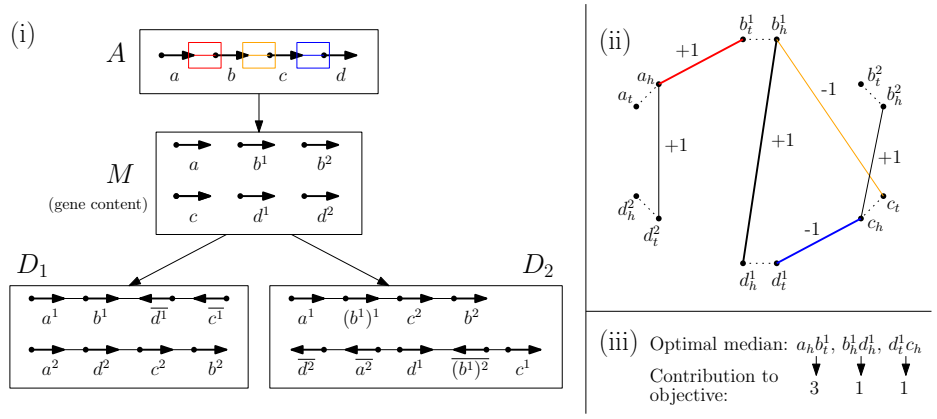

Figure 4 Part (i) of the figure shows an instance of the rooted median problem. Genomes A, D_1 and D_2 and the gene content of M are provided. Each adjacency in A has been assigned a *color family*. For instance, any adjacency $a_h b_t$ in the median belongs to the red family. Part (ii) translates the problem into an instance of the colored MWM problem. Only, candidate adjacencies, appearing in at least two out of the three given genomes can be seen in the graph. The bold edges, forming a matching on the graph, denote the optimal median. Part (iii) lists the contributions of each of the selected edges.

decision variable denoting the inclusion of edge (candidate adjacency) $e \in E(G)$ in M. Let $w(e)$ be the weight of the corresponding edge. Also, let c_i be the binary decision variable indicating the existence of at least one edge from color family E_i in the median 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 $\left\lceil \frac{n_v(g_h g_t)}{n_v(g)} \right\rceil$. Note here that $n_v(g_h g_t)$ counts the occurrences of $g_h g_t$ adjacencies in v while $n_v(g)$ counts the number of copies of g in v. 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 $x, y, z \in \{0, 1\}$.

Thus, there are $|m_M|$ binary variables $x(e)$ where $|m_M|$ is the number of candidate adjacencies for the median. Additionally, there are $|m_A|$ binary variables, to account for the color of each adjacency in A. Further, for every gene g from Γ_A or from Γ_M, there are 7 variables each, used in equations (9-12). All together, there are $2|m_M|$ constraints pertaining to existence of median adjacencies, $|m_A|$ constraints to determine the inclusion of the color of each ancestral adjacency and finally $4(|\Gamma_A| + k|\Gamma_M|)$ constraints from (9-12).

Experiments

We now describe two sets of experiments on simulated data. In the first one, we evaluate the ability of the directed distance to correctly estimate the number of evolutionary events when gene duplications occur through *segmental duplications* (a more realistic model than single-gene duplications). In the second set of experiments, we evaluate the ability of the rooted median to reconstruct an accurate ancestral gene order, again in a model where gene duplication is not restricted to single-gene duplications. We also describe the results of the rooted median ILP on real mosquito genomes data.

The pairwise distance

We ran experiments on simulated instances with the aim to evaluate the ability of the d-SCJ-TD-FD distance to correlate with the true number of syntenic events. We followed a simulation protocol inspired from [15]. The code itself was programmed

in Python and is available via github[1]. We first describe the simulation protocol, followed by the results we obtained.

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

For each instance, we compared two quantities to the true number of cuts and joins in the scenario transforming A into D, which is roughly four times the number of inversions plus three times the number of segmental duplications: first we compared the full SCJ-TD-FD distance, defined as stated in Theorem 1 and the number of cuts and joins ($|A - D| + |D - A|$). Fig. 5 illustrates the results we obtained.

We can make several observations from these results. The first one is a general trend that both measured quantities (the number of cuts and joins and the full SCJ-TD-FD distances) scale linearly with the true number of cuts and joins. The second observation is that, as expected, the slope and y-intercept of the graphs depend from both the frequency of duplications and the length of the duplicated segments. This leaves open the question of using the SCJ-TD-FD distance as an estimator of the number of cuts and joins in an evolutionary model where the probability of duplication compared to rearrangements (that can be estimated for example from reconciled gene trees and adjacency forests [23]) is given and the length of duplicated segments is expected to follow a well defined distribution.

The rooted median problem

Next, 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 [27]. 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

[1]https://github.com/cchauve/SCJ-with-SGD

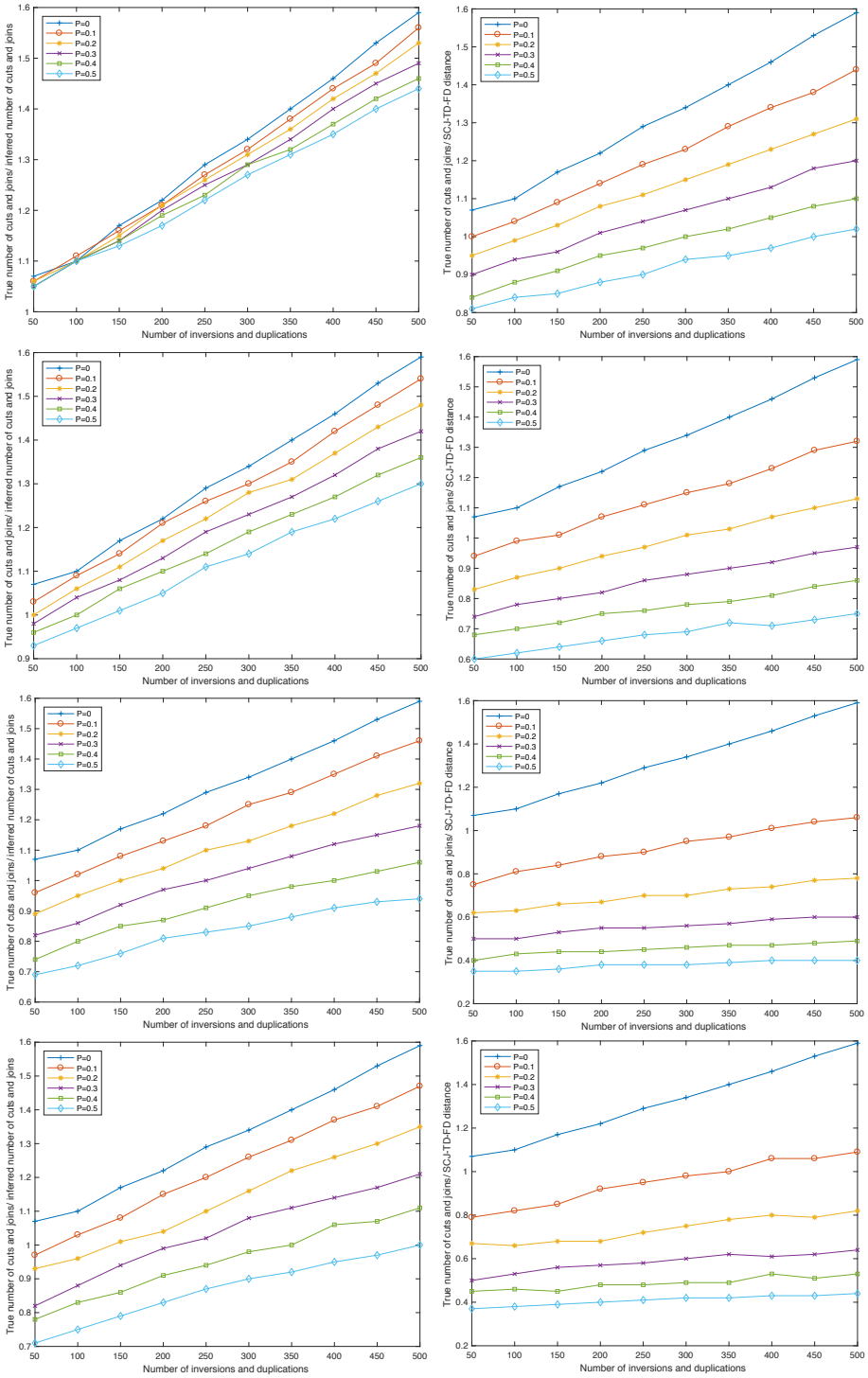

Figure 5 Experimental results, for four duplications parameters – single-gene segmental duplication (top row), two-genes segmental duplication (second row), five-genes segmental duplications (third row), variable length segmental duplications (bottom row) – and two measured quantities – inferred cuts and joins (left column) and SCJ-TD-FD distance (right column).

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.

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

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

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 6, 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.

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.

Conclusions

In this work, our first main result is the introduction of a simple variant of the SCJ model that accounts for single-gene duplications, for which computing the directed distance from a trivial ancestral genome to a non-trivial descendant genome can be done in linear time. This is a somewhat surprising tractability result as some relatively similar problems are known to be intractable, such as the $(1, 2)$-exemplar breakpoint distance [19]. The requirement of considering a trivial ancestral genome and of assuming unambiguous orthology relations is crucial toward our tractability result and is motivated by applications toward the Small Parsimony Problem. Moreover it is relevant toward applications as recent progress in reconciliation algorithms make it realistic to assume that the gene content and orthology relations are known at all nodes of a given species phylogeny; we refer to [22, 23, 28] for a series of papers describing this approach and applying it on real data. From a theoretical point of view, it remains to be seen if these assumptions can be lifted, although this makes the problem very close to general breakpoint distance with duplicated genes, that has been shown to be intractable [29]. Generally, we believe it is worthwhile, both from a theoretical point of view and an applied point of view, to push the tractability boundaries of the SCJ models toward augmented models of evolution (here accounting for duplications).

Our other results deal with the median genome; we show an intriguing tractability boundary between the directed median problem and the rooted median problem, while in the SCJ model with no duplicated genes, both problems are equivalent and the median problem is tractable [12]. 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. A consequence of the hardness of the rooted median problem is that it likely implies the hardness of the Small Parsimony Problem in augmented SCJ model, when the considered species phylogeny is rooted. Again this contrasts with the classical SCJ model for which the Small Parsimony Problem is tractable [12].

To address the intractability of the rooted median problem, 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. This suggests that in practice, the rooted median problem in our model is relatively easy to solve.

Our work leaves several open questions. The most natural one asks if our model can be extended to include other kinds of duplications, other than single-gene duplications. It was shown in [20] that Whole-Chromosome Duplications can be handled, although it is much more complicated to compute the distance. It is then relevant to ask if an intermediate model accounting for a wider range of duplication mechanisms can lead to tractable distance problems. Accounting for gene duplication naturally leads to considering gene loss. So far our results assume all genomes have equal gene family content, which combined with the requirement of unambiguous orthology relations, imply that we do not consider gene losses. It is not difficult to model gene loss in our model, using cuts and joins to extract lost genes into single-gene circular chromosomes, the symmetric operation of a floating duplication. However, in preliminary experiments on real and simulated data (not shown), this leads to a dramatic increase of the distance, driven by gene losses. The question of modeling gene losses with SCJ was previously raised in [20] and is still largely open. Last, the question of counting or sampling optimal evolutionary scenarios, both between two genomes or in the median problem comes to mind. When two genomes are considered, it was shown in [16] that the exact number of SCJ scenarios can be computed in polynomial time through simple recurrences, that also lead to a sampling algorithm; for the median problem, it follows immediately from the algorithm described in [12] that optimal medians can be counted and sampled easily (actually there is a unique optimal median if k is odd). However, both techniques do not extend immediately to our model, especially because an adjacency multi-set does not have a unique realization as a gene order with duplicated genes. So counting and sampling optimal evolutionary scenarios in our model is an open question deserving further research.

Ethics approval and consent to participate
Not applicable

Consent for Publication
Not applicable

Availability of data and materials
Data analysed during the study was generated using the ZOMBI genome simulator

Competing interests
The authors declare that they have no competing interests.

Funding
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. Publications fees are covered by the SFU Open Access Fund.

Author's contributions
CC and PCF designed the research. ACM, ML, PCF and CC designed the algorithms. ML designed the hardness proof. ACM wrote the code and performed the experiments. All authors wrote the manuscript.

Acknowledgements
Most computations were done on the Cedar system of ComputeCanada through a resource allocation to CC.

Author details
[1]Department of Mathematics, Simon Fraser University, 8888 University Drive, V5A 1S6 Burnaby, Canada. [2]Department of Computer Science, Université de Sherbrooke, Boulevard de l'Université, J1K 2R1 Sherbrooke, Canada. [3]School of Computing Science, Simon Fraser University, 8888 University Drive, V5A1S6 Burnaby, Canada.

References

1. Neafsey, D., Waterhouse, R., Abai, M., *et al.*: Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. Science **347**(6217), 1258522 (2015). doi:10.1126/science.1258522
2. Ming, R., VanBuren, R., Wai, C.M., *et al.*: The pineapple genome and the evolution of CAM photosynthesis. Nature Genetics **47**(12), 1435–1442 (2015). doi:10.1038/ng.3435
3. Fertin, G., Labarre, A., Rusu, I., Tannier, E., Vialette, S.: Combinatorics of Genome Rearrangements. Computational molecular biology. MIT Press, ??? (2009)
4. Sankoff, D., Sundaram, G., Kececioglu, J.D.: Steiner points in the space of genome rearrangements. International Journal of Foundations of Computer Science **7**(1), 1–9 (1996). doi:10.1142/S0129054196000026
5. Blanchette, M., Bourque, G., Sankoff, D.: Breakpoint phylogenies. Genome informatics **8**, 25–34 (1997)
6. Pe'er, I., Shamir, R.: The median problems for breakpoints are np-complete. Technical Report TR98-071, Electronic Colloquium on Computational Complexity (ECCC) (1998). http://eccc.hpi-web.de/eccc-reports/1998/TR98-071
7. Bryant, D.: A lower bound for the breakpoint phylogeny problem. Journal of Discrete Algorithms **2**(2), 229–255 (2004). doi:10.1016/S1570-8667(03)00077-7
8. Tannier, E., Zheng, C., Sankoff, D.: Multichromosomal median and halving problems under different genomic distances. BMC Bioinformatics **10**, 120 (2009). doi:10.1186/1471-2105-10-120
9. Boyd, S.C., Haghighi, M.: Mixed and circular multichromosomal genomic median problem. SIAM Journal on Discrete Mathematics **27**(1), 63–74 (2013). doi:10.1137/120866439
10. Kovác, J.: On the complexity of rearrangement problems under the breakpoint distance. Journal of Computational Biology **21**(1), 1–15 (2014). doi:10.1089/cmb.2013.0004
11. Doerr, D., Balaban, M., Feijão, P., Chauve, C.: The gene family-free median of three. Algorithms for Molecular Biology **12**(1), 14 (2017). doi:10.1186/s13015-017-0106-z
12. Feijão, P., Meidanis, J.: SCJ: A breakpoint-like distance that simplifies several rearrangement problems. IEEE/ACM Transactions on Computational Biology and Bioinformatics **8**(5), 1318–1329 (2011). doi:10.1109/TCBB.2011.34
13. Levasseur, A., Pontarotti, P.: The role of duplications in the evolution of genomes highlights the need for evolutionary-based approaches in comparative genomics. Biology Direct **6**(1), 11 (2011). doi:10.1186/1745-6150-6-11
14. Kondrashov, F.A.: Gene duplication as a mechanism of genomic adaptation to a changing environment. Proceedings of the Royal Society of London B: Biological Sciences **279**(1749), 5048–5057 (2012). doi:10.1098/rspb.2012.1108
15. Shao, M., Lin, Y., Moret, B.M.E.: An exact algorithm to compute the Double-Cut-and-Join distance for genomes with duplicate genes. Journal of Computational Biology **22**(5), 425–435 (2015). doi:10.1089/cmb.2014.0096
16. Bulteau, L., Jiang, M.: Inapproximability of (1,2)-exemplar distance. IEEE/ACM Transactions on Computational Biology and Bioinformatics **10**(6), 1384–1390 (2013). doi:10.1109/TCBB.2012.144
17. Rubert, D.P., Feijão, P., Braga, M.D.V., Stoye, J., Martinez, F.H.V.: Approximating the DCJ distance of balanced genomes in linear time. Algorithms for Molecular Biology **12**(1), 3 (2017). doi:10.1186/s13015-017-0095-y
18. 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). doi:10.1007/978-94-011-4309-7_19
19. Angibaud, S., Fertin, G., Rusu, I., Thévenin, A., Vialette, S.: On the approximability of comparing genomes with duplicates. Journal of Graph Algorithms and Applications **13**(1), 19–53 (2009)
20. Zeira, R., Shamir, R.: Sorting by cuts, joins, and whole chromosome duplications. Journal of Computational Biology **24**(2), 127–137 (2017). doi:10.1089/cmb.2016.0045
21. Sankoff, D., El-Mabrouk, N.: Duplication, rearrangement, and reconciliation. In: Sankoff, D., Nadeau, J.H. (eds.) Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families, pp. 537–550. Springer, Dordrecht (2000). doi:10.1007/978-94-011-4309-7_46
22. Chauve, C., El-Mabrouk, N., Guéguen, L., Semeria, M., Tannier, E.: In: Chauve, C., El-Mabrouk, N., Tannier, E. (eds.) Duplication, Rearrangement and Reconciliation: A Follow-Up 13 Years Later, pp. 47–62. Springer, London (2013). doi:10.1007/978-1-4471-5298-9_4
23. Duchemin, W., Anselmetti, Y., Patterson, M., Ponty, Y., Bérard, S., Chauve, C., Scornavacca, C., Daubin, V., Tannier, E.: Decostar: Reconstructing the ancestral organization of genes or genomes using reconciled phylogenies. Genome Biology and Evolution **9**(5), 1312–1319 (2017). doi:10.1093/gbe/evx069
24. Compeau, P.E.C.: DCJ-Indel sorting revisited. Algorithms for Molecular Biology **8**, 6 (2013). doi:10.1186/1748-7188-8-6
25. Galil, Z., Micali, S., Gabow, H.N.: Priority queues with variable priority and an O(EV log V) algorithm for finding a maximal weighted matching in general graphs. In: 23rd Annual Symposium on Foundations of Computer Science, pp. 255–261 (1982). doi:10.1109/SFCS.1982.36
26. 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).

http://eccc.hpi-web.de/eccc-reports/2003/TR03-049/index.html
27. 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). doi:10.1101/339473
28. Anselmetti, Y., Duchemin, W., Tannier, E., Chauve, C., Bérard, S.: Phylogenetic signal from rearrangements in 18 Anopheles species by joint scaffolding extant and ancestral genomes. BMC Genomics **19**(S2), 1–15 (2018). doi:10.1186/s12864-018-4466-7
29. Blin, G., Fertin, G., Chauve, C.: The breakpoint distance for signed sequences. In: 1st Conference on Algorithms and Computational Methods for Biochemical and Evolutionary Networks (CompBioNets'04). Texts in Algorithms, vol. 3, pp. 3–16. King's College London publications, London (2004)

Additional File 1

Proof of Theorem 1. First, we state an immediate result related to the reduction process:

Lemma 7 $d_{DSCJ}(A, D) = d_{DSCJ}(A, r(D)) + t(D)$.

As a consequence, we assume from now on that D has been reduced and does not contain any tandem array or any extra copy of a non-trivial family that is in a single-gene circular chromosome, and we prove that

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

For the sake of exposition, from now we denote $\delta(A, r(D))$ by d.

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

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

We now assume that $d > 0$, and pick a gene g with one copy in A and more than one copy in D. Depending on how the adjacencies of g are conserved or not in D, we have a few different subcases to consider. However, in each subcase the general strategy remains the same, as follows. We build a genome A_2 from A by applying one duplication (FD or TD) and also relabeling the original copy g as g', creating an adjacency $g_h g_t$ in the case of an FD or $g'_h g_t$ in the case of a TD. Then we build a genome D_2 from D by also relabeling one copy of g to g', thus creating a new trivial gene family and an instance of the d-SCJ-TD-FD problem with exactly $d - 1$ duplicated gene copies. We can apply the induction hypothesis, leading to the inequality

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

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

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

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

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

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

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

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

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

In this case, we build A_2 by applying an FD to create an extra copy of g and relabel the original copy of g in A as g'; we also relabel g' an arbitrary copy of g in D that has the same context than g in A, to obtain D_2 (see Fig. 6. Comparing the adjacency sets of A and D with A_2 and D_2, we can see that from A to A_2 two adjacencies where renamed from xg_t and $g_h y$ to xg'_t and $g'_h y$, and exactly the same change happened from D to D_2. Also, the adjacency $g_h g_t$ was added in A_2. As a result, $A_2 = A - \{xg_t, g_h y\} + \{xg'_t, g'_h y, g_h g_t\}$. Similarly, $D_2 = D - \{xg_t, g_h y\} + \{xg'_t, g'_h y\}$. Therefore, we have that $|A - D| + |D - A| = |A_2 - D_2| + |D_2 - A_2| - 1$. Note that this relabeling only works if we introduce a an extra copy of g in A with an FD here; if instead we introduce it with a TD, it would not be possible to get adjacencies xg'_t and $g'_h y$ in D_2, as the copy of g involved in both adjacencies would be different.

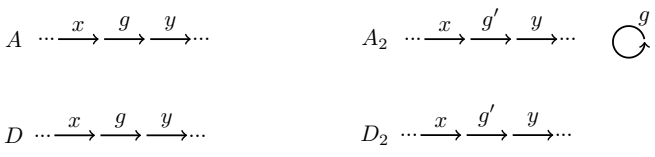

Figure 6 The context of g is strongly conserved between A and D (Case (i)).

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

In this case, we build A_2 by applying a TD on g, relabeling the gene g that has the adjacency xg_t as a new gene g' in both A_2 and D_2, as shown on Fig. 7. Comparing the adjacency sets of A and A_2, we notice that the adjacency xg_t changes to xg'_t, and $g_h g_t$ is added. Thus, $A_2 = A - \{xg_t\} + \{xg'_t, g_h g_t\}$. From D to D_2 we also have the same change, and possibly one more, depending if g'_h is a telomere in D (no change) or if g'_h has an adjacency $g'_h w$. In the former case, $D_2 = D - \{xg_t\} + \{xg'_t\}$. Otherwise, $D_2 = D - \{xg_t, g_h w\} + \{xg'_t, g'_h w\}$. In either case, the possible adjacency $g'_h w$ does not exist in A or A_2. Consequently, the equality $|A - D| + |D - A| = |A_2 - D_2| + |D_2 - A_2| - 1$ holds.

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

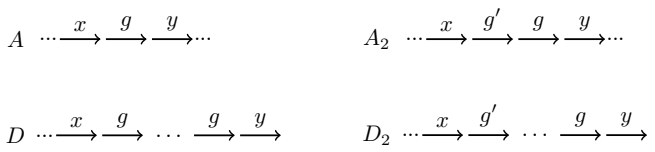

Figure 7 The context of g is weakly conserved between A and D (Case (ii)).

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

This case is similar to case (i), if we assume that either xg_t or $g_h y$, are present in D, or neither. In the same way, we apply an FD on g, labeling the original copy as g', as shown in Fig. 8. On D, we pick a gene g that has an adjacency xg_t or $g_h y$ if any or, if no adjacency involving g is conserved in D, we pick an arbitrary g, and relabel it as g'.

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

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

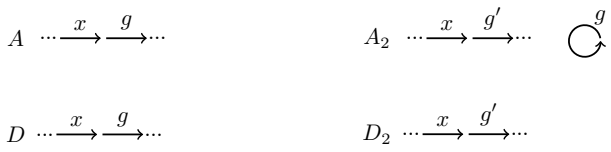

Figure 8 At most one adjacency of g is conserved (Case (iii)).

Proof of Corollary 2. From Theorem 1, we can easily transform the SCJ-TD-FD distance formula into

$$d_{\mathrm{DSCJ}}(A, B) = |A - r(D)| + |r(D) - A| + 2\delta(A, D) - t(D). \tag{13}$$

Indeed we remind that the original pairwise distance formula (eq. (3)) is

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

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 $2\delta(A,D) - t(D) = 2\delta(A, r(D)) + t(D)$.

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 $|A - r(D)|$ and $|r(D) - A|$ by suitable expressions as follows. We show that (1) $|A - r(D)| = |A - D| + \sum_{g \in \Gamma_A} \alpha_g$, and (2) $|r(D) - A| = |D - A| - t(D) + \sum_{g \in \Gamma_A} \alpha_g$. Substituting the terms in eq. (13) yields eq. (4).

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

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

Proof of Theorem 5. We show that finding the optimal gene order for M is NP-hard even for $k = 2$, by reduction from the 2P2N-3SAT problem [26][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 Figure 9 (top left).

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

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

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

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 $\gamma_i^-\alpha_j$.

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 8 *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. □

[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}".

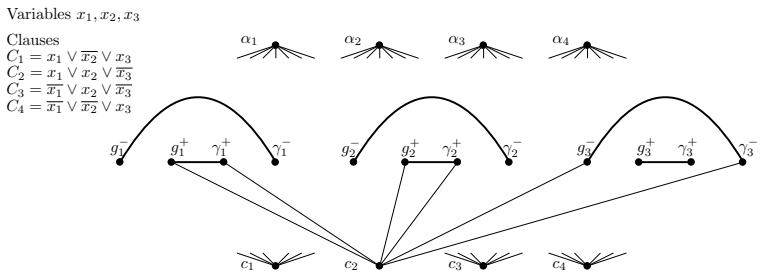

Figure 9 An example of a 2P2N-3SAT instance, with an illustration of the genes of M (only the gene tails are shown) and the adjacencies that are allowed by D_1 and D_2. The fat edges represent pairs of adjacencies of which at least one must be present according to Lemma 8

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. (13)), 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 9 *For $i \in [n]$, M contains at least one of $g_i^+\gamma_i^+$ and $g_i^-\gamma_i^-$.*

Proof 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 8. 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 8, 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 γ_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. □

We now formally prove the hardness of computing the rooted SCJ-TD-FD median.

Theorem 5 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 8, 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 2 *$|M| \leq 3n$, and $|M| = 3n$ only if every c_j and α_j extremity is in some adjacency of M.*

Proof Call an extremity e of a gene in Γ *matchable* if there exists an adjacency of D_1 that contains e. By Lemma 8, 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. □

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

Proof 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 8, 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 2, $|M| \leq 3n$ and because A has n adjacencies, $q \leq n$. Hence, this inequality is only possible if $|M| = 3n$ and $q = n$. □

Because $q = n$, Claim 3 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 3, 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 $c_j e \in M$. By Lemma 8, the adjacency $c_j e$ 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 $c_j g_i^+$ 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. □

Proof of Claim (1). 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

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

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 (6). 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) 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 $|D_{xy}| \leq \left\lfloor \frac{k+1}{2} \right\rfloor$.

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

The analysis in this case is similar to Case 1. On adding all the equations and using $|D_{xy}| \leq \left\lfloor \frac{k+1}{2} \right\rfloor$, 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. □