Mathematical Biology

An algebraic model for inversion and deletion in bacterial genome rearrangement

Chad Clark[1] · Julius Jonušas[2] · James D. Mitchell[2] · Andrew Francis[1]

Abstract

Inversions, also sometimes called reversals, are a major contributor to variation among bacterial genomes, with studies suggesting that those involving small numbers of regions are more likely than larger inversions. Deletions may arise in bacterial genomes through the same biological mechanism as inversions, and hence a model that incorporates both is desirable. However, while inversion distances between genomes have been well studied, there has yet to be a model which accounts for the combination of both deletions and inversions. To account for both of these operations, we introduce an algebraic model that utilises partial permutations. This leads to an algorithm for calculating the minimum distance to the most recent common ancestor of two bacterial genomes evolving by inversions (of adjacent regions) and deletions. The algebraic model makes the existing short inversion models more complete and realistic by including deletions, and also introduces new algebraic tools into evolutionary distance problems.

Keywords Bacterial genomics · Distance · Phylogenetics · Inversion · Deletion · Partial permutation

Mathematics Subject Classification 20M20 · 92D15 · 20M05

✉ Chad Clark
  chad.clark@westernsydney.edu.au

1   Centre for Research in Mathematics and Data Science, Western Sydney University, Penrith, NSW, Australia

2   Mathematical Institute, School of Mathematics and Statistics, University of St Andrews, St Andrews, UK

🗠 Springer

1 Introduction

Methods for computing the evolutionary distance between bacterial genomes are important for phylogenetic reconstruction, especially by way of contrast with organisms that have morphological characteristics and better defined species boundaries. Approaches to distances based on large-scale rearrangements have been widely studied in bacteria because they are often relatively quick to compute and can be used to complement, or even improve, trees based on other methods such as sequence comparisons (Bochkareva et al. 2018).

The bacterial genomes that we will consider have a single circular chromosome. During the evolution of bacterial genomes a frequent rearrangement event is the inversion, where the clockwise order of a contiguous block of conserved regions is reversed (Eisen et al. 2000). If the orientation of regions is taken into account, these events also reverse the orientations of regions in this block. While most early mathematical models assumed the probability of all inversions to be equal, evidence to the contrary has emerged which suggests shorter inversions are more likely (Seoighe et al. 2000; Dalevi et al. 2002; Lefebvre et al. 2003; Darling et al. 2008). With this in mind, throughout this paper we will be concerned with inversions of length two.

Many other large scale changes to bacterial DNA have been observed and investigated, notably insertion of novel DNA (horizontal gene transfer), deletion of segments, translocation of segments to different locations on the genome, and duplication of segments (Saier 2008). Deletions are special in the context of inversions however, because they can occur by the same mechanism, namely site-specific recombination (Plasterk et al. 1983). This means that inversions and deletion are related biologically in a way that other combinations of rearrangement operations are not.

Site-specific recombination acts on the circular genome by forming a synaptic complex around two copies of a specific sequence on the genome, that might be far apart on the sequence but close together in a three-dimensional sense in the cell. The recombinase then cuts the DNA at both sites and rejoins across the two, in effect locally replacing a trivial 2-braid with a braid generator (as an algebraic topologist might describe it). This event can result in the inversion of a segment of the genome relative to the rest of the genome, but can also result in the deletion of a segment, as shown in Fig. 1.

Most rearrangement models, with few exceptions (see Alexandrino et al. (2021a) for instance), assume that the genomes in question have the same sets of regions. While inversion models need both genomes to have the same gene content (or ignore gene content that is not shared), a model incorporating both inversions and deletions can model an evolutionary history of two genomes with differing gene content under the assumption that they both evolved from a common ancestor with the union of their sets of genes. Incorporating deletions thus enables a wider class of genomes to be compared more completely, especially since in some instances (see Raeside et al. (2014)) deletions are the most frequently observed recombination event.

By thinking of bacterial genomes as sequences of region labels or integers (see Bhatia et al. (2018) for a review of these conventions), a pair of genomes σ_1 and σ_2 can be represented by signed or unsigned permutations, assuming all regions are distinct.
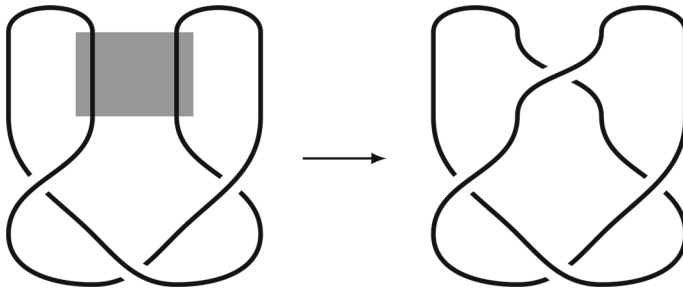
Fig. 1 Site-specific recombination giving rise to deletion, with the area of recombinase action shown shaded on the left. The result is in fact a pair of linked components (topologically, a "Hopf link"), but over time any component without the essential genes from the original genome (such as origin and terminus of replication) would degrade and the result would be a genome without the genetic material from that component (that is, a deletion). If the figure on the left had an even number of twists the result would be an inversion (see for example Francis (2014))

The minimum length sequence of operations $t_1, \cdots, t_k$ such that $\sigma_1 t_1 \cdots t_k = \sigma_2$ consequently provides an estimate of the evolutionary distance between these genomes. These distances may then be used to reconstruct phylogenetic trees using methods such as neighbor-joining (Saitou and Nei 1987).

Although finding the unsigned inversion distance between two genomes is **NP**-hard (Caprara 1997), the signed inversion distance can be found in polynomial time when all inversions (of any length) are assumed to be equally likely (Hannenhalli and Pevzner 1999). For unsigned inversions, an upper bound on the inversion distance between genomes was first provided in (Watterson et al. 1982), with polynomial time algorithms later established from a combinatorial perspective by Jerrum (1985) and an algebraic perspective by (Egri-Nagy et al. 2014). Polynomial time algorithms also exist for signed inversion distances (Galvao et al. 2017; Oliveira et al. 2018) (using terms such as "super short reversal").

When a polynomial time algorithm for a rearrangement distance exists, it is often possible to incorporate both deletions and *insertions* into the model. Polynomial time algorithms exist for calculating the minimal genomic distance under exclusively insertions and deletions (Marron et al. 2004), with insertions, deletions and signed inversions (El-Mabrouk 2000), and with inversions, transpositions, insertions and deletions (Alexandrino et al. 2021b). Insertions and deletions have also been incorporated into other models such as double cut and join (Braga et al. 2010; Shao and Lin 2012).

When insertions and deletions are both allowed, the minimum distance between any pair of genomes G_1 and G_2 with region labels R_1 and R_2 respectively always exists. Furthermore, this distance is symmetric in the sense that the distance from G_1 to G_2 is the same as the distance from G_2 to G_1, because the deletion of a region can be "undone" by inserting the deleted region back into the genome and vice versa. There is, however, little work that considers the addition of deletions without also considering insertions. When considering deletions without insertions, unless we make the assumption that $R_1 \subseteq R_2$ or $R_2 \subseteq R_1$ or both (as in El-Mabrouk (2000)), there will not necessarily be an inversion/deletion sequence that transforms one genome

into the other. To deal with this, we will provide a model for directly reconstructing the most recent common ancestor of G_1 and G_2.

This model will make use of partial analogues of the symmetric group, namely the symmetric inverse monoid and the symmetric inverse category, which will be discussed in Sect. 3. To the best of the authors' knowledge they have not yet been explicitly used in any distance-based methods. When working with these structures, it is advantageous to adopt the convention of writing maps on the right and composing from left to right. That is, we write $(x)f$ instead of $f(x)$, and fg is written instead of $g \circ f$.

Hereafter we will use the term "inversion" to mean an inversion of precisely two adjacent regions. The paper proceeds as follows. In Sect. 2 we provide an algebraic framework for describing bacterial genomes. After introducing a number of key algebraic structures in Sect. 3, these structures are then used to establish an algebraic model of the inversion/deletion process in Sect. 4. This allows us to define a problem called the region alignment problem, where it will be shown that solving this problem over all pairs of orientations of G_1 and G_2 allows for the reconstruction of a parsimonious most recent common ancestor with respect to inversion and deletions. An exact algorithm for calculating this distance is provided in Sect. 5. The paper ends with a Discussion in Sect. 6 that describes some of the important limitations of the models here, and also some of the opportunities for further development. In particular, it is to be hoped that the introduction of the semigroup models here will lead to further work by algebraists to improve applicability and utility of genome rearrangement models.

2 An algebraic model of bacterial genomes

For a circular genome G with a set R of n distinct regions, different rotations and reflections of G represent different ways of viewing the genome in three dimensional space. These symmetries are accounted for by an action of the dihedral group D_n, which consists of permutations in the symmetric group S_n (the group of permutations of $\mathbf{n} = \{1, \ldots, n\}$) representing the rotations and reflections of an n-gon. Beginning with a set X_R containing the $n!$ words of length n whose distinct letters are from R, consider the action $\cdot$ of D_n on X_R where for $\sigma \in D_n$ we have

$$\sigma \cdot x_1 \cdots x_n = x_{(1)\sigma} \cdots x_{(n)\sigma}.$$

The equivalence relation $\sim$ on X_R induced by this action (where words $u, v \in X_R$ are related if and only if there exists $\sigma \in D_n$ such that $u = \sigma \cdot v$) allows for the following algebraic definition of a circular genome.

Definition 2.1 A *genome* G with region set R is an equivalence class in the quotient set $X_R / \sim$.

For $u \in X_R$ the equivalence class of u is denoted by $[u]$, elements of each equivalence class (words in X_R) are called the *reference frames* of G, and for two genomes G_1 and G_2 a *reference pair* is an element of the Cartesian product $G_1 \times G_2$.
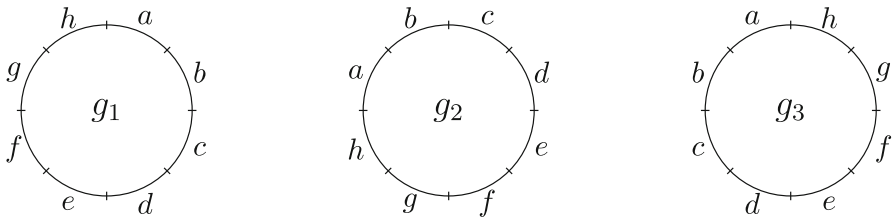
Fig. 2 Given a set of regions $R = \{a, b, c, d, e, f, g, h\}$, the reference frames $g_1 = abcdefgh$, $g_2 = cdefghab$ and $g_3 = hgfedcba$ of the genome $[abcdefgh]$ represent different ways of viewing the same circular genome in three dimensional space. The reference frame g_2 is obtained by rotating g_1 two positions anticlockwise and g_3 is obtained by reflecting g_1 in the vertical axis

To visualise the reference frames of a genome, begin with the unit circle centered at $(0, 0)$ in $\mathbb{R}^2$ and specify a distinguished point at $(0, 1)$. Subdivide the circle into n arcs of equivalent length proceeding clockwise from $(0, 1)$ where the arc immediately clockwise from $(0, 1)$ is considered to be position 1, the next arc clockwise is considered to be position 2 and so on until we reach position n (which will be the arc directly anti-clockwise from $(0, 1)$). If $x_1 \cdots x_n$ is a reference frame of G then its diagram is obtained by labelling position i by $x_i \in R$ via bijection $\lambda : R \to \mathbf{n}$ from regions to positions (see Fig. 2). With this is mind, these bijections may also be used to represent reference frames rather than elements of X_R.

We will proceed under the assumption that each genome has arisen via the minimum possible number of inversions and deletions, which is commonly known as the *parsimony criterion*. This approach allows genome rearrangement problems to be viewed as combinatorial optimisation problems whose minimised solutions represent evolutionary distances in accordance with this criterion (Fertin et al. 2009). With this assumption in mind the most recent common ancestor of genomes G_1 and G_2 with region sets R_1 and R_2 respectively will have region set $R_1 \cup R_2$, noting that it must certainly contain the union of the two sets of regions, but could possibly contain more (in which case a greater number of deletions would be required to yield G_1 and G_2, contradicting the parsimony criterion).

Figure 3 illustrates an example of how reference frames g_1 and g_2 of genomes G_1 and G_2 respectively may arise via inversions and deletions from a (not necessarily most recent) common ancestor A.

3 The symmetric group, the symmetric inverse monoid and their generalisations

To model the inversion/deletion process and formalise the notion of a distance between genomes we use the machinery of the symmetric group, the symmetric inverse monoid and their generalisations. Throughout we let $\mathbf{n} = \{1, \ldots, n\}$ for all positive integers n (where $\mathbf{0} = \varnothing$), let $\mathbb{N} = \{0, 1, \ldots\}$ and $\mathbb{N}^+ = \mathbb{N}\backslash\{0\}$, and let the restriction of a map f to a subset X of its domain be denoted by $f|_X$.
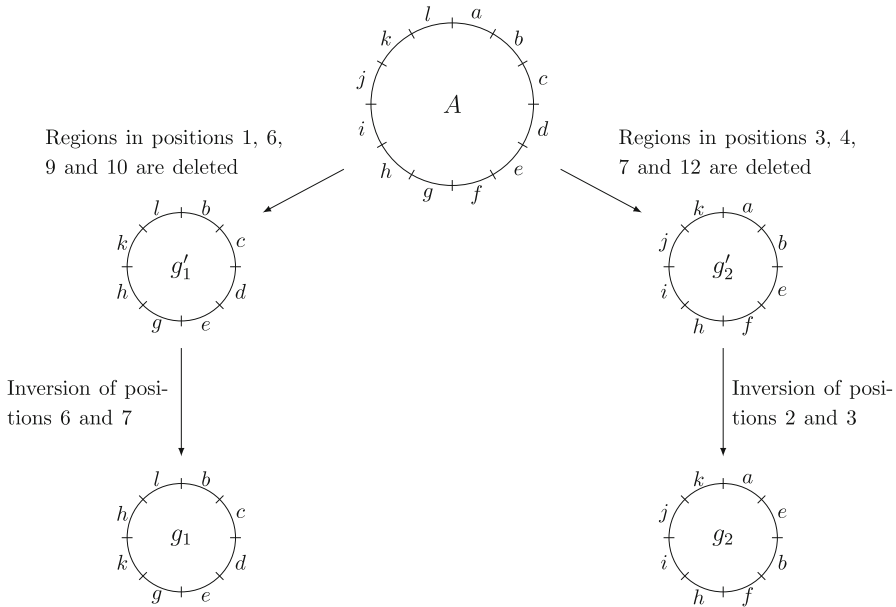
Fig. 3 An example of frames of reference g_1 and g_2 of G_1 and G_2 arising from an ancestor A with the deletions occurring first, followed by inversions. After the deletions but prior to inversions there are intermediate reference frames g'_1 and g'_2 of genomes G'_1 and G'_2

Definition 3.1 Let X, Y and X' be sets where $X' \subseteq X$. A *partial permutation* with domain X' from X to Y is an injection $f|_{X'} : X' \to Y$ where $(x)f = (x)f|_{X'}$ for all $x \in X'$ and where $(x)f$ is undefined for all $x \in X \setminus X'$. The *domain* of f is denoted $\mathrm{dom}(f)$, while $\mathrm{im}(f)$ is the *image* of X' under $f|_{X'}$.

For a monoid M the *inverse* of $m \in M$ is the unique $m^{-1} \in M$ such that $mm^{-1}m = m$ and $m^{-1}mm^{-1} = m^{-1}$. If all elements of M have an inverse in this sense, then M is an *inverse monoid*. The set of partial permutations from $\mathbf{n}$ to itself, which is denoted $\mathcal{I}_n$, is an inverse monoid called the *symmetric inverse monoid* whose identity is the identity map. We will also consider the set $\mathcal{I}_{m,n}$ of partial permutations from the set $\mathbf{m}$ to the set $\mathbf{n}$ for all $m, n \in \mathbb{N}$, where if $m = n$ we write $\mathcal{I}_n = \mathcal{I}_{n,n}$. These partial permutations will be used to represent the relative positions of conserved regions that appear in two circular bacterial genomes, and to represent inversion/deletion operations.

The *symmetric inverse category*, denoted $\mathcal{I}$, is the (small) category whose objects are the natural numbers and where the set of arrows from m to n is $\mathcal{I}_{m,n}$. For partial permutations $f \in \mathcal{I}_{m,n}$ and $g \in \mathcal{I}_{n,p}$, their composition $fg \in \mathcal{I}_{m,p}$ is such that, for all $i \in \mathrm{dom}(f)$, if $(i)f \in \mathrm{im}(f) \cap \mathrm{dom}(g)$ then $(i)fg = \big((i)f\big)g$ and if $(i)f \notin \mathrm{im}(f) \cap \mathrm{dom}(g)$ then $i \notin \mathrm{dom}(fg)$.

The *diagram* of $f \in \mathcal{I}_{m,n}$ is formed by arranging m vertices labelled by elements of $\{1, \ldots, m\}$ above n vertices labelled by elements of $\{1, \ldots, n\}$ forming two parallel rows of vertices. If $(i)f = j$ then there is an edge connecting i in the upper row with j in the lower row of the diagram (as in Fig. 4).

Fig. 4 A partial permutation f in $\mathcal{I}_{5,4}$ with $\mathrm{dom}(f) = \{2, 4, 5\}$ and $\mathrm{im}(f) = \{2, 3, 4\}$
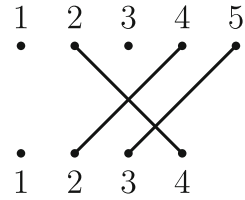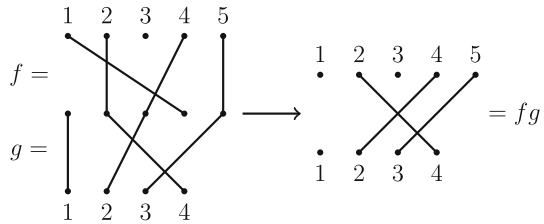

Fig. 5 Calculating the product of partial permutations $f \in \mathcal{I}_{5,5}$ and $g \in \mathcal{I}_{5,4}$

Using the diagrams of f and g it is often helpful to view their composition diagrammatically by first associating the vertices in the lower row of the diagram of f with those in the upper row of the diagram of g, forming a graph called the product graph (see Fig. 5). If there is a path from i in the upper row of the product graph to j in the lower row then $(i)fg = j$.

A partial permutation $f \in \mathcal{I}_{m,n}$ is said to be *order preserving* if, for all $i, j \in \mathrm{dom}(f)$, we have $i < j$ if and only if $(i)f < (j)f$. Instances where $i < j$ but $(i)f > (j)f$ are called *crossings*. The set of order preserving elements of $\mathcal{I}_{m,n}$ is denoted by $\mathcal{POI}_{m,n}$. A partial permutation $f \in \mathcal{I}_{m,n}$ with $\mathrm{dom}(f) = \{x_1, \ldots, x_k\}$ is said to be *orientation preserving* (cf. (McAlister 1998; Catarino and Higgins 1999)) if the sequence $((x_1)f, \ldots, (x_k)f)$ is cyclic, in the sense that there exists at most one index $i \in \mathbf{k}$ such that $(x_i)f > (x_{i+1 \mod k})f$. The set of orientation preserving elements of $\mathcal{I}_{m,n}$ is denoted $\mathcal{POPI}_{m,n}$. Order preserving partial permutations will arise when regions common to two genomes appear in the same order reading from position 1 to position n, while orientation preserving partial permutations will arise when these regions appear in the same (clockwise) cyclic order in both genomes.

4 An algebraic model of inversions and deletions

Given a reference frame of a genome G specified by a bijection $\lambda : R \to \mathbf{n}$, inversions and deletions acting on G are modelled by composing on the right of λ by certain elements of the symmetric inverse category $\mathcal{I}$. For all $n \in \mathbb{N}^+$ let $s_{i;n}$ be the adjacent transposition $(i, i+1)$ in the symmetric group S_n for all $1 \leq i \leq n-1$ and, to account for the circular nature of G, we also consider the 2-cycle $s_{n;n} = (1, n)$ since positions 1 and n are adjacent in G. Letting

$$\mathcal{T}_n = \{s_{i;n} : i \in \mathbf{n}\},$$

Fig. 6 The partial permutation diagram of $d_{2;5} \in \mathcal{I}_{5,4}$. Note that this is still an injective map between the regions that are preserved, with the vertex corresponding to the position of the deleted region having degree 0
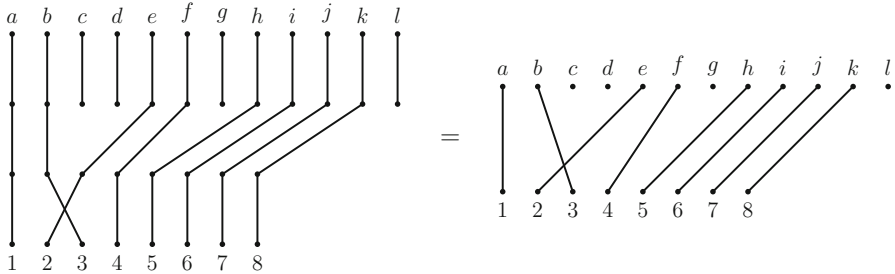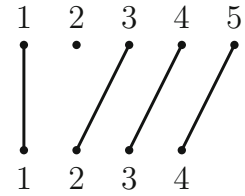

Fig. 7 Given the reference frame of A from Fig. 3, which is represented by the bijection where $a \mapsto 1$, $b \mapsto 2$ and so on, the deletions of the regions at positions 3,4,7 and 12 is given by composing on the right by the (non-unique) term $d_{12;12}d_{7;11}d_{4;10}d_{3;9}$. The subsequent inversion of positions 2 and 3 yielding g_2 is represented by composing on the right by $s_{2;8}$. After these operations a is in position 1, e is in position 2 and so on

composing on the right of λ by elements of $\mathcal{T}_n$ will represent an inversion interchanging two adjacent regions in G. Note that the term "inversion" is used to refer to elements of $\mathcal{T}_n$ as well as the evolutionary operations they represent.

To model deletions, suppose $n \geq 2$ and let $d_{i;n}$ be the unique order preserving map in $\mathcal{POI}_{n,n-1}$ with $\mathrm{dom}(d_{i;n}) = \mathbf{n}\backslash\{i\}$ and $\mathrm{im}(d_{i;n}) = \{1, \ldots, n-1\}$ (see Fig. 6 for an example).

Letting

$$\mathcal{D}_n = \{d_{i;n} : i \in \mathbf{n}\},$$

composing on the right of λ by $d_{i;n} \in \mathcal{D}_n$ will represent deleting the region appearing in position i. Composing by a deletion yields a partial permutation from R to $\mathbf{n}$, where a region x is not in the domain if it has been deleted. Note that after we compose on the right by $d_{i;n}$, for all $j > i$ the region that appeared in position j now appears in position $j - 1$. For all $j < i$ the region appearing in position j remains in that position. Figure 7 illustrates the corresponding compositions of deletions and inversions yielding the reference frame g_2 from the genome A in Fig. 3.

Let G_1 and G_2 be arbitrary genomes and suppose that G_2 can be obtained from G_1 by inversions and deletions (note that we are not considering the most recent common ancestor of G_1 and G_2 here). For a fixed reference pair $(\lambda_{G_1}, \lambda_{G_2}) \in G_1 \times G_2$ a parsimonious inversion/deletion sequence transforming λ_{G_1} into λ_{G_2}, when it exists, corresponds to a minimum length well-defined product u of elements in
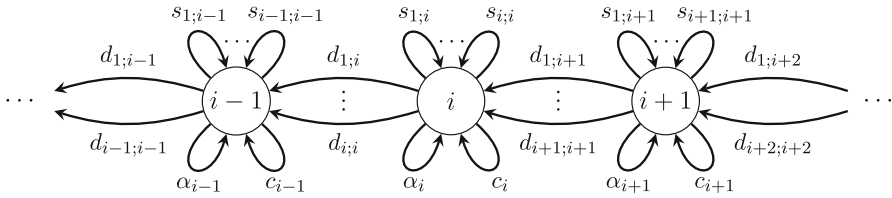
Fig. 8 A local view of the digraph Δ

$$\mathcal{X} = \bigcup_{i \in \mathbb{N}^+} (\mathcal{D}_{i+1} \cup \mathcal{T}_i)$$

such that the bijection $(\lambda_{G_1} u) |_{\mathrm{dom}(\lambda_{G_1} u)}$ is equal to λ_{G_2}. Given a reference frame λ_G of any genome G and a well-defined product u of elements in $\mathcal{X}$ we let

$$(\lambda_G u) |_{\mathrm{dom}(\lambda_G u)} = \overline{\lambda_G u}$$

and let the length of u be denoted by $\ell(u)$. For a fixed reference frame λ_{G_1} of G_1 the quantity

$$d(\lambda_{G_1}, G_2) = \min \left\{ \ell(u) : \overline{\lambda_{G_1} u} \in G_2 \right\}$$

represents the length of a parsimonious inversion/deletion sequence transforming G_1 into G_2 beginning with the reference frame λ_{G_1}, while the quantity

$$d(G_1, G_2) = \min \left\{ d(\lambda_{G_1}, G_2) : \lambda_{G_1} \in G_1 \right\}$$

represents the minimal inversion/deletion distance from G_1 to G_2.

We now work towards establishing Lemma 4.1 from which it follows, for a fixed reference frame λ_{G_1} of G_1, that there exists a reference frame λ_{G_2} of G_2 and a minimum length inversion/deletion sequence transforming λ_{G_1} into λ_{G_2} where the deletions occur first.

We proceed by first defining a digraph Δ whose paths represent the possible sequences of inversions, deletions, rotations and reflections of a genome that can occur. The digraph Δ (see Fig. 8) has

- vertex set $\mathbb{N}$;
- a directed edge from n to n for each element of $\mathcal{T}_n$ representing inversions for all $n \in \mathbb{N}^+$;
- a directed edge from $n + 1$ to n for each element of $\mathcal{D}_n$ representing deletions for all $n \in \mathbb{N}^+$.

The digraph Δ also has a directed edge from n to n for all $n \in \mathbb{N}^+$ labelled by c_n representing the n-cycle rotation $(1, \ldots, n)$ in S_n, along with an edge labelled by α_n representing a reflection where
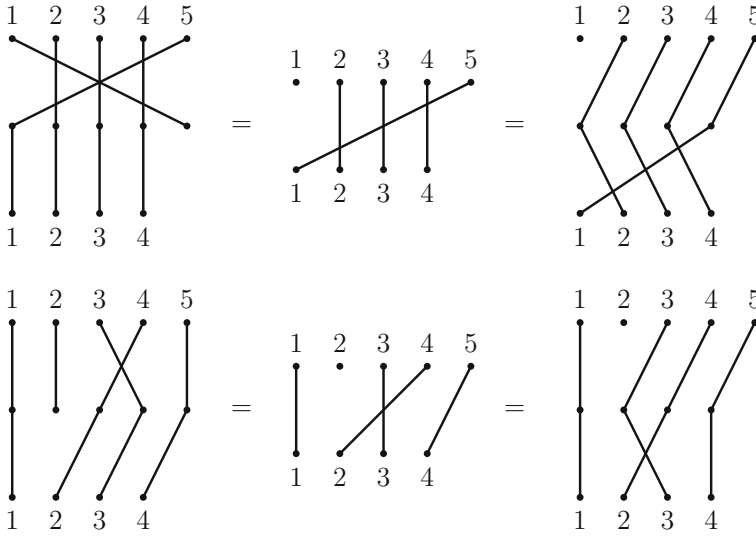
Fig. 9 Diagrammatic illustration of the relation R2 (top row) with $s_{5;5}d_{5;5} = d_{1;5}c_4$ and R4 (bottom row) with $s_{3;5}d_{2;5} = d_{2;5}s_{2;4}$

$$\alpha_n = \begin{cases} (1, n)(2, n - 1) \cdots (k, k + 1) & if\, n = 2k, \\ (1, n)(2, n - 1) \cdots (k, k + 2) & if\, n = 2k+1. \end{cases}$$

Note that the dihedral group D_n is generated by $\{c_n, \alpha_n\}$.

The *free category* Δ^* on Δ contains all words over the alphabet

$$\bigcup_{i \in \mathbb{N}^+} (\mathcal{D}_{i+1} \cup \mathcal{T}_i \cup \{c_i, \alpha_i\})$$

corresponding to paths in Δ (note that edges may be traversed more than once if possible) that represent sequences of inversions, deletions, rotations and reflections. It can be verified (with the aid of diagrams as in Fig. 9 or using the presentation of the symmetric inverse category by East (2020)) that the following relations are satisfied by the corresponding partial permutations in $\mathcal{I}$ for all meaningful values of n, subject to stated constraints:

$$s_{i;n}d_{j;n} = \begin{cases} d_{j;n}s_{n-1;n-1} & \text{if } i = n \text{ and } 1 < j < n & \text{(R1)} \\ d_{1;n}c_{n-1} & \text{if } i = j = n & \text{(R2)} \\ d_{n;n}c_{n-1}^{n-1} & \text{if } i = n \text{ and } j = 1 & \text{(R3)} \\ d_{j;n}s_{i-1;n-1} & \text{if } i > j \text{ and } i \leq n - 1 & \text{(R4)} \\ d_{j;n}s_{i;n-1} & \text{if } i + 1 < j & \text{(R5)} \\ d_{j+1;n} & \text{if } i = j \text{ and } i \leq n - 1 & \text{(R6)} \\ d_{i;n} & \text{if } i + 1 = j \text{ and } i \leq n - 1 & \text{(R7)} \end{cases}$$

$$c_n s_{i;n} = \begin{cases} s_{i-1;n} c_n & \text{if } 2 \leq i \leq n \\ s_{n;n} c_n & \text{if } i = 1 \end{cases} \tag{R8} \tag{R9}$$

$$c_n d_{i;n} = \begin{cases} d_{i-1;n} c_{n-1} & \text{if } 2 \leq i \leq n \\ d_{n;n} & \text{if } i = 1, \end{cases} \tag{R10} \tag{R11}$$

$$\alpha_n d_{i;n} = d_{n-i+1;n} \alpha_{n-1} \quad \text{if } i \in \mathbf{n} \tag{R12}$$

$$\alpha_n s_{i,n} = \begin{cases} s_{n-i;n} \alpha_n & \text{if } 1 \leq i \leq n-1 \\ s_{n;n} \alpha_n & \text{if } i = n. \end{cases} \tag{R13} \tag{R14}$$

Lemma 4.1 *Let G_1 be a circular genome with region set R_1 of size m and let G_2 be a circular genome with region set R_2 of size n where $R_2 \subset R_1$. Given a fixed reference frame $\lambda_1 : R_1 \to \mathbf{m}$ of G_1 suppose that p is a minimum length product corresponding to a path in Δ^* such that $\overline{\lambda_1 p} \in G_2$. There exists a reference frame $\lambda_2 : R_2 \to \mathbf{n}$ of G_2, a product x consisting solely of deletions and a product y consisting solely of inversions (both corresponding to paths in Δ^*) such that $\ell(xy) = \ell(p)$ and $\overline{\lambda_1 xy} = \lambda_2$.*

Proof Let p be a minimum length product in $\mathcal{I}$ corresponding to a word in Δ^* (which, by abuse of notation, we also denote by p) consisting of inversions and deletions such that $\overline{\lambda_1 p} \in G_2$. Suppose also that p contains at least one deletion. Using the relations in (R1)– (R14) it is clear that p is related to a word of the form xyr where x consists solely of deletions, y consists solely of inversions and r consists solely of dihedral symmetries. Since each application of these relations does not increase word length, it follows that $\ell(xy) \leq \ell(xyr) \leq \ell(p)$.

Now, if $\overline{\lambda_1 p}$ is in G_2 then so too is $\overline{\lambda_1 xyr}$ since p and xyr evaluate to the same partial permutation in $\mathcal{I}$. As r consists only of rotations and reflections, it then follows that $\overline{\lambda_1 xyrr^{-1}} = \overline{\lambda_1 xy}$ is also in G_2 as the partial permutation corresponding to r^{-1} is a dihedral group element. The minimality of $\ell(p)$ together with the fact that $\ell(xy) \leq \ell(p)$ implies that $\ell(xy) = \ell(p)$ which completes the proof. $\square$

Theorem 4.2 *For a fixed reference frame λ_{G_1} of G_1 there exists a product u of elements in $\mathcal{X}$ minimising $d(\lambda_{G_1}, G_2)$ where the deletions occur first.*

Proof This follows immediately from Lemma 4.1. $\square$

4.1 Reconstructing the most recent common ancestor

Given genomes G_1 and G_2, candidates for their most recent common ancestor (under the parsimony criterion) are genomes A with region set $R_1 \cup R_2$ minimising the sum $d(A, G_1) + d(A, G_2)$. While it could be the case that there are distinct reference frames λ_{A_1} and λ_{A_2} of A such that $d(A, G_1) = d(\lambda_{A_1}, G_1)$ and $d(A, G_2) = d(\lambda_{A_2}, G_2)$

where $d(A, G_1) + d(A, G_2)$ is minimal, the following theorem establishes the fact that minimum length inversion/deletion sequences yielding G_1 and G_2 can always be thought of as beginning with a fixed reference frame of A.

Theorem 4.3 *Let G_1 and G_2 be genomes with region sets R_1 and R_2 respectively and suppose, among genomes with region set $R_1 \cup R_2$, that the genome A has the property that $d(A, G_1) + d(A, G_2)$ is minimal. There exists a reference frame λ_A of A such that $d(A, G_1) = d(\lambda_A, G_1)$ and $d(A, G_2) = d(\lambda_A, G_2)$.*

Proof Suppose that $\overline{\lambda_{A_1} y} = \lambda_{G_1}$ and $\overline{\lambda_{A_2} z} = \lambda_{G_2}$ where λ_{A_1} and λ_{A_2} are reference frames of A and where y and z are minimum length sequences of inversions and deletions. Suppose also that $w \in \{\alpha_n, c_n\}^*$ (that is, the set of all words whose letters are in $\{\alpha_n, c_n\}$) is such that $\lambda_{A_1} w = \lambda_{A_2}$. Since both α_n and c_n are dihedral group elements there exists α_n^{-1} and c_n^{-1} such that $\alpha_n \alpha_n^{-1}$ and $c_n c_n^{-1}$ are the identity map at $n \in \mathbb{N}$. As such, there exists a word $u \in \{\alpha_n, c_n\}^*$ such that wu corresponds to the identity map at n and so

$$\lambda_{G_1} = \overline{\lambda_{A_1} y} = \overline{\lambda_{A_1} wuy} = \overline{\lambda_{A_2} uy}. \tag{1}$$

Using the relations (R8)– (R14) there exists a word $v \in \{\alpha_a, c_a\}^*$ for some $a \in \mathbb{N}^+$ and a word y' of inversions and deletions such that $uy \sim y'v$ where $\ell(y') \leq \ell(y)$, in which case

$$\overline{\lambda_{A_2} y' v} = \lambda_{G_1} \tag{2}$$

by Equation (1). Consider the fact that $\overline{\lambda_{A_2} y' v v^{-1}} = \overline{\lambda_{A_2} y'}$ by Equation (2). Since multiplying on the right by $v^{-1} \in \{\alpha_a, c_a\}^*$ is equivalent to changing the reference frame of $\overline{\lambda_{A_2} y' v} = \lambda_{G_1}$, there is thus a sequence of inversions and deletions of length $\ell(y')$ such that $\overline{\lambda_{A_2} y'} \in G_1$ which completes the proof since $\ell(y') \leq \ell(y)$ and y is minimal. $\qquad\square$

To find the minimal distance $d(A, G_1) + d(A, G_2)$ given G_1 and G_2 we define a problem which we will refer to as the *region alignment problem*, and show that if the solution to the region alignment problem is $k \in \mathbb{N}$ over all reference pairs in $G_1 \times G_2$ then these genomes have arisen minimally in $d(A, G_1) + d(A, G_2) = k + |R_1 \ominus R_2|$ inversions and deletions (where $\ominus$ denotes the symmetric difference of sets).

To define this problem, begin with G_1 and G_2 where $|R_1| = m$ and $|R_2| = n$. For a reference pair $(g_1, g_2) \in G_1 \times G_2$ where $g_1 = x_1 \cdots x_m$ and $g_2 = y_1 \cdots y_n$ construct a partial permutation $\sigma_{g_1, g_2} \in \mathcal{I}_{m,n}$ where $(i)\sigma_{g_1, g_2} = j$ if and only if $x_i = y_j$. Figure 10 illustrates an example of how σ_{g_1, g_2} is formed.

The elements in $\mathbf{m} \setminus \mathrm{dom}(\sigma_{g_1, g_2})$ and $\mathbf{n} \setminus \mathrm{im}(\sigma_{g_1, g_2})$ represent the regions in the symmetric difference $R_1 \ominus R_2$ that do not appear in both genomes. The crossings of σ_{g_1, g_2} represent the disorder of the labels in $R_1 \cap R_2$, in the sense that if (i, j) is a crossing in σ_{g_1, g_2} then the label $x_j \in R_1 \cap R_2$ appears before $x_i \in R_1 \cap R_2$ in the word g_2 while x_j appears after x_i in g_1. If σ_{g_1, g_2} is order preserving then regions in $R_1 \cap R_2$ appear in the same order reading from position 1 to position n around the genome.
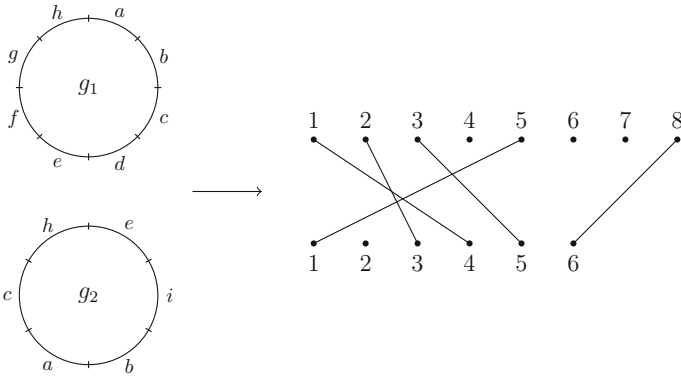
Fig. 10 Forming the partial permutation $\sigma_{g_1,g_2} \in \mathcal{I}_{8,6}$ given two reference frames $g_1 = abcdefgh$ and $g_2 = eibach$ of G_1 and G_2 with $R_1 = \{a, b, c, d, e, f, g, h\}$ and $R_2 = \{a, b, c, e, h, i\}$. The fact that a appears at position 1 in g_1 and at position 4 in g_2 means $(1)\sigma_{g_1,g_2} = 4$

Once σ_{g_1,g_2} has been constructed multiplying on the left of σ_{g_1,g_2} by an element of $\mathcal{T}_m$ represents an inversion acting on the reference frame g_1 while multiplying on the right by an element of $\mathcal{T}_n$ represents an inversion acting on g_2. In the region alignment problem we are given a reference pair $(g_1, g_2) \in G_1 \times G_2$ and ask for the minimum number of inversions acting on either g_1 or g_2 (or both) to place the regions in $R_1 \cap R_2$ in the same (clockwise) cyclic order in both genomes. The region alignment problem is stated mathematically as follows.

Problem 4.4 Let G_1 and G_2 be genomes with region sets R_1 and R_2 respectively where $|R_1| = m$ and $|R_2| = n$. For a reference pair $(g_1, g_2) \in G_1 \times G_2$, find a sequence t_m of elements in $\mathcal{T}_m$ and a sequence t_n of elements in $\mathcal{T}_n$ minimising $\ell(t_m) + \ell(t_n)$ such that $t_m \sigma_{g_1,g_2} t_n \in \mathcal{POPI}_{m,n}$.

This problem is a generalisation of the problem considered by Egri-Nagy et al. (2014) regarding the minimum inversion distance between two genomes with the same region set, which for a permutation $\sigma \in S_n$, asks for the minimum length sequence t_n of elements in $\mathcal{T}_n$ such that σt_n is the identity.

Theorem 4.5 *If G_1 and G_2 are genomes with region sets R_1 and R_2 respectively and $\mu(g_1, g_2)$ is the minimum length solution to Problem 4.4 for a fixed reference pair (g_1, g_2) then, under the parsimony criterion, G_1 and G_2 have descended from their most recent common ancestor in*

$$\ell(g_1, g_2) = |R_1 \ominus R_2| + \min\{\mu(g_1, g_2) : (g_1, g_2) \in G_1 \times G_2\}$$

inversions and deletions.

Proof Let $k = \min\{\mu(g_1, g_2) : (g_1, g_2) \in G_1 \times G_2\}$. We begin by showing that $d(A, G_1) + d(A, G_2)$ is bounded below by $k + |R_1 \ominus R_2|$ over all genomes A with region set $R_1 \cup R_2$. To do this, suppose with the aim of obtaining a contradiction that there exists a reference pair $(g_1, g_2) \in G_1 \times G_2$ and products t_m and t_n of elements

in $\mathcal{T}_m$ and $\mathcal{T}_n$ respectively with $t_m\sigma_{g_1,g_2}t_n \in \mathcal{POPI}_{m,n}$ (i.e. $\ell(t_m) + \ell(t_n) = k$), but where G_1 and G_2 have descended from their most recent common ancestor in strictly less than $k + |R_1 \ominus R_2|$ inversions and deletions.

By Theorem 4.2 there exists a genome A with region set $R_1 \cup R_2$ minimising $d(A, G_1) + d(A, G_2)$ where $|R_1 \ominus R_2|$ deletions occur first. Further, by Theorem 4.3 there exists a fixed reference frame λ_A of A where $d(A, G_1) = d(\lambda_A, G_1)$ and $d(A, G_2) = d(\lambda_A, G_2)$ in a minimal sum $d(A, G_1) + d(A, G_2)$. With these facts in mind and using Fig. 3 as a guide, there exists a parsimonious inversion/deletion sequence yielding G_1 that proceeds by first deleting the regions in $R_2 \backslash R_1$ from a reference frame of A. This gives rise to a reference frame of an intermediate genome G_1'. Likewise for G_2, the regions in $R_1 \setminus R_2$ are deleted first from A to yield a reference frame of an intermediate genome G_2'. Since G_1' and G_2' have been obtained via deletions from the same reference frame of A, for all reference pairs $(g_1', g_2') \in G_1' \times G_2'$ the regions in $R_1 \cap R_2$ appear in the same clockwise cyclic order in both genomes. Thus, the partial permutation $\sigma_{g_1',g_2'}$ is orientation preserving (that is, $\sigma_{g_1',g_2'} \in \mathcal{POPI}_{m,n}$). Equivalently, there exists $(g_1', g_2') \in G_1' \times G_2'$ (possibly after rotating one of the genomes) such that $\sigma_{g_1',g_2'}$ is order preserving (that is, $\sigma_{g_1',g_2'} \in \mathcal{POI}_{m,n}$).

If G_1 and G_2 subsequently arise by sequences of inversions p and q in $\mathcal{T}_m$ and $\mathcal{T}_n$ acting on g_1' and g_2' respectively, then there exists $(g_1, g_2) \in G_1 \times G_2$ such that $p\sigma_{g_1',g_2'}q = \sigma_{g_1,g_2}$. However, it would then follow that $p^{-1}\sigma_{g_1,g_2}q^{-1}$ is orientation preserving where $\ell(p^{-1})+\ell(q^{-1}) = \ell(p)+\ell(q)$. Since we have $\sigma_{g_1',g_2'} = p^{-1}\sigma_{g_1,g_2}q^{-1}$, the assumption that $k = \min\{\mu(g_1, g_2) : (g_1, g_2) \in G_1 \times G_2\}$ is contradicted if $\ell(p) + \ell(q) < \ell(t_m) + \ell(t_n)$. As such, $d(A, G_1) + d(A, G_2)$ is bounded below by $k + |R_1 \ominus R_2|$ over all genomes A with region set $R_1 \cup R_2$.

To complete the proof, we show that if $k = \min\{\mu(g_1, g_2) : (g_1, g_2) \in G_1 \times G_2\}$ then there exists a genome A with region set $R_1 \cup R_2$ such that $d(A, G_1)+d(A, G_2) = k+|R_1 \ominus R_2|$. Beginning with a reference pair $(g_1, g_2) \in G_1 \times G_2$ with $\mu(g_1, g_2) = k$, suppose there exists sequences t_m and t_n of inversions from $\mathcal{T}_m$ and $\mathcal{T}_n$ respectively such that $t_m\sigma_{g_1,g_2}t_n \in \mathcal{POPI}_{m,n}$ (where $\ell(t_m) + \ell(t_n) = k$). Further, suppose that reference frames g_1' and g_2' of G_1' and G_2' are the result of these sequences of inversions acting on g_1 and g_2 respectively. By the circularity of the genomes (performing a rotation if necessary), it may be assumed without loss of generality that the regions in $R_1 \cap R_2 = \{r_1, \ldots, r_h\}$ appear in the same order reading from 1 to n in both g_1' and g_2'.

Let U_i be the set of regions appearing between r_i and r_{i+1} in g_2' reading from 1 to n for all $1 \leq i \leq h - 1$, let U_h be the set of regions appearing after r_h (up to and including position n) in g_2' and let U_0 be the set of regions appearing before r_1 (from position 1 onward) in g_2'. Beginning with g_1', form a genome A with region set $R_1 \cup R_2$ by first inserting the regions in U_0 before r_1 in g_1' where the minimum element of U_i is at position 1 in A. If U_0 is empty, then r_1 is in position 1 in A. Next, for all $1 \leq i \leq h - 1$ insert regions from U_i into g_1' between r_i and r_{i+1} in any way that ensures the regions in U_i appear in the same order that they do in g_2' reading from position 1 to n. Finally, insert regions U_h after r_h in any way that ensures their appropriate order reading from 1 to n where the maximal element of U_h is position n in the resulting genome A. If U_h is empty, then r_h appears in position n in A. Figure 11 illustrates an example of these insertions.
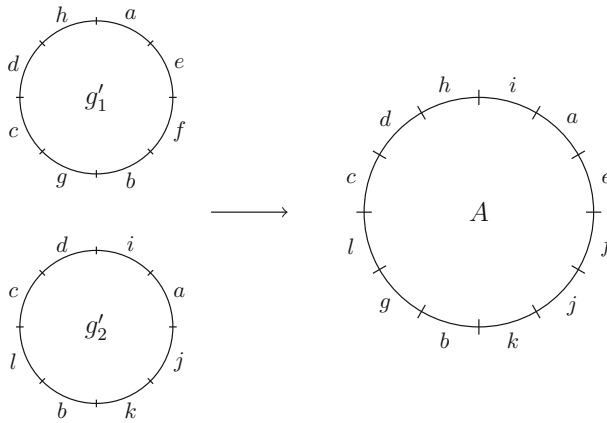
Fig. 11 Given g_1' and g_2', we form the genome A by inserting regions from $R_2 \setminus R_1$ into g_1' in their appropriate positions (with respect to elements of $R_1 \cap R_2$) and appropriate order from 1 to n

Given the construction of A, it is easily verified that deleting the regions in $R_2 \setminus R_1$ from A yields g_1' and that deleting the regions in $R_1 \setminus R_2$ from A yields g_2'. The inverses of the sequences t_m and t_n of inversions then act on g_1' and g_2' respectively to yield g_1 and g_2 in a total of $k + |R_1 \ominus R_2|$ inversions and deletions, as required. □

5 Exact algorithm and complexity

We continue to assume that G_1 and G_2 are genomes with region sets R_1 and R_2, respectively with $|R_1| = m$ and $|R_2| = n$. In this section we provide an exact algorithm for computing sequences t_m in $\mathcal{T}_m$ and t_n in $\mathcal{T}_n$ from Problem 4.4 such that $\ell(t_m) + \ell(t_n)$ is minimised and $t_m \sigma_{g_1, g_2} t_n \in \mathcal{POPI}_{m,n}$. As previously, given σ_{g_1, g_2} this minimised value is denoted by $\mu(\sigma_{g_1, g_2})$. Additionally, we describe the asymptotic time and space complexity of the algorithm, and the limits of its practical applicability on currently available computer hardware.

We denote the identity partial permutation on the set X by id_X. For the purposes of this section, a *graph* Γ is a triple (V, E, X) where V is a set whose elements are called the *vertices* of Γ; X is the set of *edge labels* of Γ; and $E \subseteq V \times X \times V$ is the set of *edges* of Γ.

If S is a semigroup and X is a subset of S, then we define the *left Cayley graph* of S with respect to X to be the graph with nodes S and edges $(s, x, xs) \in S \times X \times S$ for all $s \in S$ and for all $x \in X$; we denote this by $\Gamma_L(S, X)$. The *right Cayley graph* is defined dually, and is denoted $\Gamma_R(S, X)$. If $\Gamma_L(S_1, X_1)$ and $\Gamma_R(S_2, X_2)$ are the left and right Cayley graphs (respectively) of semigroups S_1 and S_2 with respect to subsets X_1 and X_2 then, given $S_1 \subseteq S_2$, we define the *union* of these graphs to be the graph with nodes S_2 and all of the edges belonging to $\Gamma_L(S_1, X_1)$ and $\Gamma_R(S_2, X_2)$. If $\Gamma = (V, E, X)$ is any graph and A is a subset of the vertices V of Γ, then the *subgraph* induced by A is the graph $(A, E \cap (A \times X \times A), X)$.

Let $\sigma_{g_1,g_2} \in \mathcal{I}_{m,n}$ and suppose without loss of generality that $m \leq n$. We consider $\mathcal{I}_{m,m}$ and $\mathcal{POPI}_{m,n}$ to be embedded in $\mathcal{I}_{n,n}$ via an embedding f where $(i)\sigma = j$ for σ in $\mathcal{I}_{m,m}$ or $\mathcal{POPI}_{m,n}$ if and only if $(\sigma)f$ in $\mathcal{I}_{n,n}$ maps i to j. The algorithm for determining $\mu(\sigma_{g_1,g_2})$ has the following steps:

(i) suppose that $\sigma_{g_1,g_2} \in \mathcal{I}_{m,n}$ where $|\operatorname{dom}(\sigma_{g_1,g_2})| = r$ and $m \leq n$;
(ii) let X_j denote the generating set for $\mathcal{I}_{j,j}$ consisting of $\mathcal{T}_j$ and $\operatorname{id}_{\{1,\ldots,j-1\}}$;
(iii) compute the left $\Gamma_L(\mathcal{I}_{m,m}, X_m)$ and right $\Gamma_R(\mathcal{I}_{n,n}, X_n)$ Cayley graphs of $\mathcal{I}_{m,m}$ and $\mathcal{I}_{n,n}$ with respect to the sets X_m and X_n respectively;
(iv) compute the union $\Gamma_{m,n}$ of $\Gamma_L(\mathcal{I}_{m,m}, X_m)$ and $\Gamma_R(\mathcal{I}_{n,n}, X_n)$;
(v) compute the set $\mathscr{D}_r = \{\alpha \in \mathcal{I}_{n,n} : |\operatorname{dom}(\alpha)| = r\}$ in $\Gamma_{m,n}$.

Given that the relation $\mathscr{D}$ on $\mathcal{I}_{n,n}$ where $\alpha \mathscr{D} \beta$ if and only if $|\operatorname{dom}(\alpha)| = |\operatorname{dom}(\beta)|$ is an equivalence relation (called Green's $\mathscr{D}$-relation), the subgraph $\Delta_{n,r}$ induced by $\mathscr{D}_r$ is strongly connected (in the sense that there is a path in both directions between all pairs of vertices). Paths in this strongly connected component will traverse edges from $\Gamma_L(\mathcal{I}_{m,m}, X_m)$ representing inversions from $\mathcal{T}_m$ acting on the genome G_1 with m regions, and edges from $\Gamma_R(\mathcal{I}_{n,n}, X_n)$ representing inversions from $\mathcal{T}_n$ acting on the genome G_2 with n regions.

(vi) compute the subgraph $\Delta_{n,r}$ of $\Gamma_{m,n}$ induced by $\mathscr{D}_r$;
(vii) $\mu(\sigma_{g_1,g_2})$ is then the minimum distance in $\Delta_{n,r}$ between σ_{g_1,g_2} and any element of $\mathcal{POPI}_{m,n}$ in $\mathscr{D}_r$.

Note that steps (i) to (vi) need only be computed once for each m,n and r, and the resulting value of $\Delta_{n,r}$ can be memoised.

Steps (i) and (ii) have combined time complexity $\mathcal{O}(n)$; step (iii) has time and space complexity

$$\mathcal{O}(|X_n||\mathcal{I}_{n,n}|) = \mathcal{O}\left(n \sum_{r=0}^{n} \binom{n}{r}^2 r!\right)$$

(using the Froidure-Pin Algorithm described by Froidure and Pin (1997) for example). Hence the time and space complexity for this step is at best $\mathcal{O}(n!)$. Steps (iv) and (v) also have time complexity $\mathcal{O}(|X_n||\mathcal{I}_{n,n}|)$ since the number of vertices and edges in $\Gamma_L(\mathcal{I}_{m,m}, X_M)$ and $\Gamma_R(\mathcal{I}_{n,n}, X_n)$ is $\mathcal{O}(|X_n||\mathcal{I}_{n,n}|)$. Hence steps (i) to (vi) overall have time and space complexity at best $\mathcal{O}(n!)$.

For step (vii), the distance between any two vertices in a graph can be found in a number of ways. One approach would be to apply the Floyd-Warshall Algorithm to compute the shortest path between every pair of vertices in $\Delta_{n,r}$; the time complexity of Floyd-Warshall is $\mathcal{O}(n^3)$ where n is the number of vertices in the graph. Another approach is to perform a depth or breadth first search. The version implemented by Beule et al. (2022) uses a breadth first search that also utilises the automorphism group of the graph to avoid visiting multiple identical branches. The automorphism groups of the graphs $\Delta_{n,r}$ are non-trivial when $r \neq 0$ and this approach seems to offer the best performance; see Table 1. Due to its high complexity the exact algorithm given above is only applicable for relatively small values of n; see Table 2.

Table 1 Sizes of the automorphism groups of the graph $\Delta_{n,r}$

n r	2	3	4	5	6	7	8
0	1	1	1	1	1	1	1
1	8	72	384	200	288	392	512
2	2	144	1024	400	576	784	1024
3	–	72	128	200	288	392	512
4	–	–	128	200	288	392	512
5	–	–	–	200	288	392	512
6	–	–	–	–	288	392	512
7	–	–	–	–	–	392	512
8	–	–	–	–	–	–	512

To the best of the authors' knowledge, it is not clear whether there exists a polynomial time algorithm for Problem 4.4. This problem is potentially a computationally difficult problem, and so from a practical perspective it appears that approximation based approaches, or variations, offer the most promise moving forward.

To highlight this, we finish this section by showing that a variation of Problem 4.4—whether two genomes of equal size are an equivalent inversion/deletion distance from their most recent common ancestor—is **NP**-complete. Since genomes of equal size arise from their common ancestor via the same number of deletions, by Theorem 4.5 this problem is equivalent to a problem called BALANCEDSORT. BALANCEDSORT takes a partial permutation $\sigma \in \mathcal{I}_{m,n}$ and $k \in \mathbb{N}$, and asks whether there exist sequences t_m and t_n of inversions in $\mathcal{T}_m$ and $\mathcal{T}_n$ respectively with $\ell(t_m) + \ell(t_n) \leq k$ such that $t_m \sigma t_n \in \mathcal{POI}_{m,n}$ and $\ell(t_m) = \ell(t_n)$. Note that we may consider $\mathcal{POI}_{m,n}$ instead of $\mathcal{POPI}_{m,n}$ here as, if there exists $(g_1, g_2) \in G_1 \times G_2$ such that $t_m \sigma_{g_1,g_2} t_n \in \mathcal{POPI}_{m,n}$, then there exists a reference pair $(h_1, h_2) \in G_1 \times G_2$ obtained by rotating at least one of the genomes such that $t_m \sigma_{h_1,h_2} t_n \in \mathcal{POI}_{m,n}$.

Theorem 5.1 *Determining whether two bacterial genomes of equal size are an equivalent inversion/deletion distance from their most recent common ancestor is NP-complete.*

Proof We proceed by showing that BALANCEDSORT, which is clearly in **NP**, is **NP**-complete. Consider an instance of the well known **NP**-complete problem PARTITION, which consists of a multiset $A = \{a_1, \ldots, a_n\}$ of positive integers and asks if there exists a partition of A into disjoint multisets X and Y such that $\sum_{x \in X} x = \sum_{y \in Y} y$. Construct an instance of BALANCEDSORT from an instance of PARTITION by letting $m = n + \sum_{i=1}^{n} a_i$ and by defining a partial permutation $\sigma \in \mathcal{I}_m$ to be such that

- $(1)\sigma = a_1 + 1$ and $(a_1 + 1)\sigma = 1$,
- $\left(j + \sum_{i=1}^{j-1} a_i \right) \sigma = j + \sum_{i=1}^{j} a_i$ for all $2 \leq j \leq n - 1$ and

Table 2 Time for the various steps in the exact algorithm when applied to every pair of genomes with n and k regions where $n \geq k$

n	3	4	5	6	7	8
$|\mathcal{I}_{n,n}|$	34	209	1,546	13,327	130,922	1,441,729
# of pairs of genomes	153	2,704	69,225	2,503,836	122,783,857	7,859,043,648
Time for (i) to (vi)—(s)	3.880×10^{-8}	1.228×10^{-6}	7.227×10^{-3}	8.556×10^{-2}	2.808×10^{0}	2.595×10^{2}
Time for (vii)—total (s)	4.083×10^{-4}	7.594×10^{-3}	1.938×10^{-1}	7.402×10^{0}	$\sim 6\,\text{min}$	$\sim 7\,\text{h}$
Time for (vii)—mean (s)	2.669×10^{-6}	2.809×10^{-6}	2.800×10^{-6}	2.956×10^{-6}	?	?
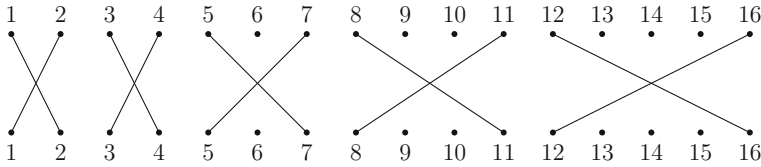
Fig. 12 For an instance $A = \{1, 1, 2, 3, 4\}$ of PARTITION, an instance of BALANCEDSORT is constructed with $\sigma \in \mathcal{I}_{16}$ and $k = 11$

- $$\left(j + \sum_{i=1}^{j} a_i \right) \sigma = \left(j + \sum_{i=1}^{j-1} a_i \right) \text{ for all } 2 \leq j \leq n - 1.$$

The value of k is the sum of all elements in A. Figure 12 illustrates an example of this reduction, which is easily seen to run in polynomial time.

We first show that if there exists a partition of A into X and Y such that $\sum_{x \in X} x = \sum_{y \in Y} y$ then there exists sequences p and q of inversions in $\mathcal{I}_m$ with $\ell(p) + \ell(q) \leq k$ such that $p \sigma q \in \mathcal{POI}_m$ where $\ell(p) = \ell(q)$. Define, without loss of generality, sequences C_{a_j} of inversions (represented by 2-cycles) for all $a_j \in A$ where

$$C_{a_j} = \left(j + \sum_{i=1}^{j-1} a_i, 1 + j + \sum_{i=1}^{j-1} a_i \right) \cdots \left(j - 1 + \sum_{i=1}^{j} a_i, j + \sum_{i=1}^{j} a_i \right)$$

and note that the sequence C_{a_j} removes the crossing consisting of domain elements $j + \sum_{i=1}^{j-1} a_i$ and $j + \sum_{i=1}^{j} a_i$ in a minimal way by left or right multiplication (but not both) without creating additional crossings. Given a partition of A into $X = \{a_{x_1}, \ldots, a_{x_b}\}$ and $Y = \{a_{y_1}, \ldots, a_{y_c}\}$ there is thus sequences $p = C_{a_{x_1}} \cdots C_{a_{x_b}}$ and $q = C_{a_{y_1}} \cdots C_{a_{y_c}}$ of inversions in $\mathcal{I}_m$ (where $\ell(p) + \ell(q) = k$ by construction) such that $p \sigma q \in \mathcal{POI}_m$ and where $\ell(p) = \ell(q)$ since $\sum_{x \in X} x = \sum_{y \in Y} y$.

Conversely, suppose that for the constructed instance of BALANCEDSORT there exists sequences p and q of inversions in $\mathcal{I}_m$ with $\ell(p) + \ell(q) \leq k$ such that $p \sigma q \in \mathcal{POI}_m$ where $\ell(p) = \ell(q)$. By construction $\ell(p) + \ell(q) = k$ where $k = \sum_{a \in A} a$. The sequence p can be written in the form $C_{a_{x_1}} \cdots C_{a_{x_b}}$ and the sequence q can be written in the form $C_{a_{y_1}} \cdots C_{a_{y_c}}$ where the sets $\{x_1, \ldots, x_b\}$ and $\{y_1, \ldots, y_c\}$ are disjoint. This is because each crossing is removed minimally by exclusively left or right multiplication of inversions without creating additional crossings. In other words, these sequences determine a partition of A into $X = \{a_{x_1}, \ldots, a_{x_b}\}$ and $Y = \{a_{y_1}, \ldots, a_{y_c}\}$ where $\sum_{x \in X} x = \sum_{y \in Y} y$ follows from the fact that p and q are such that $\ell(p) = \ell(q)$. □

6 Discussion

This paper has introduced an algebraic framework for modelling two genome rearrangements, inversion and deletion, that are known to occur through the same biological process, namely site-specific recombination. This framework involves the

use of the symmetric inverse monoid, and appears to be the first usage of this type of semigroup model in the study of genome rearrangements. As such a first step, there are on the one hand clear limitations of the model presented, and on the other, clear opportunities for further development.

The most significant limitation involves the scope of the allowable rearrangements. While the model treats a genome as a circular sequence of preserved regions of DNA (a standard way to view genomes in the rearrangement literature), it only permits inversions of adjacent regions, and only permits deletions of a single region at a time. These two simplifying restrictions make the algebra more manageable by restricting the generating sets of the monoids involved. But they are also broadly consistent with each other, since the underlying biological argument behind restricting the length of DNA sequence inverted or deleted is the same, because both arise from the same mechanism. As noted in the Introduction, traditional rearrangement models do not restrict the length of the inverted region, and those that incorporate deletion (such as El-Mabrouk (2000)) allow any length to be deleted, and with equal probability. They also generally allow the opposite operation, insertion, which typically occurs via different biological mechanism and so the savings in the computational simplicity come at an arguable cost to biological faithfulness—as indeed they do in the present paper.

A natural extension to the model presented here would be to allow longer regions to be inverted and/or deleted, perhaps along the lines attempted in Bhatia et al. (2020), which allows longer inversions in a group-theoretic model, but imposes a cost by length. Indeed, some results here, such as Theorem 4.2, apply regardless of the generating set for S_n, or the number of regions being deleted.

Other generalisations may become available as a direct result of the algebraic framework. For instance, the algebraic formalisation using the symmetric inverse monoid can be generalised further by using monoids and categories of binary relations or partial functions. The use of certain binary relations $\lambda : R \to \mathbf{n}$ (or partial functions $\mathbf{n} \to R$ using the convention of positions to regions) allows one to account for repeated region labels, where an ordered pair (r, n) is in λ if and only if the region r appears in position n in a sequences of genome regions. For instance, the sequence $r_1 r_2 r_1 r_3$ of regions where $\{r_1, r_2, r_3\} \subseteq R$ would correspond to the relation $\{(r_1, 1), (r_2, 2), (r_1, 3), (r_3, 4)\}$.

Given sets $\mathbf{m}$ and $\mathbf{n}$ where $m, n \in \mathbb{N}$, the set of relations $\{(x_1, y_1), \ldots, (x_j, y_j)\}$ such that $\{x_1, \ldots, x_j\} \subseteq \mathbf{m}$, $y_i \in \mathbf{n}$ for all $1 \leq i \leq j$ and $y_i \neq y_j$ when $i \neq j$ is denoted by $\widehat{\mathcal{PT}}_{m,n}$, while the set of (analogously defined) partial functions from $\mathbf{m} \to \mathbf{n}$ is denoted $\mathcal{PT}_{m,n}$. One can define a (small) category $\widehat{\mathcal{PT}}$ whose objects are the natural numbers, and where the set of arrows from m to n is the set $\widehat{\mathcal{PT}}_{m,n}$ under the composition of binary relations. Given a relation λ from R to n described above, we can compose on the right by elements of $\widehat{\mathcal{PT}}$ to represent not only inversions and deletions (since $\widehat{\mathcal{PT}}$ contains $\mathcal{I}$), but also to represent duplications of regions. To model a duplication we multiply on the right by relations of the form $\lambda_{i;n} \in \widehat{\mathcal{PT}}_{n,n+1}$ where, without loss of generality (as in Fig. 13), we have

$$\lambda_{i;n} = \{(1, 1), \ldots, (i, i), (i, i + 1), (i + 1, i + 2), \ldots, (n, n + 1)\}.$$
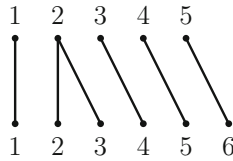
Fig. 13 The relation diagram of $\Lambda_{2;5} \in \widehat{\mathcal{PT}}_{5,6}$. Note that unlike the deletion in Fig. 6 where the degree of the vertex labelled by 2 in the upper row was 0 (to represent the fact the region in position 2 was deleted), the upper row vertex labelled by 2 in this instance has degree 2 to represent the fact that the region in position 2 has been duplicated

With this algebraic framework in mind, it is possible to consider the new problem of constructing the most recent common ancestor of two bacterial genomes (which may have repeated regions) under the three operations of inversions, deletions and duplications. Since the problem of reconstructing the most recent common ancestor of two genomes under exclusively inversions and deletions is a special case of this new problem, the same asymmetry present in the inversion/deletion model is also present in the inversion/deletion/duplication model given that only pre-existing genome regions may be duplicated. It is then natural to investigate whether similar combinatorial optimization problems regarding elements of $\widehat{\mathcal{PT}}$ have analogous interpretations to those presented here, such as Problem 4.4.

Finally, it would be interesting to explore whether the framework developed here could be cast in the representation-theoretic framework designed for maximum likelihood estimates for genome rearrangement models (Serdoz et al. 2017), that is presented in Sumner et al. (2017); Terauds and Sumner (2022). Indeed, on one hand, Terauds and Sumner (2022) remark that it may be generalised to models using semigroups, while on the other hand, the representation theory of finite monoids including that of the symmetric inverse monoid has been well studied (Steinberg et al. 2016; Munn 1964; Solomon 2002).

Declarations

References

Alexandrino AO, Brito KL, Oliveira AR, Dias U, Dias Z (2021a) Reversal distance on genomes with different gene content and intergenic regions information. In: Martín-Vide C, Vega-Rodríguez MA, Wheeler T (eds) Algorithms for Computational Biology. Springer, pp 121–133

Alexandrino AO, Oliveira AR, Dias U, Dias Z (2021) Genome rearrangement distance with reversals, transpositions, and indels. J Comput Biol 28(3):235–247

Beule JD, Jonušas J, Mitchell JD, Torpey M, Tsalakou M, Wilson WA (2022) Digraphs—GAP package, version 1.5.2, Mar. https://digraphs.github.io/Digraphs

Bhatia S, Feijão P, Francis AR (2018) Position and content paradigms in genome rearrangements: the wild and crazy world of permutations in genomics. Bull Math Biol 80(12):3227–3246

Bhatia S, Egri-Nagy A, Serdoz S, Praeger CE, Gebhardt V, Francis A (2020) A path-deformation framework for determining weighted genome rearrangement distance. Front Genet 11:1035

Bochkareva OO, Dranenko NO, Ocheredko ES, Kanevsky GM, Lozinsky YN, Khalaycheva VA, Artamonova II, Gelfand MS (2018) Genome rearrangements and phylogeny reconstruction in *Yersinia pestis*. PeerJ 6:e4545

Braga MDV, Willing E, Stoye J (2010) Genomic distance with dcj and indels. In: International workshop on algorithms in bioinformatics. Springer, pp 90–101

Caprara A (1997) Sorting by reversals is difficult. In: Proceedings of the first annual international conference on computational molecular biology. ACM New York, pp 75–83

Catarino PM, Higgins PM (1999) The monoid of orientation-perserving mappings on a chain. In: Semigroup Forum, vol 58

Dalevi DA, Niklas E, Kimmo E, Andersson SGE (2002) Measuring genome divergence in bacteria: a case study using Chlamydian data. J Mol Evol 55(1):24–36. https://doi.org/10.1007/s00239-001-0087-9

Darling AE, Miklós I, Ragan MA (2008) Dynamics of genome rearrangement in bacterial populations. PLoS Genet 4(7):1

East J (2020) Presentations for tensor categories. arXiv preprint, arXiv: 2005.01953

Egri-Nagy A, Gebhardt V, Tanaka MM, Francis AR (2014) Group-theoretic models of the inversion process in bacterial genomes. J Math Biol 69(1):243–265

Eisen JA, Heidelberg JF, White O, Salzberg SL (2000) Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol 1(6):1

El-Mabrouk N (2000) Genome rearrangement by reversals and insertions/deletions of contiguous segments. In: Annual Symposium on Combinatorial Pattern Matching. Springer, pp 222–234

Fertin G, Labarre A, Rusu I, Tannier É, Vialette S (2009). Combinatorics of genome rearrangements. MIT Press, Cambridge

Francis AR (2014) An algebraic view of bacterial genome evolution. J Math Biol 69(6–7):1693–1718

Froidure V, Pin J-E (1997) Algorithms for computing finite semigroups. In: Foundations of computational mathematics (Rio de Janeiro, 1997). Springer, Berlin, pp 112–126

Galvao GR, Baudet C, Dias Z (2017) Sorting circular permutations by super short reversals. IEEE/ACM Trans Comput Biol Bioinf (TCBB) 14(3):620–633

Hannenhalli S, Pevzner PA (1999) Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. J ACM (JACM) 46(1):1–27

Jerrum MR (1985) The complexity of finding minimum-length generator sequences. Theoret Comput Sci 36:265–289

Lefebvre JF, El-Mabrouk N, Tillier E, Sankoff D (2003) Detection and validation of single gene inversions. Bioinformatics 19(Suppl 1):i190–i196

Marron M, Swenson KM, Moret BME (2004) Genomic distances under deletions and insertions. Theoret Comput Sci 325(3):347–360

McAlister DB (1998) Semigroups generated by a group and an idempotent. Comm Algebra 26(2):243–254

Munn WD (1964) Matrix representations of inverse semigroups. Proc. Lond. Math. Soc. 3(14):165–181

Oliveira AR, Brito KL, Dias Z, Dias U (2018) Sorting by weighted reversals and transpositions. In: Brazilian Symposium on Bioinformatics. Springer, pp 38–49

Plasterk RHA, Ilmer TAM, Van de Putte P (1983) Site-specific recombination by Gin of bacteriophage Mu: inversions and deletions. Virology 127(1):24–36

Raeside C, Gaffé J, Deatherage DE, Tenaillon O, Briska AM, Ptashkin RN, Cruveiller S, Médigue C, Lenski RE, Barrick JE et al (2014) Large chromosomal rearrangements during a long-term evolution experiment with *Escherichia coli*. MBio 5(5):e01377–14

Saier MH (2008) The bacterial chromosome. Crit Rev Biochem Mol Biol 43(2):89–134

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425

Seoighe C, Federspiel N, Jones T, Hansen N, Bivolarovic V, Surzycki R, Tamse R, Komp C, Huizar L, Davis RW et al (2000) Prevalence of small inversions in yeast gene order evolution. Proc Natl Acad Sci 97(26):14433–14437

Serdoz S, Egri-Nagy A, Sumner J, Holland BR, Jarvis PD, Tanaka MM, Francis AR (2017) Maximum likelihood estimates of pairwise rearrangement distances. J Theor Biol 423:31–40

Shao M, Lin Y (2012) Approximating the edit distance for genomes with duplicate genes under DCJ, insertion and deletion. In: BMC Bioinformatics, vol 13. Springer, pp 1–9

Solomon L (2002) Representations of the rook monoid. J Algebra 256(2):309–342

Steinberg B et al (2016) Representation theory of finite monoids. Springer, Berlin

Sumner JG, Jarvis PD, Francis AR (2017) A representation-theoretic approach to the calculation of evolutionary distance in bacteria. J Phys A: Math Theor 50(33):335601

Terauds V, Sumner J (2022) A new algebraic approach to genome rearrangement models. J Math Biol 84(6):1–32

Watterson GA, Ewens WJ, Hall TE, Morgan A (1982) The chromosome inversion problem. J Theoret Biol 99(1):1–7