A mean first passage time genome rearrangement distance

Andrew R. Francis · Henry P. Wynn

Received: 12 April 2019 / Revised: 11 March 2020 / Published online: 6 April 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract
This paper introduces a new way to define a genome rearrangement distance, using the concept of mean first passage time from probability theory. Crucially, this distance provides a genuine metric on genome space. We develop the theory and introduce a link to a graph-based zeta function. The approach is very general and can be applied to a wide variety of group-theoretic models of genome evolution.

Mathematics Subject Classification 92B99

1 Introduction

Estimating evolutionary distances between organisms is a key ingredient in most approaches to phylogenetic reconstruction. Making such estimates broadly involves two steps: specifying an evolutionary model (the way in which the organisms can change), and deciding what metric to use. The most straightforward approach to the latter is to ask for the least number of steps between the two organisms under the model; this is the minimal distance.

By definition, a minimal distance can only underestimate the true distance, and there is considerable interest in finding ways to estimate distance that are less problematic (Jukes and Cantor 1969; Wang and Warnow 2001; Serdoz et al. 2017). In this paper, we take a new approach to this problem in the context of genome rearrangements, by adapting a construction from probability theory within a framework that also exploits group theory (Egri-Nagy et al. 2014; Francis 2014). Thus, the methods in this paper bring together mathematical tools from disparate fields to address a problem.
in molecular evolution. Specifically, we show how one may be able to calculate the \textit{mean first passage time (MFPT)} between two organisms, and we put this forward as an alternative to the minimal distance.

The “true distance” is the actual number of inversions that take place on the path between two genomes that have the same gene content. As said, the minimal distance is an underestimate of this, and unlikely to be exhibited by a random process. The distance in this paper can be considered as an alternative to the maximum likelihood estimate approach, and has some differences and potential advantages over it. The main point of difference is that the MLE does not always exist in the sense that the likelihood function may not have a maximum (evidence suggests this occurs over 40\% of the time Serdoz et al. 2017; Terauds and Sumner 2019), and may conceivably not have a \textit{unique} maximum (though this has yet to be observed). The mean first passage time always exists. Second, in principal the MLE has an infinite series computation, although this has been reduced to the computation of the eigenvalues of the irreducible representations of certain group elements (Sumner et al. 2017). The formulae for the mean first passage time in this paper are closed form, though like the representation theory approach are constrained by linear algebra to compute. A third point of difference is that looking at the \textit{first} passage time of a path from genome \( A \) to genome \( B \) avoids the issue of multiple visits of paths that are considered in the MLE approach.

The mean first passage time to a target is the average time it takes for a random process to reach the target for the first time, along any path. This will depend typically on the structure of the underlying graph that represents an evolutionary process on genome space; in our case a group structure. The first passage time is a well-studied quantity in the theory of Markov processes and network flows. It can be defined on any strongly connected directed graph (that is, for which each pair of vertices is connected by a directed path). A moment generating function for the first passage time can be computed using the structure of the graph, including closed loops within the graph. This is an approach due to Mason in the 1950s in the system theory literature (Mason 1953). It can also be computed using determinants of matrices associated with the adjacency matrix of the graph (Butler and Huzurbazar 1997).

We apply the latter approach, using the adjacency matrix of the Cayley graph of a group, \( G \), that represents genome space under a particular model of evolution. The Cayley graph of a group is a graph whose vertices represent the group elements, and each directed edge corresponds to multiplication by one of the generators of the group, with the reverse arrow being multiplication by the inverse element (here we use right multiplication). Note that typically a list of generators is not unique, in which case nor is the Cayley graph or our metric. The method builds on the algebraic models of bacterial inversion introduced in Egri-Nagy et al. (2014). In this framework, each vertex of the Cayley graph corresponds to a unique genome arrangement and the edges to possible evolutionary events. While we will consider the general situation, in most examples arising from this framework the generators of the group are self-inverse (for instance an inversion done twice returns the genome to the original state), which means that the edges of the corresponding Cayley graph are often bi-directed.

Now, assume each (directed) edge \((i, j)\) has a Markov transition probability \( p_{ij} > 0 \) from \( i \) to \( j \). In addition, for each edge \((i, j)\) there is an independent random “passage time” \( X_{ij} \), from \( i \) to \( j \). The \textit{first} passage time is then the time taken for a random walk
starting at state \( i \) to reach the target state \( j \), *for the first time*. We assume that the graph is strongly connected, so that there is at least one path between any two vertices. Once the probability distribution function of this first passage time is given, our distance \( d_{i,j} \) is defined as its mean \( \mu_{i,j} \). We assume:

1. The moment generating functions, \( \{m_{ij}(s)\} \) of the passage times \( \{X_{ij}\} \) are known, and
2. The Markov transition probabilities \( \{p_{ij}\} \) are known.

In the terminology of Markov processes, a Cayley graph may not yield an aperiodic system, and therefore the underlying Markov chain will not be positive recurrent. Thus, that part of the stochastic process theory which requires aperiodicity will not apply. But the formulae here are quite general and can also be written in terms of purely combinatorial zeta functions for graphs and sub-graphs associated with paths (described in Sect. 6).

The structure of the paper is as follows. We begin in Sect. 2 with an introduction to the motivating problem from bacterial genome rearrangements, together with an explanation of the group-theoretic framework with which we study it. This is followed by a straightforward account of the Markov flow theory in Sect. 3. The definition of our mean first passage time distance is given in Sect. 4, followed by some fully worked examples in Sect. 5. We return to the Mason rule in Sect. 6 giving links to zeta functions on graphs. A discussion section points to further research.

# 2 Background to bacterial genome rearrangements and group-theoretic models

In this section we explain the basic biological information about bacterial genome rearrangements, and we present some necessary details of the algebraic models from Egri-Nagy et al. (2014) that this paper relies on.

Large scale rearrangements, in which whole regions of the chromosome are moved around relative to each other, are a significant driver of evolutionary adaptation in the case of bacteria. Large scale rearrangements are uncommon in eukaryotic nuclear DNA, though they are a feature of mitochondrial DNA, probably because of its heritage as an ancestral bacterial invasion (the DNA of almost all bacterial genomes and mitochondria is circular). The mechanisms within the cell that give rise to these rearrangements revolve around enzymes called *site-specific recombinases*, which cut two double-helical strands that are adjacent in the cell, and rejoin them in a new way, changing the sequence of the chromosome. They facilitate the movement of genes around a chromosome (which can have a fitness effect) as well as the acquisition of new genetic material (through horizontal gene transfer), and deletion of redundant DNA.

We will focus on single-celled organisms and large scale rearrangements on a single chromosome because they play an important role in establishing phylogenetic relationships in these cases. This set-up includes all bacteria, but omits models such as the influential double-cut-and-join (DCJ) (Yancopoulos et al. 2005; Bergeron et al. 2006). The rearrangement events we will focus on are inversion and translocation, because both of these events are “invertible” (can be undone), and so are suited to a
group-theoretic treatment (Egri-Nagy et al. 2014; Francis 2014). By a model of rearrangement, we mean three things: a genome structure; a set of allowable operations that rearrange the regions on the genome; and a probability distribution on the operations. This slightly generalizes the algebraic structure described in Egri-Nagy et al. (2014), in the inclusion of the probability distribution, as independently done in Terauds and Sumner (2019). We will assume no deletion or insertion of regions occurs, so that all possible rearrangements have the same set of regions.

The two models of general interest we consider and in which we include or omit the orientation of the DNA, are illustrated in Fig. 1 for the circular case. The other genome geometry we will mention is a lineal chromosome, which has regions arranged along a line, again either including orientation or not (we use the word lineal to avoid confusion from use of the word “linear” for this geometry).

Inversion takes a region or set of contiguous regions and reverses their relative positions, while translocation takes a region or set of contiguous regions and moves it past another set of regions, as illustrated in Fig. 2. Such rearrangements are called signed if the orientation of the regions is considered.

The dominant method for establishing distance in any of these models is by calculating the minimal distance. In some cases, this can be done very fast. In particular, there is a large body of literature establishing the minimal distance under the model in which all inversions occur with equal probability (Sankoff et al. 1992; Bafna and Pevzner 1993). These methods are chiefly combinatorial, constructing a “breakpoint graph” whose features give the distance. Recently, group-theoretic methods have been introduced that rely on results in Coxeter groups, and are effective for inversions of only two adjacent regions (we will call these 2-inversions) (Egri-Nagy et al. 2014; Francis 2014).

Several alternatives to the minimal distance to estimate the evolutionary distance in rearrangement models have been proposed. These are summarized in Serdoz et al. (2017), which develops a way to calculate a maximum likelihood estimate for the evolutionary distance, further developed in Sumner et al. (2017). Aside from the MLE of Serdoz et al. (2017), these are dominated by approaches that use the relationships between the minimal distance and the true distance obtained using simulation studies (for instance Wang and Warnow 2001; Dalevi and Eriksen 2008).

In this paper we focus on rearrangement models that are suited to a group-theoretic approach, namely those for which the permitted evolutionary operations are “invertible”: there is a permitted operation that undoes each one. Usually in this context such operations are self-inverse: they undo themselves. The core of this approach is the
observation that if a family of rearrangement events are each invertible, then, because
the composition of such events is associative \((a \circ (b \circ c) = (a \circ b) \circ c)\), they
generate a group that is acting on the set of genome arrangements. A sequence of such
rearrangement events is then a composition of a sequence of group generators. This
gives an interpretation of the distance, and of evolutionary history, in terms of paths
and path lengths on the Cayley graph of the group (Clark et al. 2019), which we now
describe.

If one particular genome arrangement is chosen as a “reference genome”, then every
other genome arrangement can be associated with a unique group element defined by
the product of the generators on a path to it from the reference genome. Despite
there being possibly many paths to the genome, the product is well-defined regardless
of which path is chosen, giving a single group element. That is, there is a one-to-
one correspondence between the set of genomes, and the set of elements of the group
generated by the rearrangements. This one-to-one correspondence depends on a choice
of rearrangement operations—which give the generators of the group—and it maps
the chosen reference genome to the identity element of the group.

The correspondence between genome arrangements and group elements means
that the space of genomes can now be considered as a Cayley graph: the graph whose
vertices are elements of the group, and in which there is an edge from group element
g to h if \(h = gs\) for some generator s. The minimal distance between two genomes is
then simply the length of a minimal path on the Cayley graph, and the true evolutionary history is a walk on the Cayley graph.

The methods we develop here are applicable to several models of rearrangement. First, the widely studied cases of lineal or circular genomes in which regions of DNA are considered as either oriented (signed) or unoriented. Second, one may allow different evolutionary events, including inversions of different lengths of DNA (measured by the number of regions inverted in a single event), as well as translocations. Some of the models that are relevant are listed in Table 2, in Appendix A.

3 Markov flow models

Consider a directed simple graph $G(V, E)$ with vertex set $V$, $|V| = n$ and edge set $E$ consisting of ordered pairs $(i, j)$. By “simple” we mean that $G$ has no parallel edges (so that $(i, j)$ defines at most one edge) and no loops $(i, i)$ (edges from a vertex to itself).

It helps intuition to consider transition as the movement of a hypothetical particle. We assume that given any vertices $i$ and $j$, a particle starting at vertex $i$ can reach vertex $j$ along a directed path (this strongly connected condition is satisfied for Cayley graphs, for instance). If $(i, j) \in E$ and the particle is at vertex $i$, the particle travels directly to vertex $j$ with probability $p_{ij} > 0$, and $\sum_{j \neq i} p_{ij} = 1$ for all $i$ (with $p_{ij} = 0$ if $(i, j) \notin E$); so the particle having reached vertex $i$ must move immediately away from $i$. Let the random variable $X_{ij}$ denote the inter-arrival time (passage time) along edge $(i, j) \in E$. We assume that all the $X_{ij}$ for the travel of the particle are independent and that each $X_{ij}$ has a well-defined moment generating function:

$$m_{ij}(s) = \mathbb{E} X_{ij} \{ \exp(s X_{ij}) \}.$$  

Let $Y_{ij}$ be the first passage time from vertex $i$ to vertex $j$. To study this we consider the directed subgraph $G(V, E_{ij})$ where

$$E_{ij} = E \setminus \{(j, k) : (j, k) \in E\}.$$  

That is, we remove all edges out of $j$, thereby turning $j$ into an absorbing state. The first passage time $Y_{ij}$ is the sum of all $X_{ij}$ realised from vertex $i$ till the particle arrives at vertex $j$ for the first time, for the graph $G(V, E_{ij})$. Note that the particle can spend an arbitrary large but finite length of time in circuits (if there are circuits) and different visits to an edge, say $(i, j)$, are assumed to give independent copies of $X_{ij}$.

What we call here the Mason rule (Theorem 3.1) is a version of the original rule which is also available via Markov renewal theory (see Butler and Huzurbazar 1997; Pyke 1961; Howard 1960; Howard and Matheson 1971). It is sometimes referred to as the cofactor rule. The rule gives the moment generating function, $\tilde{M}_{ij}(s)$ of the $Y_{ij}$ in terms of the $p_{ij}$ and $m_{ij}(s)$.

Let $P = \{p_{ij}\}$ denote the Markov matrix for the process, and let $Q(s) = \{p_{ij}m_{ij}(s)\}$ denote the transmittance matrix, both of size $|V| \times |V|$, whose entries we will write $q_{ij}(s) := p_{ij}m_{ij}(s)$. Note that $q_{ij}(s) = 0$ if $p_{ij} = 0$. The following theorem is a version of Mason’s rule.
Theorem 3.1 (Mason’s rule (Mason 1953)) For a Markov flow model with transmittance matrix $Q(s)$, the moment generating function $\tilde{M}_{ij}(s)$ of the first passage time $Y_{ij}$ from vertex $i$ to vertex $j$, is given by the $i, j$ entry in the matrix $(I - Q_{[j]})^{-1}(s)$, where $Q_{[j]}$ is obtained from $Q(s)$ by setting all transmittances $q_{jk}$, $(j, k) \in E$ equal to zero.

Note that the matrix inverse in the theorem exists via the theory of absorbing states for Markov chains.

4 The mean first passage time distance

We define our distance $d_{ij}, i \rightarrow j$ as follows.

Definition 4.1 For a directed graph $G(V, E)$ with Markov transition matrix $P = \{p_{ij}\}$ and inter-arrival moment generating function matrix $M(s) = \{m_{ij}(s)\}$, we define the mean first passage time (MFPT) distance as

$$d_{ij} = \mathbb{E}_{Y_{ij}}(Y_{ij}),$$

where $Y_{ij}$ is the first passage time from vertex $i$ to vertex $j$.

The distance depends on $M(s)$ only via the edge means $\mu_{ij} = \mathbb{E}(X_{ij}) = m'_{ij}(0)$. Let $e_i$ be the $i$-th unit vector (1 in entry $i$ and 0 elsewhere). Then, using the $e_i$ to pick out the entries, Mason’s rule (Theorem 3.1) immediately gives us:

Proposition 4.2

$$d_{ij} = \frac{\partial}{\partial s} \left( e_i^T (I - Q_{[j]}(s))^{-1} e_j \right) \bigg|_{s=0}. $$

While this is already a practically useful expression, we can further develop an explicit expression for $d_{ij}$ as follows. By the formula for differentiating inverses,

$$d_{ij} = \frac{\partial}{\partial s} \left( e_i^T (I - Q_{[j]}(s))^{-1} e_j \right) \bigg|_{s=0} = -\left\{ e_i^T (I - Q_{[j]}(s))^{-1} \left( \frac{\partial}{\partial s} (I - Q_{[j]}(s)) \right) (I - Q_{[j]}(s))^{-1} e_j \right\} \bigg|_{s=0}. $$

Note that

$$I - Q_{[j]}(s) = I - P_{[j]} \circ M_{[j]}(s),$$

where “$\circ$” means the Schur (Hadamard, entry-wise) product, and $P_{[j]}$ is the matrix $P$ with the $j$-th row set to zero.

Define $\mathbb{M} := M'(0) = \{\mu_{ij}\}$ to be the matrix of passage time edge means and write $\mathbb{M}_{[j]} = M'_{[j]}(0)$ for the matrix formed from $M$ by excluding means for edges out of
vertex $j$ (that is, “zero-ing out” the $j$-th row). Then, using 3.1 we have an explicit formula for the distances (noting that $m_{ij}(0) = 1$ for all $(i, j) \in E$ with $i \neq j$):

$$d_{ij} = e_i (I - P_{[j]})^{-1} (P_{[j]} \circ M_{[j]}) (I - P_{[j]})^{-1} e_j.$$ 

Examples of this computation can be found in Sect. 5, exploiting the efficiency of the linear algebra version in Proposition 4.2.

Note that the $(r, t)$ entry of $(I - P_{[j]})^{-1}$ is the passage time from $r$ to $t$ for the graph in which $j$ has been made an absorbing state and in which every edge has fixed unit travel time. Let us label these entries $\{z_{rt}\}$. Then expanding the matrices we have

$$d_{ij} = \sum_{(r, t)} z_{ir} p_{rt} \mu_{rt} z_{tj}.$$ 

Each term $\mu_{rt}$ in this expansion has coefficient

$$\theta_{r, t : i, j} = z_{ir} z_{tj} p_{rt},$$ 

whose informal interpretation is as follows. Each distinct path from $i$ to $j$ may or may not use the edge $(r, t)$. Those that use $(r, t)$ each contribute a weight to the mean $\mu_{rt}$. Each such path must enter at $r$ and leave at $t$. The paths into $r$ have total passage time $z_{ir}$ and those out of $r$ and being absorbed at $j$ have total passage time $z_{tj}$. Independence gives the contribution as $z_{ir} z_{tj} p_{rt}$. A technical point is that because the matrix $P_{[j]}$ has an absorbing state, and the full chain is connected, all the terms are finite.

As defined, the distance $d_{ij}$ satisfies the directional triangular inequality: $d_{ik} \leq d_{ij} + d_{jk}$ for distinct vertices $i, j, k$. This is proved by splitting events into two disjoint types: (i) starting in $i$ we reach $j$ first before $k$ and (ii) starting in $i$ we reach $k$ first before $j$. In both cases the inequality holds: in (i) we have equality and in (ii) by assumption $d_{ik} \leq d_{ij}$. The symmetry needed to be an ordinary distance requires $d_{ij} = d_{ji}$ for all $i \neq j$. The Cayley graph version defined next satisfies this condition, because generators being involutions mean that each edge is bi-directional with the same weight in each direction, and so paths in one direction correspond to those in the other, carrying the same weight.

5 The Cayley graph case

As explained in the introduction, the distance $d_{ij}$ between group elements $g_i$ and $g_j$ is defined to be the mean of the first passage time, when the full directed graph is a Cayley graph, $C(G, S)$ of a group $G$ generated by the elements $S$ corresponding to the evolutionary events of interest.

Although the theory allows the moment generating function $m_{ij}(s)$ to depend on $i$ and $j$, for simplicity in the Cayley graph case we set all the $m_{ij}$ equal to a function $m(s)$, and all the $p_{ij}$ equal. If there are $k$ generators for the group, then for any edge $(i, j)$ of $C(G, S)$, 

Springer
\[ q_{ij}(s) = \frac{1}{k} m(s), \]

and

\[ Q(s) = \frac{1}{k} m(s) A(G), \]

where \( A(G) \) is the adjacency matrix of the Cayley graph \( C(G, S) \) of \( G \) generated by \( S \). Then we have:

\[ I - Q_{[j]}(s) = I - z A_{[j]} \]  \hspace{1cm} (5.1)

with \( z = \frac{1}{k} m(s) \) and \( A_{[j]} \) is the adjacency matrix of the graph obtained from \( C(G, S) \), by deleting all the arrows out of \( j \).

We begin with an elementary result that describes when two group elements have the same mean first passage time (considered as distances from the identity element), contingent on properties related to the normaliser of the set of generators \( S \) of the group \( G \). Recall that the normaliser of \( S \) in \( G \) is the subgroup of \( G \) defined by \( N_G(S) := \{ \sigma \in G \mid \sigma^{-1} S \sigma = S \} \). The normaliser acts on the group \( G \) by conjugation, and the orbits of this action partition both the set of generators, by definition of the normaliser, and the group itself, by definition of an orbit.

**Proposition 5.1** Suppose we have a random walk on a Cayley graph \( C(G, S) \) beginning at the identity element, in which the generators \( S \) of \( G \) are all equally probable. Suppose in addition that the inter-arrival times \( X_{ij} \) have the same distribution for all edges \((i, j)\) corresponding to generators that are in the same orbit of the action of \( N_G(S) \). Then, if two group elements are in the same orbit of \( N_G(S) \), they will have the same mean first passage time.

**Proof** Theorem 4.2 of Clark et al. (2019) shows that two group elements conjugate by the normaliser of the generators will have order-isomorphic “intervals”, meaning that the path structures from the identity to each of them are isomorphic as partially ordered sets. The additional requirements here about the probabilities of the generators, and the inter-arrival times, ensures that the distribution of first passage times to each element will be the same, and in particular the mean first passage times will be the same. \( \Box \)

A key property of a Cayley graph is vertex transitivity, which, heuristically, means that the graph looks the same from any vertex. Thus, any edge \( g \rightarrow gs \), where \( s \) is a generator can be transformed to \( e \rightarrow s \) by left multiplying by \( g^{-1} \). We see, more generally, that the whole graph remains the same if we premultiply by a fixed \( g^{-1} \) at every vertex.

For simplicity, in the examples below we will use \( m_{ij}(s) = e^{s} \), for all \((i, j) \in E\), which corresponds to a fixed (non-random) time between vertices of one unit. That is, we take \( X_{ij} \) to be a discrete random variable which takes the value 1 with probability 1, so that the moment generating function is just \( e^{s} \).
5.1 Example: $S_3$ with standard (Coxeter) generators

We carry out the computations for the Cayley graph of the symmetric (permutation) group on three elements. With identity $e$ and two generators $g_1$ and $g_2$, respectively (transpositions $(1\,2)$ and $(2\,3)$ as shown in Fig. 3, acting on the right), the elements, labeled as vertices are

\[ \{v_1 = e, v_2 = g_1 = (1\,2), v_3 = g_2 = (2\,3), v_4 = g_1g_2 = (1\,2\,3), v_5 = g_2g_1 = (1\,3\,2), v_6 = g_1g_2g_1 = g_2g_1g_2 = (1\,3) \}. \]

The adjacency matrix and its absorbing form based on $v_6$ are

\[
A = \begin{pmatrix}
0 & 1 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 1 & 0
\end{pmatrix},
A_{[6]} = \begin{pmatrix}
0 & 1 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.
\]

The expressions below give the entries of the matrix $(I - zA_{[6]})^{-1}$.

\[
((I - zA_{[2]})^{-1})_{1,2} = ((I - zA_{[6]})^{-1})_{4,6} = \frac{z - 2z^3}{1 - 3z^2},
\]
\[
((I - zA_{[4]})^{-1})_{1,4} = ((I - zA_{[6]})^{-1})_{2,6} = \frac{z^2}{1 - 3z^2},
\]
\[
((I - zA_{[6]})^{-1})_{1,6} = \frac{2z^3}{1 - 3z^2}.
\]

Note that by vertex transitivity, we do not need to calculate $A_{[2]}$ or $A_{[4]}$: the set of paths from vertex 1 to vertex 4 ($e$ to $g_1g_2$) is in one-to-one correspondence with the set of paths from vertex 2 to vertex 6 ($g_1$ to $g_1g_2g_1$).

We now use the simple form of Equation (5.1), setting $k = 2$, $z = \frac{1}{2}e^s$, differentiating with respect to $s$ and setting $s = 0$. This gives the distinct mean first passage
times from $e$ to (i) $v_2$ or $v_3$, (ii) $v_4$ or $v_5$ and (iii) $v_6$, as, respectively, 5, 8 and 9. This compares to the shortest (geodesic) distances of 1, 2 and 3 respectively.

That some distances are equal supports the fact that in the group theoretic sense elements which are conjugate by the normaliser of the generators will have identical mean first passage time, as in Proposition 5.1.

Typically, the MFTP distance has more distinct values than minimal distance (indicating a finer discrimination). But even here, with the same number of distinct values, the MPFT distance changes the scale to provide more information about the relative random walk length between elements.

5.2 Example: $S_3$ with circular generators

Working with circular generators adds a generator across the top of the appropriate circle diagram (representing the circular genome with three regions), namely (1 3), so that the generators are $\{(1 2), (2 3), (1 3)\}$. The Cayley graph is a complete bipartite graph, namely $K_{3,3}$ (Fig. 3).

Ordering the elements $e$, (12), (23), (13), (123), (132), the adjacency matrix with the last state absorbing is

$$A_{[6]} = \begin{pmatrix}
0 & 1 & 1 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 & 1 \\
0 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 
\end{pmatrix}.$$ 

The entries in the last column of $(I - zA_{[6]})^{-1}$ are $3z^2/(1 - 6z^2)$, $z/(1 - 6z^2)$, $z/(1 - 6z^2)$, $3z^2/(1 - 6z^2)$, 1 respectively. By substituting $k = 3, z = \frac{1}{3}e^s$, differentiating with respect to $s$ and then setting $s = 0$, we obtain mean first passage times from $e$ of $(6, 5, 5, 6, 0)$ (which compare to $(2, 1, 1, 1, 2, 0)$ for the shortest distance). This means the MFPT distance between any two elements on opposite sides of the bipartite graph is 5, and between two distinct elements on the same side is 6.

5.3 Example: $S_4$ with standard and circular generators

The Cayley graphs for $S_4$, for both standard and circular generating sets, are shown in Figs. 4 and 5, Appendix B.

Instead of calculating the inverse of the matrix $(I - zA_{[24]})$ (which is computationally difficult), we calculate the terms we need by using the fact that any distance on the Cayley graph is equivalent to a distance to the longest word (see for instance Humphreys (1990), Part I). So we only need to compute that final column of $(I - zA_{[24]})^{-1}$, which can be done by a simple row reduction. That is to say, if the final column is given by the vector $v$ then it is the solution to the matrix equation $(I - zA_{[24]})v = e_{24}$, (recalling that $e_{24} \in \mathbb{Q}_{24}$ is the vector whose entries are zero except for a 1 in the last position).
Table 1 Mean first passage time distances in $S_4$ to the longest word, using Coxeter generators (lineal genome) on the left ($\{(1 \ 2), \ (2 \ 3), \ (3 \ 4)\}$) and circular generators on the right ($\{(1 \ 2), \ (2 \ 3), \ (3 \ 4), \ (4 \ 1)\}$), in comparison to the minimal distances in the middle columns. Ordering by minimal length in each case.

| Lineal generators | Minimal distance | MFPT distance | Circular generators | Minimal distance | MFPT distance |
|-------------------|-----------------|---------------|---------------------|-----------------|---------------|
| \() \ | 6 | $1296/28 = 46.3$ | \() \ | 4 | 32 |
| \((1 \ 2)\) \ | 5 | $1273/28 = 45.5$ | \((1 \ 2)\) \ | 3 | 31 |
| \((3 \ 4)\) \ | 5 | $1273/28 = 45.5$ | \((2 \ 3)\) \ | 3 | 31 |
| \((2 \ 3)\) \ | 5 | $1258/28 = 44.9$ | \((1 \ 4)\) \ | 3 | 31 |
| \((1 \ 2)(3 \ 4)\) \ | 4 | $1242/28 = 44.4$ | \((1 \ 3)\) \ | 3 | 31 |
| \((1 \ 3 \ 2)\) \ | 4 | $1197/28 = 42.8$ | \((3 \ 4)\) \ | 3 | 31 |
| \((1 \ 2 \ 3)\) \ | 4 | $1197/28 = 42.8$ | \((2 \ 4)\) \ | 3 | 31 |
| \((2 \ 4 \ 3)\) \ | 4 | $1197/28 = 42.8$ | \((1 \ 4 \ 3 \ 2)\) \ | 3 | 31 |
| \((2 \ 3 \ 4)\) \ | 4 | $1197/28 = 42.8$ | \((1 \ 2 \ 3 \ 4)\) \ | 3 | 31 |
| \((1 \ 3)\) \ | 3 | $1153/28 = 41.2$ | \((1 \ 3 \ 2)\) \ | 2 | 30 |
| \((2 \ 4)\) \ | 3 | $1153/28 = 41.2$ | \((1 \ 2 \ 3)\) \ | 2 | 30 |
| \((1 \ 3 \ 4 \ 2)\) \ | 3 | $1096/28 = 39.1$ | \((2 \ 4 \ 3)\) \ | 2 | 30 |
| \((1 \ 2 \ 4 \ 3)\) \ | 3 | $1096/28 = 39.1$ | \((1 \ 4 \ 3)\) \ | 2 | 30 |
| \((1 \ 2 \ 3 \ 4)\) \ | 3 | $1081/28 = 38.6$ | \((2 \ 3 \ 4)\) \ | 2 | 30 |
| \((1 \ 4 \ 3 \ 2)\) \ | 3 | $1081/28 = 38.6$ | \((1 \ 4 \ 2)\) \ | 2 | 30 |
| \((1 \ 4 \ 3)\) \ | 2 | $981/28 = 35.0$ | \((1 \ 3 \ 4)\) \ | 2 | 30 |
| \((1 \ 4 \ 2)\) \ | 2 | $981/28 = 35.0$ | \((1 \ 2 \ 4)\) \ | 2 | 30 |
| \((1 \ 3 \ 4)\) \ | 2 | $981/28 = 35.0$ | \((1 \ 2 \ 3)(4 \ 1)\) \ | 2 | 28 |
| \((1 \ 2 \ 4)\) \ | 2 | $981/28 = 35.0$ | \((1 \ 4 \ 2 \ 3)\) \ | 2 | 28 |
| \((1 \ 3 \ 2 \ 4)\) \ | 2 | $810/28 = 28.9$ | \((1 \ 2 \ 4 \ 3)\) \ | 1 | 23 |
| \((1 \ 4)\) \ | 1 | $682/28 = 24.4$ | \((1 \ 3 \ 4 \ 2)\) \ | 1 | 23 |
| \((1 \ 4 \ 2 \ 3)\) \ | 1 | $625/28 = 22.3$ | \((1 \ 4 \ 2)\) \ | 1 | 23 |
| \((1 \ 3 \ 2 \ 4)\) \ | 1 | $625/28 = 22.3$ | \((1 \ 3 \ 2 \ 4)\) \ | 1 | 23 |
| \((1 \ 4)(2 \ 3)\) \ | 0 | 0 | \((1 \ 3)(2 \ 4)\) \ | 0 | 0 |

Substituting $z = \frac{1}{2} e^s$, differentiating with respect to $s$, and evaluating at $s = 0$, we have the distances to the longest word shown in Table 1. For instance, for the circular generators, the first entry in the final column of $(I - zA_{24})^{-1}$ is the rational polynomial

$$\frac{16z^4}{16z^4 - 16z^2 + 1},$$

and this procedure gives 32: the mean first passage time from the identity to the longest word (the top entry in the right hand column of Table 1).
In Table 1, one can also observe the additional symmetry in the generating set provided by the circular generators (corresponding to inversions on a circular genome as in Egri-Nagy et al. (2014)), as follows. Recalling Proposition 5.1, observe that with the lineal generators \{(1 2), (2 3), (3 4)\}, the normaliser of these is trivial, namely \{e\}. This is in contrast to the case of circular generators, \{(1 2), (2 3), (3 4), (4 1)\}, for which the normaliser is

\[\{e, (1 3), (2 4), (1 3)(2 4), (1 2 3 4), (1 4 3 2), (1 4 2 3), (1 2)(3 4)\}.\]

More elements are conjugate by elements of the (larger) normalizer in the circular case and so we see in Table 1 that there are fewer distinct mean first passage times using the circular generators than the lineal.

### 5.4 Abelian groups

For some general classes of groups we are able to obtain exact combinatorial formulae for the distance. Here we give the result for the abelian group of order $2^k$ with $k$ generators $a_1, \ldots, a_k$ and each element of order two: $a_1^2 = a_2^2 = \cdots = a_k^2 = e$, where $e$ is the identity, and an example to show the method (the full proof will appear in a separate paper).

Consider the case $k = 4$. The elements of the Cayley graph can be divided into 5 sets, according to the lengths of the elements in terms of the generators:

\[
L_0 = \{e\}, \quad L_1 = \{a_1, a_2, a_3, a_4\}, \quad L_2 = \{a_1a_2, a_1a_3, a_1a_4, a_2a_3, a_2a_4, a_3a_4\},
\]

\[
L_3 = \{a_1a_2a_3, a_1a_2a_4, a_1a_3a_4, a_2a_3a_4\}, \quad L_4 = \{a_1a_2a_3a_4\}.
\]

To understand the structure of the Cayley graph it is convenient to work inductively, doubling the size of the matrix every time we add a new generator. Thus for $k = 4$ the rows and columns are ordered as follows:

\[
\begin{array}{cccccccc}
e, & a_1 & a_2 & a_3 & a_4, & a_1a_2, & a_1a_3, & a_1a_4, & a_2a_3, a_1a_2a_3, a_1a_4a_2, a_1a_2a_4, a_1a_3a_4, a_1a_2a_3a_4.
\end{array}
\]

The adjacency matrix is then:
If $A_{[1]}$ is obtained from $A$ by setting all elements in the first row equal to zero and $I_{16}$ is the $16 \times 16$ identity matrix then the functions we need are the entries of the first column of $(I - z\tilde{A})^{-1}$. They come in four sets corresponding to $L_1, L_2, L_3, L_4$, above, respectively:

$$
(1 - 10z^2)z, \quad 2(1 - 4z^2)z^2, \quad 6z^3, \quad 24z^2
$$

If $\zeta(z)$ is one of these functions then using $k = 4$ the distances are given by $\frac{1}{4}\zeta'(\frac{1}{4})$, and are respectively:

$$
15, \quad \frac{56}{3}, \quad \frac{61}{3}, \quad \frac{64}{3}.
$$

The following proof method can be made general but we again restrict it to the case $k = 4$. By considering the sets $L_0, \ldots, L_4$ as equivalence classes we can replace $A$ by a $5 \times 5$ (in general $(k + 1) \times (k + 1)$) matrix $B$ of the following form representing the transition between the $L_i$:

$$
L_0 \Leftrightarrow L_1 \Leftrightarrow L_2 \Leftrightarrow L_3 \Leftrightarrow L_4.
$$

The matrix has an interesting form:

$$
B = \begin{bmatrix}
0 & 4 & 0 & 0 & 0 \\
1 & 0 & 3 & 0 & 0 \\
0 & 2 & 0 & 2 & 0 \\
0 & 0 & 3 & 0 & 1 \\
0 & 0 & 0 & 4 & 0 \\
\end{bmatrix}.
$$
If we carry out the same procedure as we used for $A$, namely take $(I_5 - z \tilde{B})^{-1}$, where $\tilde{B}$ is $B$ with entries in the first row set to zero, we find the same $\zeta(z)$ functions (now distinct) as obtained by using the full incidence matrix.

The general result uses the so-called “group algebra” which quotients out by the equivalence relation, described above, replacing the set of generators in each $L_i$ by their sum. The group’s action is then, essentially, on the whole equivalence class with a new (pseudo) adjacency matrix of the form $B$ above. By careful study of the structure of such matrices, we are able to derive the formula for the distances from an element $L_t, (t = 0, \ldots, k - 1)$ to $L_k$ to be:

$$d_{t,k} = \sum_{s=1}^{k-t} \left( \binom{k-1}{s-1}^{-1} \sum_{r=s}^{k} \binom{k}{r} \right).$$

(5.2)

A simple check confirms that $d_{0,4}, d_{1,4}, d_{2,4},$ and $d_{3,4}$ give the distances $\frac{64}{3}, \frac{61}{3}, \frac{56}{3}, 15$ as calculated above.

### 6 Mason’s rule and zeta functions

The original Mason rule (Mason 1953) writes the formula for $\tilde{M}_{ij}(s)$ in terms of specifically defined paths and cycles (called “loops” in the original because it was written in the context for feedback loops). For completeness we present the essence of the original construction.

For a vertex $i$, a bundle, $B_i$ of $k(B_i)$ cycles is defined as a set of disjoint cycles which do not pass through vertex $i$. Define the weight of any collection, $C$ of edges $w(C) = \prod_{(i,j) \in C} q(i, j)$,

where $q(i, j)$ is the transmittance of $(i, j)$. By Sylvester’s rule for matrix inversion, the required moment generating function matrix has entries:

$$\tilde{M}_{ij}(s) = \left( (I - Q)_{ij} \right)^{-1} = \frac{\det((I - Q)_{ij})}{\det(I - Q_{ij})}$$

where $\left( (I - Q)_{ij} \right)^{-1}_{ij}$ indicates the $(i, j)$ entry of the matrix, and where the numerator of the right hand expression is the cofactor of the $(j, i)$ element of $I - Q_{ij}$ (this is the source of the term “co-factor rule”).

Properties of determinants give the denominator and numerator in the Mason rule as originally expressed:

$$\det(I - Q_{ij}(s)) = 1 + \sum_t (-1)^t \sum_{B_j : k(B_j) = t} w(B_j)$$

Springer
\[
\det \left( (I - Q_{\{j\}})_{ij} \right) = \sum_R w(R_{ij}) \sum_t \left( 1 + (-1)^t \sum_{B_j: k(B_j) = t, B_j \cap R = \emptyset, i \not\in B_j} w(B_j) \right),
\]

where \( R = \{R_{ij}\} \), and \( R_{ij} \) is a direct (non self-intersecting) path from \( i \) to \( j \) (note that in this situation \( B_j \cap R \) is always empty, since \( B_j \) is a set of cycles that avoids vertex \( j \), while \( R \) is the set of direct paths from \( i \) to \( j \)).

Recall that in the case of a Cayley graph \( G \) we start with a regular graph with incidence matrix \( A \) in which \( p_{ij} = \frac{1}{k} \) and \( m_{ij}(s) = m(s) \) and use the generic notation \( Q = zA \).

We feel that it is of independent interest that the quantity

\[
\zeta_G(z) = \det(I - zA)^{-1}
\]

is a type of zeta function. There are several different zeta functions for graphs, and this special type arises in the theory of dynamical systems in which the edge \( i \to j \) is referred to as a shift. It is related to the Bowen-Lanford theory (Bowen and Lanford 1970) as follows. Under a suitable definition of a closed path and the condition that \( A \) is aperiodic it can be shown that

\[
\zeta_G(z) = \prod_\tau (1 - z^{|\tau|}),
\]

where \( \tau \) is a simple circuit.

However, in our case the graphs defined by the matrices \( A_{[i]} \) are absorbing and therefore are not captured by this formula. Even for the full graph, \( A \) can be periodic. For example in the case of \( B_3 \) (the hyperoctahedral group, or signed permutation group, on three letters):

\[
\det(I - zA)^{-1} = (3z - 1)(3z + 1)(z - 1)^3(z + 1)^3(2z - 1)^6(2z + 1)^6,
\]

showing eigenvalue multiplicities and hence periodicities.

Despite these multiplicity issues we can still give a description of the quantities of interest in the style given in the Mason rule. In the same way that zeta functions count circuits, it seems that they are a natural vehicle in this area to capture the cycles and paths of the original theorem. Just as for several types of zeta function, determinants play a key role.

Thus, define

\[
\zeta_G(z)^{-1} = 1 + \sum_j (-1)^j \sum_{B_j: k(B_j) = j} z^{k(B_j)}.
\]
This suggests defining a vertex-specific zeta function for the Cayley graph $G$ by

$$
\zeta_{G,i}(z)^{-1} = \det(I - zA_{[i]}) = 1 + \sum_j (-1)^j \sum_{B_i: k(B_i)=j; j \neq B_i} z^{k(B_i)}.
$$

In a similar way, for an edge $(i, j)$ we have

$$
\zeta_{G,ij}(z)^{-1} = \det((I - zA_{[i]}{_{[ij]}})) = \sum_R z^{|R|} \sum_t \left(1 + (-1)^t \sum_{B_i: k(B_i)=t, B_i \cap R = \emptyset; j \neq B_i} z^{|B_i|}\right),
$$

where $R$ is defined as before.

### 7 Discussion

We collect here some questions, comments, and ideas for further investigation which arose in the gestation of this project, and of which some will be covered in our own future work.

1. It is clear that from a purely algebraic viewpoint the distance we propose houses information about the groups. This is very analogous to the way a zeta function holds information. One could say we have a special type of zeta function theory.

2. Since the Cayley graph depends not just on the group but the choice of generators for the group, so then does the distance. Thus the distances may be useful in separating out different biological processes by considering group generators corresponding to different sets of inversions, or other invertible operations such as translocations.

3. We should make clear that the distance is \textit{linear} in the interarrival means $\mu_{ij}$. One could ask whether linearity is a useful property which may motivate further study.

4. That Cayley graphs are typically not aperiodic has been pointed out, and this is also clear from the circular structures in some examples. By adding additional generators they can be made aperiodic, which should make the steady state (ergodic) properties easier to study.

5. We have made some simple assumptions about the interarrival moment generating function and the transitions, for our examples in Sect. 5. But these can be made more general, for example by allocating different transition values to different types of biological event. An example of this may be a group-based model including both inversions and translocations, or one with different weights for inversions of different numbers of regions, as in Bhatia et al. (2019).

6. It is clear from examples, and the Abelian group example, that derivation of general formulae such as (5.2) may be achieved by a reduction using conjugacy, and a dummy Cayley graph with a pseudo-incidence structure such as in the matrix $B$ in Sect. 5.4.
(7) The mean first passage time (MFPT) distance provides a partial order on the elements of the group, much like the minimal distance or the maximum likelihood estimate (MLE) distance (although this may only exist for a subset of the group). The MLE distance has been shown to reverse the ordering on some group elements, relative to the minimal distance (Serdoz et al. 2017), which has a concrete implication for phylogenetic reconstruction using algorithms like Neighbour-Joining (Saitou and Nei 1987). It would be very interesting to understand whether any reversals of the minimal order under the MFPT are the same as those for the MLE.

(8) The MFPT distance seems to discriminate better among shorter distances, whereas longer distances tend to be closer together, which could be seen as a drawback and may suggest some asymptotic behaviour as group order increases (something for further investigation). One positive side of this preference for short distances is that in general one would expect biological samples to be relatively closely related, among the space of all possible genomes, that is, from a relatively contained region of the Cayley graph. In other words, the MFPT distance may be effective precisely where it is more likely to be used.

Another point in this theme is that many measures that go beyond minimal distance also run into limitations with longer distances. For instance the Jukes–Cantor correction for single nucleotide polymorphism distances is only applicable if the minimal distance is less than $3/4$ of the maximal possible distance (Jukes and Cantor 1969), and the likelihood approaches do not have maxima for pairs of elements that are far apart on the Cayley graph (Serdoz et al. 2017) (accounting for over 40% of pairs of genomes, according to Serdoz et al. (2017); Terauds and Sumner (2019)).

(9) Finally, the main challenge for using the MFPT approach to distance is computability. In principal the key computational obstacle is simple: the method requires finding a particular element of the inverse of a matrix. The naïve implementation that we have used here involves a matrix that has size $n! \times n!$, for genomes with $n$ regions. However, these matrices are sparse (only $\sim n$ non-zero terms in each row), and there may be ways to significantly improve the efficiency.

Acknowledgements ARF acknowledges the support of the Australian Research Council via Discovery Project DP180102215.

Appendix A. Evolutionary models

Table 2 shows a range of group-based models that this approach can be applied to. Each corresponds to a particular group and generating set.
### Table 2  Models and corresponding group features

| Name          | Model                                      | Group and Generators                                                                 |
|---------------|--------------------------------------------|--------------------------------------------------------------------------------------|
| **Unsigned inversions**  |                                           |                                                                                      |
| $U_2^{[L]}$   | 2-Inversions on a lineal chromosome        | The group is the symmetric group $S_n$, or Coxeter group of type $A_{n-1}$           |
| $U_2$         | 2-Inversions on a circular chromosome      | Generators (allowable inversions) are the Coxeter generators $(1 \; 2), (2 \; 3),...,(n - 1 \; n)$ |
| $U_3$         | 2- and 3-inversions, circular chromosome   | Generators are $(i \; i + 1)$ and $(i \; i + 2) \mod n$                           |
| $U_4$         | 2-, 3- and 4-inversions, circular chromosome | Generators are $(i \; i + 1), (i \; i + 2),$ and $(i \; i + 3) (i + 1 \; i + 2) \mod n$ |
| $U_k$         | etc                                        |                                                                                      |
| $U_{full}$    | All inversions, circular chromosome        | This is the dominant model in inversion distance literature                           |
| **Signed inversions**  |                                           |                                                                                      |
| $SCox$        | The Coxeter model with generators $t, s_1, \ldots, s_{n-1}$, where $t$ swaps $1$ and $-1$ and $s_i = (i \; i + 1)$. Not biologically sensible, but of interest algebraically |
| $S_2$         | Signed 1- and 2-inversions                 | Generators $(i \; -i), (1 \; -2)$ etc. With signed inversions we assume $\pi(-i) = -\pi(i)$, so $(1 \; -2)$ denotes $(1 \; -2)(-1 \; 2)$ |
| $S_k$         | Signed inversions up to $k$ regions        |                                                                                      |
| **Combined inversion and translocation model**  |                                           | For any of the above, include translocations that shift $i$ past $j$. Write $t_{i, j} = (i \; i + 1 \ldots j)$ |

**Appendix B. Cayley graphs for $S_4$**

The Cayley graphs of $S_4$ with standard and with circular generators are shown for reference in Figs. 4 and 5 respectively.
Fig. 4 Cayley graph of $S_4$ with standard Coxeter generators. Edges are coded dashed blue for multiplication on the right by $(1\,2)$, dotted red for $(2\,3)$, and black for $(3\,4)$ (group action is also on the right) (color figure online)
Fig. 5 Cayley graph of $S_4$ with circular generators. Edges are coded dashed blue for multiplication on the right by $[1,2]$, dotted red for $[2,3]$, black for $[3,4]$, and dash-dotted green for $[1,4]$ (group action is also on the right) (color figure online)

References

Bafna V, Pevzner PA (1993) Genome rearrangements and sorting by reversals. In: 34th annual symposium on foundations of computer science, 1993. Proceedings. IEEE, pp 148–157
Bergeron A, Mixtacki J, Stoye J (2006) A unifying view of genome rearrangements. In: Bücher P, Moret BME (eds) Algorithms in bioinformatics. WABI 2006. Lecture notes in computer science, vol 4175. Springer, Berlin, Heidelberg, pp 163–173
Bhatia S, Egri-Nagy A, Serdoz S, Praeger C, Gebhardt V, Francis A (2019) A flexible framework for determining weighted genome rearrangement distance. preprint
Bowen R, Lanford III OE (1970) Zeta functions of restrictions of the shift transformation. In: Global analysis, proceedings of symposia in pure mathematics. Volume XIV, Berkeley, California, 1968
Butler RW, Huzurbazar AV (1997) Stochastic network models for survival analysis. J Am Stat Assoc 92(437):246–257
Clark C, Egri-Nagy A, Francis A, Gebhardt V (2019) Bacterial phylogeny in the Cayley graph. Discrete Math Algorithms Appl 11(05):1950059
Dalevi D, Eriksen N (2008) Expected gene-order distances and model selection in bacteria. Bioinformatics 24(11):1332–1338
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
Francis AR (2014) An algebraic view of bacterial genome evolution. J Math Biol 69(6–7):1693–1718
Howard RA (1960) Dynamic programming and Markov processes. The Technology Press of M.I.T, Cambridge
Howard RA, Matheson JE (1971) Risk-sensitive Markov decision processes. Manag Sci 18:356–369
Humphreys JE (1990) Reflection groups and Coxeter groups, vol 29. Cambridge University Press, Cambridge
Jukes TH, Cantor CR (1969) Evolution of protein molecules. Mamm Protein Metab 3(21):132
Mason SJ (1953) Feedback theory-some properties of signal flow graphs. Proc IRE 41(9):1144–1156
Pyke R (1961) Markov renewal processes: definitions and preliminary properties. Ann Math Stat 32:1231–1242
Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425
Sankoff D, Leduc G, Antoine N, Paquin B, Lang BF, Cedergren R (1992) Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. Proc Natl Acad Sci 89(14):6575–6579
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
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 (2019) Maximum likelihood estimates of rearrangement distance: implementing a representation-theoretic approach. Bull Math Biol 81(2):535–567
Wang L-S, Warnow T (2001) Estimating true evolutionary distances between genomes. In: Proceedings of the thirty-third annual ACM symposium on theory of computing. ACM, pp 637–646
Yancopoulos S, Attie O, Friedberg R (2005) Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21(16):3340–3346

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.