

## The ‘Butterfly effect’ in Cayley graphs with applications to genomics

Vincent Moulton · Mike Steel

Received: 15 March 2011 / Revised: 2 November 2011  
© Springer-Verlag 2011

**Abstract** Suppose a finite set  $X$  is repeatedly transformed by a sequence of permutations of a certain type acting on an initial element  $x$  to produce a final state  $y$ . For example, in genomics applications,  $X$  could be a set of genomes and the permutations certain genome ‘rearrangements’ or, in group theory,  $X$  could be the set of configurations of a Rubik’s cube and the permutations certain specified moves. We investigate how ‘different’ the resulting state  $y'$  to  $y$  can be if a slight change is made to the sequence, either by deleting one permutation, or replacing it with another. Here the ‘difference’ between  $y$  and  $y'$  might be measured by the minimum number of permutations of the permitted type required to transform  $y$  to  $y'$ , or by some other metric. We discuss this first in the general setting of sensitivity to perturbation of walks in Cayley graphs of groups with a specified set of generators. We then investigate some permutation groups and generators arising in computational genomics, and the statistical implications of the findings.

**Keywords** Evolutionary distance · Permutation · Metric · Group action · Genome rearrangements

**Mathematics Subject Classification (2000)** 20B05 · 92B10 · 92D15

---

VM thanks the Royal Society for supporting his visit to University of Canterbury, where most of this work was undertaken. MS thanks the Royal Society of New Zealand under its James Cook Fellowship scheme.

---

V. Moulton (✉)  
School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK  
e-mail: vincent.moulton@cmp.uea.ac.uk

M. Steel  
Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand  
e-mail: m.steel@math.canterbury.ac.nz

## 1 Introduction and background

Mathematical techniques play an important role in developing methods to infer evolutionary relationships between species, based on differences between their genomes [Fertin et al. \(2009\)](#), [Pevzner \(2000\)](#), [Semple and Steel \(2003\)](#), [Wang and Warnow \(2005\)](#). For the purposes of this paper a *genome* is simply an ordered sequence of objects—usually taken from the DNA alphabet or a collection of genes—which may occur with or without repetition, and with or without an orientation (+, −).

In evolutionary genomics, two genomes are frequently compared by the minimum number of ‘rearrangements’ (of various types, such as transpositions or reversals) required to transform one genome into another ([Fertin et al. 2009](#)). This minimum number is then used as an estimate of the actual number of events and thereby the ‘evolutionary distance’ between the species involved.

Since both the precise number and the actual rearrangement events that occurred in the evolution of the two genomes from a common ancestor are unknown, it is pertinent to have some idea of how sensitive this distance estimate might be to the sequence of events (not just the number) that really took place ([Sinha and Meller 2008](#)).

This question has important implications for the accurate inference of evolutionary relationships between species from their genomes, and we discuss some of these further in Sect. 3. However, we begin by framing the type of mathematical questions that we will be considering in a general algebraic context.

Let  $G$  be a finite group, whose identity element we write as  $1_G$ , and let  $S \subset G$  be a subset of generators, that is *symmetric* (i.e. closed under inverses, so  $x \in S \Rightarrow x^{-1} \in S$ ) and with  $1_G \notin S$ . In addition, let  $\Gamma = \text{Cay}(G, S)$  be the associated (undirected) *Cayley graph*, with vertex set  $G$  and an edge connecting  $g$  and  $g'$  if there exists  $s \in S$  with  $g' = gs$  (unless otherwise stated, we use the convention of multiplying group elements from left to right). For any two elements  $g, g' \in G$ , the distance  $d_S(g, g')$  in  $\text{Cay}(G, S)$  is the minimum value of  $k$  for which there exist elements  $s_1, \dots, s_k$  of  $S$  so that  $g' = gs_1 \cdots s_k$  (for  $g = g'$ , we set  $d_S(g, g') = 0$ ). Note that  $d_S$  is a metric<sup>1</sup> (in fact it is just the usual shortest-path metric on  $\Gamma$ ), in particular,  $d_S(g, g') = d_S(g', g)$ , since  $S$  is symmetric.

In this paper, our focus is on the following two quantities:

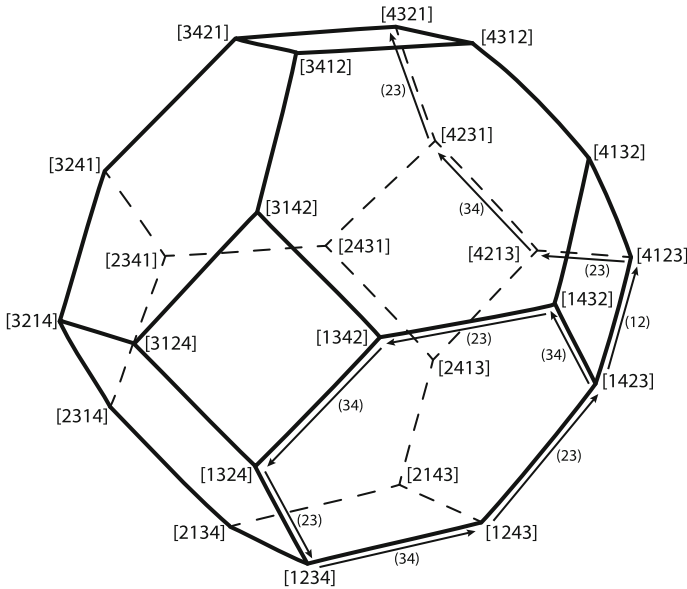
$$\lambda_1(G, S) := \max_{g \in G, s \in S} \{d_S(sg, g)\},$$

and

$$\lambda_2(G, S) := \max_{g \in G, s, s' \in S} \{d_S(sg, s'g)\}.$$

One way to view these quantities is via the following result which is easily proved.

<sup>1</sup> Recall that a metric is a non-negative real-valued function  $d$  defined on  $X \times X$  which satisfies the following properties for all values  $x, y, z \in X$ :  $d(x, y) = d(y, x)$ ,  $d(x, y) = 0 \Leftrightarrow x = y$ , and  $d(x, z) \leq d(x, y) + d(y, z)$ .



**Fig. 1** The Cayley graph  $Cay(G, S)$  for  $G = \Sigma_4$  (the permutation group on  $\{1, 2, 3, 4\}$ ) and the set of transpositions  $S = \{(12), (23), (34)\}$ . Substituting just one element—namely (34) for (12)—in the product corresponding to the walk in the lower front face (which starts and returns to the lower-most point [1234]) results in a walk that ends at a point ([4321], top) that is very distant (under  $d_S$ ) from the end-point of the original walk. In fact, the two end-points are at maximal distance in this example

**Lemma 1** *Let  $S$  be a symmetric set of generators for a finite group  $G$ . Then:*

- (a)  $\lambda_1(G, S)$  is the maximum value of  $d_S(g, g')$  between any pair of elements  $g$  and  $g'$  of  $G$  for which  $g = h_1 h_2 \cdots h_k$ , and  $g' = h'_1 h'_2 \cdots h'_k$ , where  $h'_i = h_i \in S$  for all but at most one value (say  $j$ ) for  $i$ , and  $h'_j = 1_G, h_j \in S$ .
- (b)  $\lambda_2(G, S)$  is the maximum value of  $d_S(g, g')$  between any pair of elements  $g$  and  $g'$  of  $G$  for which  $g = h_1 h_2 \cdots h_k$  and  $g' = h'_1 h'_2 \cdots h'_k$  where  $h'_i = h_i \in S$  for all but at most one value (say  $j$ ) for  $i$ , and  $h_j, h'_j \in S, h'_j \neq h_j$ .

Thus,  $\lambda_1(G, S)$  tells us how much (under  $d_S$ ) a product of generators can change if we omit one generator at some position in the product sequence, whilst  $\lambda_2(G, S)$  tells us how much (again under  $d_S$ ) a product of generators in  $S$  can change if we substitute one generator for another at some position in the product sequence (see Fig. 1 for an example where  $\lambda_2(G, S) = 6$ ; this follows by Lemma 1(b) since the maximal distance between any pair of vertices in  $Cay(G, S)$  is 6 and, as indicated in the figure caption, the substitution of the third generator (34) in the product (34) (23) (34) (23) (34) (23) with the generator (12) converts a closed cycle in  $Cay(G, S)$  into a path from [1234] to the maximally-distant vertex [4321]).

As such,  $\lambda_m$  ( $m = 1, 2$ ) is a measure of the ‘sensitivity’ of walks in the Cayley graph to a switch in, or deletion of, a generator at some point. Moreover, if  $G$  acts transitively and freely<sup>2</sup> on any arbitrary finite set  $X$  then  $\lambda_m$  provides a corresponding measure of

<sup>2</sup>  $G$  acts transitively on  $X$  if for any pair  $x, y \in X$  there exists  $g \in G$  with  $xg = y$ ; the action is free if  $xg = xh \Rightarrow g = h$ , for all  $g, h \in G$  and  $x \in X$ .

sensitivity of this action to a switch in or deletion of a generator (since a transitive, free action of  $G$  on  $X$  is isomorphic to the action of  $G$  on itself by right multiplication). Actions with large  $\lambda_m$  values can thus be viewed as exhibiting a group-theoretical analogue of the ‘butterfly effect’ in dynamical systems theory (Hilborn 2004), where a small change in the initial state of a system can result in large differences to later states (cf. Holmgren (1994, Part IV) where the concept of sensitivity in dynamical systems with a discrete state space is reviewed).

In the genomics applications that we consider,  $X$  is a set of DNA sequences or genomes, elements of  $G$  correspond to evolutionary events such as site substitution or genome rearrangement, and the metric  $d_S$  to some ‘evolutionary distance’ between DNA sequences or genomes, respectively. In this context, we are interested in understanding whether small changes in the sequence of evolutionary events acting on some initial DNA sequence or genome can lead to large differences in the resulting DNA sequence or genome relative to  $d_S$  (cf. Holmgren (1994, Sect. 11.1) for related concepts arising in symbolic dynamics). After presenting some general results concerning  $\lambda_m$ , in Sect. 2 we discuss some applications arising for various choices of  $G$  and  $S$ . These include the Klein four group, which arises in evolutionary models of DNA sequence evolution, and the permutation group, which typically appears when studying rearrangement distances between genomes. In Sect. 3 we outline some statistical implications of our results, followed by some brief concluding comments.

One can imagine many other settings besides genomics where similar questions arise—for example, in a sequence of moves that should unscramble the Rubik’s cube from a given position (Kunkle and Cooperman 2009), what will be the consequences (in terms of the number of moves required) for completing the unscrambling if a mistake is made at some point (or one move is forgotten)? In addition, related questions arise in the study of ‘automatic’ groups, where the group under consideration is typically infinite (Eppstein 1992).

In genomics, there has been a great deal of work over the last two decades on the combinatorial and algorithmic aspects of genome rearrangements. Particular highlights have included Hannenhalli and Pevner’s celebrated polynomial-time algorithms for calculating the minimal reversal distance between signed genomes (Hannenhalli and Pevzner 1999), and a more recent and general approach based on the combinatorics of the ‘double cut and join’ operation (Bergeron et al. 2009) [other related results are also described in Bafna and Pevzner (1996), Kececioğlu and Sankoff (1995) and Pevzner (2000)]. However, to the best of our knowledge, there has been no attempt to define or study two quantities ( $\lambda_1, \lambda_2$ ) that we consider in this paper, and we view our results as providing some modest initial results, motivated by statistical considerations that are outlined in Sect. 3.

### 1.1 Algebraic background and general inequalities

We first make some basic observations about Cayley graphs and the metric  $d_S$  [further background on basic group theory, Cayley graphs, and group actions can be found in Rotman (1995)]. It is well known that  $\Gamma$  is a connected regular graph of degree equal to the cardinality of  $S$  and that  $\Gamma$  is also vertex-transitive [see, for example, Kostantinova

(2008, Proposition 1)]. Consider the function  $l_S : G \rightarrow \{0, 1, 2, 3 \dots |G|\}$ , where,  $l_S(1_G) = 0$  and, for each  $g \in G - \{1_G\}$ ,  $l_S(g)$  is the smallest number  $l$  of elements  $s_1, \dots, s_l$  from  $S$  for which we can write  $g = s_1 \cdots s_l$ . In terms of the Cayley graph,  $l_S(g)$  is the shortest path from  $g$  to the identity element. The function  $l_S$  clearly satisfies the subadditivity property that, for all  $g, g' \in G$ :

$$l_S(gg') \leq l_S(g) + l_S(g').$$

In addition,

$$l_S(g^{-1}) = l_S(g),$$

and

$$l_S(g) = 1 \Leftrightarrow g \in S, \quad l_S(g) = 0 \Leftrightarrow g = 1_G.$$

Note that  $l_S(gg')$  is generally not equal to  $l_S(g'g)$ . The metric  $d_S$ , described in the previous section, is related to  $l_S$  as follows:

$$d_S(g, g') = l_S(g^{-1}g').$$

Consequently, by definition:

$$\lambda_1(G, S) = \max_{g \in G, s \in S} \{l_S(g^{-1}sg)\}, \tag{1}$$

and

$$\lambda_2(G, S) = \max_{g \in G, s, s' \in S} \{l_S(g^{-1}ss'g)\}. \tag{2}$$

Let  $l_S(G) = \max\{l_S(g) : g \in G\}$ , which is the diameter of  $Cay(G, S)$ , that is, the maximum length shortest path connecting any two elements of  $G$ . Clearly,  $\lambda_1(G, S) \leq l_S(G)$  and  $\lambda_2(G, S) \leq l_S(G)$ . Moreover:

$$\lambda_2(G, S) \leq 2 \cdot \lambda_1(G, S), \tag{3}$$

since, for any  $g \in G$  and  $s, s' \in S$ , we have:

$$d_S(sg, s'g) \leq d_S(sg, g) + d_S(g, s'g).$$

A partial converse to Inequality (3) is provided by the following:

$$\lambda_1(G, S) \leq \lambda_2(G, S) + \lambda'_1(G, S), \tag{4}$$

where  $\lambda'_1(G, S) = \max_{g \in G} \min_{s \in S} \{l_S(g^{-1}sg)\}$ . To verify (4), select a pair  $g$  from  $G$  and  $s$  from  $S$  so that  $l_S(g^{-1}sg) = \lambda_1(G, S)$ . Then:

$$\lambda_1(G, S) = d_S(sg, g) \leq d_S(sg, s_1g) + d_S(s_1g, g),$$

where  $s_1$  is an element  $s'$  (possibly equal to  $s$ ) in  $S$  that minimizes  $l_S(g^{-1}s'g)$ . Now,  $d_S(sg, s_1g) \leq \lambda_2(G, S)$  (even if  $s' = s$ ) and  $d_S(s_1g, g) \leq \lambda'_1(G, S)$ , and so we obtain (4).

Note also that if  $G$  is Abelian, then  $\lambda_1(G, S) = 1$ , and  $\lambda_2(G, S) \leq 2$  for any symmetric set  $S$  of generators. Moreover, for the Abelian 2-group  $G = \mathbb{Z}_2^n$  and with the symmetric set  $S$  of generators consisting of all  $n$  elements with the identity at all but one position, we have  $l_S(G) = n$  and  $\lambda_1(G, S) = 1$ . This shows that the ratio  $\lambda_1(G, S)/l_S(G)$  can be made as close to zero as desired by selecting an appropriate pair  $(G, S)$ . Our next result generalizes this observation further.

**Lemma 2** *Let  $G_1, G_2, \dots, G_k$  be finite groups, and let  $S_i$  be a symmetric set of generators of  $G_i$  for  $i = 1, \dots, k$ . Consider the direct product  $G = G_1 \times G_2 \times \dots \times G_k$  along with the symmetric set of generators  $S$  of  $G$  consisting of all possible  $k$ -tuples which consist of the identity element of  $G_i$  at all but one co-ordinate  $i$ , where it takes some value in  $S_i$ . Then:*

- (i)  $\lambda_1(G, S) \leq \max_{1 \leq i \leq k} \{l_{S_i}(G_i)\}$ , and
- (ii)  $l_S(G) = \sum_{i=1}^k l_{S_i}(G_i)$ .

*Proof* For Part (i), let  $\lambda_1(G, S) = l_S(g^{-1}sg)$ , where  $s \in S$  is a non-identity element at some co-ordinate  $v$ . Notice that  $(g^{-1}sg)_j = 1_{G_j}$  for all  $j \neq v$ . Moreover,  $(g^{-1}sg)_v = s_1 \dots s_l$  where  $l \leq l_{S_v}(G_v)$ . Thus  $l_S(g^{-1}sg) \leq l_{S_v}(G_v)$ , as claimed.

For Part (ii), the inequality  $l_S(G) \leq \sum_{i=1}^k l_{S_i}(G_i)$  is clear; to establish the reverse inequality, for any  $1 \leq i \leq k$  let  $g_i$  be an element of  $G_i$  with  $l_{S_i}(g_i) = l_{S_i}(G_i)$ , and  $g = (g_1, \dots, g_k) \in G$ . Then  $l_S(g) = \sum_{i=1}^k l_{S_i}(G_i)$ , and so  $l_S(G) \geq \sum_{i=1}^k l_{S_i}(G_i)$ . □

We now consider how  $\lambda_m$  ( $m = 1, 2$ ) behaves under group homomorphisms. Suppose  $H$  is the homomorphic image of a group  $G$  under a map  $p$ . Let  $N = Ker(p)$  be the kernel of  $p$ , which is a normal subgroup of  $G$ , and with  $H \cong G/N$ . Thus we have a short exact sequence:

$$1_G \rightarrow N \rightarrow G \xrightarrow{p} H \rightarrow 1_G. \tag{5}$$

Let  $S$  be a symmetric set of generators of  $G$ . Then  $S_H = \{p(s) : s \in S - N\}$  is a symmetric set of generators of  $H$ .

**Lemma 3** *For  $m = 1, 2$ ,  $\lambda_m(H, S_H) \leq \lambda_m(G, S)$ .*

*Proof* First suppose that  $m = 1$ . For  $x \in S_H$  and  $h \in H$ , consider  $h^{-1}xh$ . There exist elements  $g \in G$  and  $s \in S - N$  for which  $f(g) = h$  and  $f(s) = x$ . Now the element  $g^{-1}sg \in G$  can be written as a product of at most  $l = \lambda_1(G, S)$  elements of  $S$ , that is  $g^{-1}sg = s_1s_2 \dots s_k$  for  $k \leq l$ . Applying  $p$  to both sides of this equation gives:  $h^{-1}xh = p(s_1)p(s_2) \dots p(s_k)$ . Notice that some of the elements on right may equal the identity element of  $H$  (since  $p(s_i) = 1_H \Leftrightarrow s_i \in N$ ), but they are elements of  $S_H$  otherwise. Thus  $l_{S_H}(h^{-1}xh) \leq l$ . Since this holds for all such elements  $h, x$ , Eq. (1)

shows that  $\lambda_1(H, S_H) \leq \lambda_1(G, S)$ . The corresponding result for  $m = 2$  follows by an analogous argument.  $\square$

To obtain a lower bound for  $\lambda_m(G, S)$  suppose that the short exact sequence (5) is a *split extension*, i.e. there is a homomorphism  $\iota : H \rightarrow G$  so that  $p\iota$  is the identity map on  $H$ , which [by the splitting lemma (Rotman 1995)] is equivalent to the condition that  $G$  is the semidirect product<sup>3</sup> of  $N$  with a subgroup  $H'$  isomorphic to  $H$ . In this case we have the following bounds.

**Proposition 1** *Suppose a finite group  $G$  is a semidirect product of subgroups  $N$  (normal) and  $H$ . Let  $S_N, S_H$  be symmetric generator sets for  $N$  and  $H$  respectively, and let  $S = S_N \cup S_H$  which is a symmetric generator set for  $G$ . Then:*

$$\lambda_1(H, S_H) \leq \lambda_1(G, S) \leq \lambda_1(H, S_H) + l_{S_N}(N).$$

In particular, by (3),  $\lambda_2(G, S) \leq 2\lambda_1(H, S_H) + 2l_{S_N}(N)$ .

*Proof* The lower bound on  $\lambda_1(G, S)$  follows from Lemma 3. For the upper bound we must show that for all  $s \in S$  and  $g \in G$ ,  $d_S(sg, g) \leq \lambda_1(H, S_H) + l_{S_N}(N)$  holds. We consider two cases: (i)  $s \in N$ , and (ii)  $s \in H$ . In Case (i), note that the conjugate element  $g^{-1}sg$  is also an element of  $N$ ; in this case we have the tighter bound  $d_S(sg, g) \leq l_{S_N}(N)$ . In Case (ii), write  $g = hn$  where  $n \in N$  and  $h \in H$ . Consider the word

$$w = g^{-1}sg = n^{-1}h^{-1}shn.$$

Since  $N$  is normal we have  $n^{-1}(h^{-1}sh) = (h^{-1}sh)n'$  for some element  $n' \in N$ . Thus  $w = h^{-1}shn'n$ . Write  $w = w_1w_2$  where  $w_1 = h^{-1}sh \in H$  and  $w_2 = n'n \in N$ . We can select  $w_2$  to be a product of terms of  $S_N$  of length at most  $l_{S_N}(N)$  and, by Inequality (3), we can select  $w_1$  to be a product of terms of  $S_H$  of length at most  $\lambda_1(H, S_H)$ . Thus  $w$  can be written as a product of, at most,  $\lambda_1(H, S_H) + l_{S_N}(N)$  elements of  $S$ .  $\square$

Finally, as it can sometimes be useful to consider metrics on a group other than those arising from some Cayley graph, given an arbitrary metric  $d$  and a finite group  $G$  with symmetric generator set  $S$ , in analogy with (1) and (2) we define:

$$\lambda_1(G, S, d) := \max_{g \in G, s \in S} \{d(sg, g)\} \quad \text{and} \quad \lambda_2(G, S, d) := \max_{g \in G, s, s' \in S} \{d(sg, s'g)\}.$$

In particular,  $\lambda_m(G, S) = \lambda_m(G, S, d_S)$  and  $\lambda_m(G, S, d) \leq \max_{g, g' \in G} \{d(g, g')\}$ ,  $m = 1, 2$ . Moreover, the following analogue of Inequality (3) for an arbitrary metric  $d$  on  $G$  is easily seen to hold:

<sup>3</sup> The statement that  $G$  is the semidirect product of a normal subgroup  $N$  of  $G$  with another subgroup  $H'$  of  $G$  means that each element  $g$  of  $G$  can be written uniquely as the product  $nh$  (and as  $hn$ ) where  $n \in N$  and  $h \in H$ .

$$\lambda_2(G, S, d) \leq 2 \cdot \lambda_1(G, S, d). \quad (6)$$

Note that, although the quantities  $\lambda_m(G, S)$  and  $\lambda_m(G, S, d)$  need not be directly related to one another, in certain circumstances, they are. For example, if  $d$  has the property that  $d(g, gs) \leq c$  for some constant  $c$  it is an easy exercise to show that  $\lambda_m(G, S, d) \leq c \cdot \lambda_m(G, S)$ , for  $m = 1, 2$ .

## 2 Permutation groups and genomic applications

In this section, we describe some applications of the above observations in computational genomics.

### 2.1 Evolution of DNA sequences by site-wise permutations

We first describe a direct application that is relevant to the evolution of a DNA sequence under a simple model of site substitution (Kimura's 3ST model) (Kimura 1981). Consider the four-letter DNA alphabet  $\mathcal{A} = \{A, C, G, T\}$  and the Klein four-group  $K = \mathbb{Z}_2 \times \mathbb{Z}_2$  with an action on  $\mathcal{A}$  in which the three non-zero elements of  $K$  correspond to 'transitions' ( $A \leftrightarrow G, C \leftrightarrow T$ ) and the two types of 'transversions' ( $A \leftrightarrow C, G \leftrightarrow T$ ; and  $A \leftrightarrow T, G \leftrightarrow C$ ). This representation of the Kimura 3ST model was first described and exploited by Evans and Speed (1993).

For  $g \in K$  and  $x \in \mathcal{A}$ , let  $xg$  denote the element of  $\mathcal{A}$  obtained by the (right) action of  $g$  on  $x$  (the identity element fixes each element of  $\mathcal{A}$ ). The resulting component-wise action of  $K^n$  on  $\mathcal{A}^n$ , defined by:  $(x_1, \dots, x_n)(g_1, \dots, g_n) = (x_1g_1, \dots, x_ng_n)$ , can be regarded as the set of all changes that can occur to a DNA sequence over a period of time under site substitutions.

Now, under any continuous-time Markovian process these change events ('site substitutions') occur just one at a time and so a natural generating set of  $K^n$  is the set  $S_n$  of all elements of  $K^n$  that consist of  $1_K$  at all but one co-ordinate. Moreover, since the action of  $K^n$  on  $\mathcal{A}^n$  is transitive and free (and so is isomorphic to the action of  $K^n$  on itself by right multiplication),  $\lambda_m(K^n, S_n)$  measures the impact of ignoring (for  $m = 1$ ) or replacing (for  $m = 2$ ) one substitution in a chain of such events over time. As  $K^n$  is Abelian, one has  $\lambda_1(K^n, S_n) = 1$  and  $\lambda_2(K^n, S_n) = 2$ , which implies that this impact is minor, and, more significantly, is independent of  $n$ ; this has important statistical implications which we will describe further in Sect. 3.

For a related example, consider the ordered sequence of distinct genes  $(g_1, g_2, \dots, g_n)$  partitioned into regions  $R_1, R_2, \dots, R_k$  so that genomic rearrangements occur within each region, but not between regions (e.g.  $R_i$  for  $1 \leq i \leq k$  might refer to  $k$  different chromosomes). This situation can be modelled by the setting of Lemma 2 in which  $G_i$  is a permutation group on the genes within  $R_i$ , and  $S_i$  is set of elementary gene order rearrangement events that generates  $G_i$  (we discuss some examples below). In this case, Lemmas 2(i) and (3) provide a bound on  $\lambda_1$  and  $\lambda_2$  that is independent of the number of regions  $k$  (specifically, if  $l_{S_i}(G_i) \leq r$  then  $\lambda_1 \leq r$  and  $\lambda_2 \leq 2r$ ).

## 2.2 Genome rearrangements and permutations

We turn now to the calculation of  $\lambda_m(\Sigma_n, S)$  for the permutation group  $\Sigma_n$  on  $n!$  elements and various sets  $S$  of generators. This group commonly arises when studying (unsigned) genome rearrangements (Kostantinova 2008). Our main interest is to determine, for each instance of  $S$ , whether there is a constant  $C$  (independent of  $n$ ) for which  $\lambda_m(\Sigma_n, S) \leq C$ , for  $m = 1, 2$ .

A permutation  $g$  on the set  $[n] := \{1, 2, \dots, n\}$  is a bijective mapping from  $[n]$  to itself. We will also write  $g$  as  $g = [g_1, g_2, \dots, g_n]$  where  $g_i = g(i)$  is the image of the map  $g$  for  $i \in [n]$ . Note that, following the usual convention, the product  $gg'$  of two permutations  $g, g' \in \Sigma_n$  will be considered as the composition of the functions  $g$  and  $g'$ . In particular,  $gg'(i) = g(g'(i))$  for all  $i \in [n]$ .

When studying genomes, each entry  $g_i$  of a permutation  $g$  corresponds to a gene and the full list  $[g_1, g_2, \dots, g_n]$  to a genome. Multiplying  $g$  by a permutation leads to a rearrangement of the genome. For example, multiplying by a *transposition*  $t_{i,j}$  interchanges the values at positions  $i$  and  $j$  of  $g$ , i.e.  $[\dots, g_i, \dots, g_j, \dots]t_{i,j} = [\dots, g_j, \dots, g_i, \dots]$ , and multiplying by a *reversal*  $r_{i,j}$  reverses the segment  $[g_i, g_j]$ ,  $1 \leq i < j \leq n$ , of  $g$ , i.e.

$$[\dots, g_i, g_{i+1}, \dots, g_{j-1}, g_j, \dots]r_{i,j} = [\dots, g_j, g_{j-1}, \dots, g_{i+1}, g_i, \dots].$$

Such rearrangements are widely observed and studied in molecular biology (Fertin et al. 2009).

In genomics applications, we are often interested in defining some distance between genomes. One distance that is commonly used in the context of permutations is the *breakpoint* distance (Setubal and Meidanis 1997, 7.3). For  $g, g' \in \Sigma_n$ ,  $d_{BP}(g, g')$  is defined as the number of pairs of elements that are adjacent in the list  $[0, g_1, g_2, \dots, g_n, n + 1]$ , but not in the list  $[0, g'_1, g'_2, \dots, g'_n, n + 1]$ . For example, if  $g = [1, 2, 3, 4, 5]$ ,  $g' = [1, 4, 3, 2, 5] \in \Sigma_5$ , we have  $d_{BP}(g, g') = 2$ . It is clear that  $\max\{d_{BP}(g, g') : g, g' \in \Sigma_n\} = n + 1$ .

Alternatively, one can consider the *rearrangement distance* between two genomes, i.e. the minimal number of operations of a certain type (such as transpositions or reversals) that can be applied to one of the genomes to obtain the other (Fertin et al. 2009). In terms of Cayley graphs, this distance can be conveniently expressed for transpositions and reversals as follows. For  $t_{i,j}$  and  $r_{i,j}$  as defined above, let

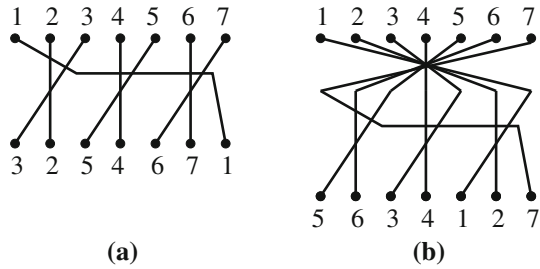
$$\begin{aligned} T_n &:= \{t_{i,j} : 1 \leq i < j \leq n\}, \\ C_n &:= \{t_{i,i+1} : 1 \leq i \leq n - 1\}, \end{aligned}$$

(the *Coxeter* generators), and

$$R_n := \{r_{i,j} : 1 \leq i < j \leq n\}.$$

Note that each one of these sets generates  $\Sigma_n$  (Kostantinova 2008) and that each of them is symmetric, since each generator is its own inverse. The metric  $d_S$ , for  $S = T_n, C_n$  or  $R_n$ , is precisely the rearrangement distance (relative to  $S$ ).

**Fig. 2** **a** A diagrammatic representation of the element  $g = [3, 2, 5, 4, 6, 7, 1]$  in  $\Sigma_7$ , defined in the proof of Theorem 1 (iii). **b** The product  $r_{1,7}g = [5, 6, 3, 4, 1, 2, 7]$ . Note that  $d_{BP}(r_{1,7}g, g) = 8$



The diameters of  $Cay(\Sigma_n, T_n)$  and  $Cay(\Sigma_n, R_n)$  are both  $n - 1$ , and the diameter of  $Cay(\Sigma_n, C_n)$  is  $\binom{n}{2}$  (Kostantinova 2008).

Regarding the quantities  $\lambda_m(\Sigma_n, S)$  for  $S = T_n, C_n, R_n$  we have the following result:

**Theorem 1** For  $n \geq 7$  the following hold:

- (i)  $\lambda_1(\Sigma_n, T_n) = 1$  and  $\lambda_2(\Sigma_n, T_n) = 2$ .
- (ii)  $\lambda_1(\Sigma_n, C_n) = 2n - 3$  and  $2n - 2 \leq \lambda_2(\Sigma_n, C_n) \leq 4n - 6$ .
- (iii)  $\frac{n+1}{2} \leq \lambda_m(\Sigma_n, R_n) \leq n - 1, m = 1, 2$ .

*Proof* (i) Note that if  $g \in \Sigma_n$  and  $t_{i,j} \in T_n$ , then:

$$g^{-1}t_{i,j}g = t_{g^{-1}(i),g^{-1}(j)}. \tag{7}$$

Therefore  $\lambda_1(\Sigma_n, T_n) = 1$  by (1). Thus, by Inequality (3), we have  $\lambda_2(\Sigma_n, T_n) \leq 2$ . The equality  $\lambda_2(\Sigma_n, T_n) = 2$  follows by (2) and the fact that  $g^{-1}t_{k,l}t_{i,j}g = t_{g^{-1}(i),g^{-1}(j)}t_{g^{-1}(k),g^{-1}(l)}$  holds for any  $g \in \Sigma_n$  and  $1 \leq i < j < k < l \leq n$ .

- (ii) Consider the permutation  $g \in \Sigma_n$  given by  $g = [2, 3, \dots, n - 1, n, 1]$ . Then  $g^{-1}t_{1,2}g = [n, 2, 3, \dots, n - 1, 1]$ . Therefore,  $l_C(g^{-1}t_{1,2}g) \geq 2n - 3$  (since to transform  $[n, 2, 3, \dots, n - 1, 1]$  to  $1_{\Sigma_n}$  requires moving 1 and  $n$  back to their original positions). Therefore,  $\lambda_1(\Sigma_n, C_n) \geq 2n - 3$  by (1). But, by Equality (7),  $\lambda_1(\Sigma_n, C_n) \leq 2n - 3$ , since any transposition is the product of at most  $2n - 3$  elements in  $C$ . In particular,  $\lambda_1(\Sigma_n, C_n) = 2n - 3$ . Similarly,  $l_C(g^{-1}t_{1,2}t_{3,4}g) \geq 2n - 2$ , and so  $\lambda_2(\Sigma_n, C_n) \geq 2n - 2$  by (2). Hence, by Inequality (3), we have  $\lambda_2(\Sigma_n, C_n) \leq 2(2n - 3)$ .

- (iii) The inequality  $\lambda_m(\Sigma_n, R_n) \leq n - 1, m = 1, 2$  follows as the diameter of  $Cay(\Sigma_n, R_n)$  is at most  $n - 1$ .

Now, suppose  $n$  is odd. Let  $g \in \Sigma_n$  be given by  $g = [3, 2, 5, 4, 7, 6, \dots, n - 3, n, n - 1, 1]$ . Then it is straight-forward to check that  $d_{BP}(r_{1,n}g, g) = n + 1$  (see Fig. 2 for the case  $n = 7$ ). In particular, since the length of any shortest path in  $Cay(\Sigma_n, R_n)$  joining any  $g, h \in \Sigma_n$  is at least  $d_{BP}(h, g)/2$  by (Setubal and Meidanis 1997, p. 238), we have  $\lambda_1(\Sigma_n, R_n) \geq \frac{n+1}{2}$ . Similarly,  $d_{BP}(r_{2,3}r_{1,n}g, g) = n + 1$  for any  $g \in \Sigma_n$ , and so  $\lambda_2(\Sigma_n, R_n) \geq \frac{n+1}{2}$ .

In case  $n$  is even, consider  $g = [3, 2, 5, 4, 7, 6, \dots, n - 4, n - 1, n - 2, 1, n]$ . Then  $d_{BP}(r_{2,n}g, g) = n + 1$  and  $d_{BP}(r_{3,4}r_{2,n}g, g) = n + 1$ . Similar reasoning yields the desired result.  $\square$

### 2.3 Signed permutations

In genomics, the direction in which a gene is oriented in a genome can also provide useful information to incorporate in rearrangement models (Hannenhalli and Pevzner 1999), which can be expressed as follows in terms of Cayley graphs (Kostantinova 2008). The *hyperoctahedral group*  $B_n$  is defined as the group of all permutations  $g^\sigma$  acting on the set  $\{\pm 1, \dots, \pm n\}$  such that  $g^\sigma(-i) = -g^\sigma(i)$  for all  $i \in [n]$ . An element of  $B_n$  is a *signed permutation*. Signed versions of transpositions and reversals can be defined in the obvious way; a sign change transposition  $t_{i,j}^\sigma$  ( $i \leq j$ ) switches the values in the  $i$ th and  $j$ th positions of a signed permutation as well as both of their signs and so forth. Note that we also allow  $i = j$  for signed transpositions and reversals so that  $t_{i,i} = r_{i,i}$ ,  $i \in [n]$ , simply switches the sign of the  $i$ th value. We denote the set of signed elements corresponding to those in  $S = T_n, C_n, R_n$ , together with the elements  $t_{i,i}$ ,  $1 \leq i \leq n$ , by  $S^\sigma$ . Note that the diameter of  $Cay(B_n, R_n^\sigma)$  is  $n + 1$  (Kostantinova 2008).

Now, the group  $B_n$  a semidirect product of  $\mathbb{Z}_2^n$  and a subgroup isomorphic to  $\Sigma_n$  and so we have a short exact sequence:

$$1_G \rightarrow \mathbb{Z}_2^n \rightarrow B_n \xrightarrow{p} \Sigma_n \rightarrow 1_G, \tag{8}$$

where the homomorphism  $p : B_n \rightarrow \Sigma_n$  sends  $g^\sigma \in B_n$  to the permutation of  $[n]$  that maps  $i$  to  $|g^\sigma(i)|$  (i.e. it ignores the sign). Notice that  $p$  maps  $S^\sigma$  onto  $S$  when  $S = T_n, C_n, R_n$ . In particular, from Lemma 3, (8) furnishes the following inequalities for  $S = T_n, C_n, R_n$  and  $m = 1, 2$ :

$$\lambda_m(B_n, S^\sigma) \geq \lambda_m(\Sigma_n, S). \tag{9}$$

This leads to the following bounds.

**Corollary 1** *For  $n \geq 7$ , the following hold:*

- (i)  $\lambda_1(B_n, T_n^\sigma) \leq 3$  and  $\lambda_2(B_n, T_n^\sigma) \leq 6$ .
- (ii)  $2n - 3 \leq \lambda_1(B_n, C_n^\sigma) \leq 2n - 1$  and  $2n - 2 \leq \lambda_2(B_n, C_n^\sigma) \leq 4n - 2$ .
- (iii)  $\frac{n+1}{2} \leq \lambda_m(B_n, R_n^\sigma) \leq n + 1$ ,  $m = 1, 2$ .

*Proof* The inequalities  $\lambda_1(B_n, T_n^\sigma) \leq 3$  and  $\lambda_1(B_n, C_n^\sigma) \leq 2n - 1$  follow from similar arguments to those used in the proof of Theorem 1 (i) and (ii), using the signed analogue of Eq. (7). Inequality (3) then implies that inequalities  $\lambda_2(B_n, T_n^\sigma) \leq 6$  and  $\lambda_2(B_n, C_n^\sigma) \leq 4n - 2$  both hold. The inequality  $\lambda_m(B_n, R_n^\sigma) \leq n + 1$ ,  $m = 1, 2$ , follows as the diameter of  $Cay(B_n, R_n^\sigma)$  is at most  $n + 1$ . The inequalities  $2n - 3 \leq \lambda_1(B_n, C_n^\sigma)$  and  $2n - 2 \leq \lambda_2(B_n, C_n^\sigma)$ , and the remaining ones in (iii) follow by Inequality (9) and Theorem 1.  $\square$

### 2.4 Beyond $d_S$ : properties of breakpoint distance

As we have seen in Sect. 2.2, it can sometimes be useful to consider the breakpoint distance  $d_{BP}$  when studying genome rearrangements. In genomics, this distance is commonly used as a proxy for rearrangement distances since, for example, it is easier to compute (Sankoff and Blanchette 1997, 1998). Thus it is of interest to note:

**Lemma 4** *For  $n \geq 7$ , the following hold:*

- (i)  $\lambda_1(\Sigma_n, T_n, d_{BP}) \leq 4$  and  $\lambda_2(\Sigma_n, T_n, d_{BP}) \leq 8$ .
- (ii)  $\lambda_1(\Sigma_n, C_n, d_{BP}) \leq 4$  and  $\lambda_2(\Sigma_n, C_n, d_{BP}) \leq 8$ .
- (iii)  $\frac{n+1}{2} \leq \lambda_m(\Sigma_n, R_n, d_{BP}) \leq n + 1, m = 1, 2$ .

*Proof* Suppose  $t = t_{i,j} \in T_n, 1 \leq i < j \leq n$ . Using Eq. (7), it is straightforward to see that  $d_{BP}(tg, g) \leq 4$  holds for any  $g \in \Sigma_n$ . Therefore  $\lambda_1(\Sigma_n, T_n, d_{BP}), \lambda_1(\Sigma_n, C_n, d_{BP}) \leq 4$ . The inequalities in (i) and (ii) involving  $\lambda_2$  now follow from Inequality (6).

The Inequalities in (iii) follow from the argument used in the proof of Theorem 1 (iii) and the diameter of  $d_{BP}$  on  $\Sigma_n$ . □

In particular, for  $C_n$ , the set of Coxeter generators of  $\Sigma_n$ , and  $m = 1, 2$ , we have  $\lambda_m(\Sigma_n, C_n) \geq 2n - 3$ , but  $\lambda_m(\Sigma_n, C_n, d_{BP}) \leq 4$ . Intriguingly, this observation can be extended as follows. For  $k \geq 1$ , let  $R_n^{(k)}$ , denote the set of reversals of the form  $\{r_{i,j} : 1 \leq i < j \leq n, |i - j| \leq k\}$ . Such ‘fixed-length’ reversals have been considered in the context of genome rearrangements in e.g. Chen and Skiena (1996). Note that  $R_n^{(1)} = C_n$  and  $R_n^{(k)} \subseteq R_n^{(k+1)}$ , so that  $R_n^{(k)}$  generates  $\Sigma_n$ .

**Proposition 2** *For  $n \geq 7, n \geq k \geq 1$  and  $m = 1, 2$ ,*

$$\lambda_m(\Sigma_n, R_n^{(k)}) \geq 2 \left\lceil \frac{n}{k} \right\rceil - 2,$$

and

$$\lambda_m(\Sigma_n, R_n^{(k)}, d_{BP}) \leq 4(k + 1).$$

*Proof* As in the proof of Theorem 1 (ii), let  $g \in \Sigma_n$  be given by  $g = [2, 3, \dots, n - 1, n, 1]$ , so that  $g^{-1}r_{1,2}g = [n, 2, 3, \dots, n - 1, 1]$ . Then we have  $l_{R_n^{(k)}}(g^{-1}r_{1,2}g) \geq 2\lceil \frac{n}{k} \rceil - 3$ , since to transform  $[n, 2, 3, \dots, 1]$  to  $1_{\Sigma_n}$  requires moving 1 and  $n$  back to their original positions. Similarly,  $l_C(g^{-1}r_{1,2}r_{3,4}g) \geq 2\lceil \frac{n}{k} \rceil - 2$ . This gives the first inequality in the proposition. Moreover, if  $r_{i,j}, r_{p,q} \in R_n^{(k)}$ , then it is straight-forward to see that  $d_{BP}(r_{i,j}g, g) \leq 2(k + 1)$  and  $d_{BP}(r_{p,q}r_{i,j}g, g) \leq 4(k + 1)$  holds, which gives the second inequality in the proposition. □

This proposition implies that in genomics applications, adding or substituting a single reversal in a sequence of reversals in  $R_n^{(k)}$  could potentially have a large effect on  $d_{R_n^{(k)}}$ , but a relatively small effect on  $d_{BP}$  (especially for large values of  $n$ , e.g. there

are  $n \geq 20,000$  genes in the human genome). It could be of interest to see whether other combinations of generating sets and metrics for  $\Sigma_n$  commonly used in genomics [such as transpositions (Labarre 2006) and the  $k$ -mer distance (Trifonov and Rabadan 2010)] exhibit a similar type of behaviour.

### 3 Statistical implications

So far we have considered metric sensitivity from a purely combinatorial and deterministic perspective. But it is also of interest to investigate the sensitivity of the metrics discussed above when the elements of  $S$  are randomly assigned by some stochastic process. Again, the motivation for this question comes from genomics, where stochastic models often play a central role [see, for example, Mossel and Steel (2005); Wang and Warnow (2005)]. In this section, we establish a result (Proposition 3) in which the quantity  $\lambda_2$  plays a crucial role in allowing underlying parameters in such stochastic models to be estimated accurately given sufficiently long genome sequences. Our motivation here is to provide some basis for eventually extending the well-developed (and tight) results on the sequence length requirements for tree reconstruction under site-substitution models [see e.g. Daskalakis et al. (2010); Erdős et al. (1999); Gronau et al. (2008); Mossel and Steel (2005)] to more general models of genome evolution.

As before, suppose a finite group  $G = G_n$  acts transitively and freely on a set  $X = X_n$  of genomes of length  $n$ , and that  $S = S_n$  is a symmetric set of generators for  $G_n$ . Consider a stationary stochastic process that generates elements of  $S_n$  according to a Poisson process. We allow the Poisson intensity at which each element  $s$  is generated to vary with  $s$ , requiring only that it matches the Poisson intensity of  $s^{-1}$  (this last requirement applies by default if  $s^{-1} = s$ , as is the case in most examples we have described thus far).<sup>4</sup> Starting from any initial genome, such a stochastic process provides a mechanism to describe the evolution of this genome - simply apply the process for a certain duration, and each time an element  $s \in S_n$  is generated it acts on the current genome.

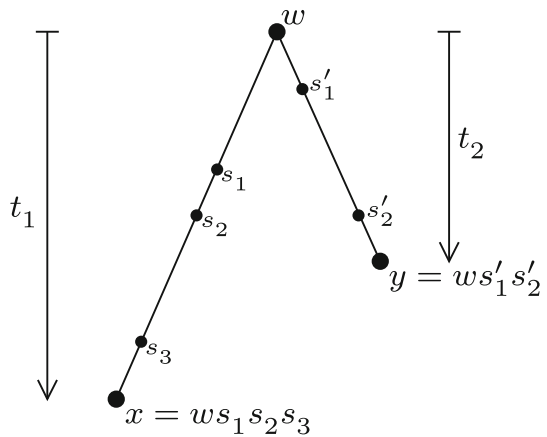
Suppose two genomes  $x$  and  $y$  are generated by two independent applications of this process to an ancestral genome  $w$  for durations  $t_1$  and  $t_2$  (see Fig. 3). We would like to estimate the 'evolutionary distance' ( $t_1 + t_2$ ) between  $x$  and  $y$ , which we will denote by  $\mu(x, y)$ , by comparing  $x$  and  $y$ , and without knowledge of the ancestral genome  $w$ .

First observe that, since the action of  $G = G_n$  on  $X_n$  is transitive and free, there is a unique value of  $g = g_{xy} \in G$  for which  $y = xg$ , and we will show how one can use (a normalization of)

$$\delta(x, y) := d(g_{xy}, 1_G)$$

<sup>4</sup> An example a process that allows Poisson intensity to vary with each generator  $s$  is provided by the Kimura 3ST model of site substitution, in which site nucleotide transitions and the two types of nucleotide transversions typically proceed at different rates.

**Fig. 3** An example of two genomes  $x$  and  $y$  that have evolved independently from some (unknown) ancestral genome  $w$ . Each branch represents an independent application of a continuous-time stochastic process that generates elements of  $S$ , each of which acts sequentially on the current genome (see text for details)



to estimate  $\mu(x, y)$ , for any metric  $d$  on  $G$  that satisfies the conditions:

- (i)  $d(g^{-1}, 1_G) = d(g, 1_G)$ , and
- (ii) for some  $\lambda > 0$ , we have  $\lambda_2(G_n, S_n, d) \leq \lambda$  for all  $n$ .

Note that all metrics considered thus far satisfy condition (i), which ensures that  $\delta(x, y) = \delta(y, x)$ , for all  $x, y \in X_n$ .

Let  $\mu_n(x, y)$  denote the expected total number of times elements of  $S_n$  are selected by the stochastic process when it is run for duration  $\mu(x, y)$ . Then  $\mu_n(x, y)$  is proportional to  $\mu(x, y)$ ; we will assume that  $\mu_n(x, y)$  is also proportional to the genome length  $n$ . That is, we impose the condition:

- (iii)  $\mu_n(x, y) = n \cdot r \mu(x, y)$ , where  $r > 0$  is a constant.

We also assume one further condition:

- (iv)  $\bar{\delta} = n \cdot f(\mu(x, y))$ , where  $\bar{\delta}$  is the expected value in the model of  $\delta(x, y)$ , and  $f$  is a function with strictly positive but bounded first derivative on  $[0, \infty)$ .

A specific example that satisfies these four conditions (i)–(iv) is site substitutions under the Kimura 3ST model (as described in Sect. 2.1 and taking  $d = d_{S_n}$ ). Observe that Properties (i) and (ii) hold (in this case,  $\delta(x, y)$  is simply the ‘Hamming distance’ between the sequences which counts the number of sites at which  $x$  and  $y$  differ). Property (iii) also holds, as does Property (iv) since

$$\bar{\delta} = n \cdot \frac{3}{4}(1 - \exp(-4\rho\mu(x, y)/3)),$$

where  $\mu(x, y)$  is the time separating  $x$  and  $y$  and  $\rho$  is the rate of substitution.

Turning to permutation-type genome rearrangements, note that both breakpoint distance  $d_{BP}$  and  $d_{S_n}$  (for  $S_n = T_n, C_n$  or  $R_n$ ) satisfy (i), and we have described above some cases where (ii) is satisfied. Whether conditions (iii) and (iv) hold depends on the details of the underlying stochastic process of genome rearrangement. For example, if reversals of fixed length occur randomly at a constant rate along a genome then condition (iii) will hold, and under the approximation to the Nadeau–Taylor model of

genome rearrangement studied in Sect. 2 of Wang (2002), Property (iv) would also hold for breakpoint distance [the proof relies on Corollary 1(a) of Wang (2002)].

The following result shows how  $\delta(x, y)/n$  can be used to estimate  $f(\mu(x, y))$  accurately, and thereby  $\mu(x, y)$  (by the assumptions regarding  $f$ ). The ability to estimate  $\mu(x, y)$  accurately provides a direct route to accurate phylogenetic tree reconstruction by standard phylogenetic methods [such as ‘neighbor-joining’ (Saitou and Nei 1987)] since  $\mu(x, y)$  is ‘additive’ on the underlying tree [for details, see Semple and Steel (2003)].

**Proposition 3** *Consider a stochastic model of genome evolution and a metric  $d$  on  $G_n$  as described above, that together satisfy conditions (i)–(iv). The probability that  $\delta(x, y)/n$  differs from  $f(\mu(x, y))$  by more than any fixed value  $z > 0$  converges to zero exponentially quickly with increasing  $n$ . More precisely, for constants  $b > 0$  (dependent on  $\mu(x, y)$ ), and  $c > 0$  (dependent on  $\mu(x, y)$  and  $\lambda$ ) we have:*

$$\mathbb{P}(|\delta/n - f(\mu(x, y))| \geq z) \leq 2 \exp(-bn) + 2 \exp(-cz^2n),$$

for  $\delta = \delta(x, y)$ .

*Proof* Suppose that  $W_1, W_2, \dots, W_k$  are independent random variables taking values in the set  $S$ , and  $h$  is any real-valued function defined on  $S$  that satisfies the following property for some constant  $\xi$ :

$$|h(w_1, w_2, \dots, w_k) - h(w'_1, w'_2, \dots, w'_k)| \leq \xi,$$

whenever  $(w_i)$  and  $(w'_i)$  differ at just one coordinate. In this general setting, a form of the Azuma–Hoeffding inequality [see e.g. Alon and Spencer (1992)] states that the random variable  $Y := h(W_1, W_2, \dots, W_k)$  satisfies the following tight concentration bound about its mean (for all  $k \geq 1$ ):

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq z) \leq 2 \exp\left(-\frac{z^2}{2\xi^2k}\right). \tag{10}$$

We apply this general result as follows.

Let random variable  $K \geq 0$  denote the total number of times elements of  $S_n$  are selected during the stochastic process that generates  $x$  and  $y$  from an ancestral genome  $w$ . By assumption,  $K$  has a Poisson distribution, with mean equal to  $n \cdot r\mu(x, y)$  (by condition (iii)).

Now, if  $K = k$  where  $k \geq 1$  we have  $x = wg_x$  and  $y = wg_y$ , hence  $y = xg_x^{-1}g_y$ , where  $g_x$  and  $g_y$  are products of a total of  $k$  elements of  $S_n$ . Thus,  $g_{xy} = g_x^{-1}g_y$  and we will regard this as a sequence consisting of  $k$  elements of  $S_n$  that begins with the inverses of the elements of  $S_n$  that make up  $g_x$ , in reverse order, followed by the elements of  $S_n$  that make up  $g_y$ . For the example in Fig. 3, this sequence would be  $s_3^{-1}, s_2^{-1}, s_1^{-1}, s'_1, s'_2$  (the reason for this transformation is that it converts a process on three genomes  $(x, y, w)$  into one on just two genomes  $(x, y)$  separated by duration  $\mu(x, y)$ ).

Let random variables  $W_1, \dots, W_k$  denote these  $k$  elements of  $S_n$  in this (transformed) order under the stochastic model. By the assumptions of the model, the  $W_i$  are independent; moreover,  $\delta(x, y) = d(g_{xy}, 1_G) = d(W_1W_2 \cdots W_k, 1_G)$  and so, by

(ii), this function satisfies the requirements of the Azuma–Hoeffding inequality for  $\xi = \lambda$ , upon taking  $h(w_1, \dots, w_k) = d(w_1 w_2 \dots w_k, 1_G)$ .

Thus (10) furnishes the following inequality:

$$\mathbb{P}(|\delta/n - \bar{\delta}/n| \geq z \mid K = k) \leq 2 \exp\left(-\frac{z^2 n^2}{2\lambda^2 k}\right). \tag{11}$$

Invoking condition (iv) and the law of total probability, we obtain:

$$\mathbb{P}(|\delta/n - f(\mu(x, y))| \geq z) = \sum_{k \geq 0} \mathbb{P}(|\delta/n - \bar{\delta}/n| \geq z \mid K = k) \mathbb{P}(K = k),$$

from which (11) ensures the inequality:

$$\mathbb{P}(|\delta/n - f(\mu(x, y))| \geq z) \leq 2\mathbb{E}\left[\exp\left(-\frac{z^2 n^2}{2\lambda^2 K}\right)\right], \tag{12}$$

where  $\mathbb{E}$  denotes expectation with respect to  $K$ . Let us write  $\mathbb{E}[\exp(-\frac{z^2 n^2}{2\lambda^2 K})]$  as a weighted sum of two conditional expectations:

$$\begin{aligned} &\mathbb{E}\left[\exp\left(-\frac{z^2 n^2}{2\lambda^2 K}\right) \mid K > 2nr\mu(x, y)\right] \cdot p \\ &+ \mathbb{E}\left[\exp\left(-\frac{z^2 n^2}{2\lambda^2 K}\right) \mid K \leq 2nr\mu(x, y)\right] \cdot (1 - p), \end{aligned} \tag{13}$$

where  $p = \mathbb{P}(K > 2nr\mu(x, y))$ . The first term in (13) is bounded above by  $\mathbb{P}(K > 2nr\mu(x, y))$  since  $\exp(-\frac{z^2 n^2}{2\lambda^2 K}) \leq 1$ ; moreover, since  $K$  has a Poisson distribution with mean  $nr\mu(x, y)$  (and so is asymptotically normally distributed with mean and variance equal to  $nr\mu(x, y)$ ), the quantity  $\mathbb{P}(K > 2nr\mu(x, y))$  is bounded above by a term of the form  $\exp(-bn)$  where  $b$  depends just on  $\mu(x, y)$ .

The second term in (13) is bounded above by  $\exp(-\frac{z^2 n}{4\lambda^2 r\mu(x, y)})$ , where  $\lambda = \lambda_2(G, S, d)$ , since the function  $x \mapsto \exp(-A/x)$  increases monotonically on  $[0, \infty)$ .

Combining these two bounds in (13), the result now follows from (12).  $\square$

*Remark* Referring again to the particular case of site substitutions under the Kimura 3ST model, Proposition 3 can be strengthened to:

$$\mathbb{P}(|\delta/n - f(\mu(x, y))| \geq z) \leq 2 \exp(-c' z^2 n),$$

where  $c' > 0$  can be chosen to be independent of  $\mu(x, y)$ . This stronger result is the basis of numerous results in the phylogenetic literature that show that large trees can be reconstructed from remarkably short sequences under simple site-substitution models (Erdős et al. 1999). Although the bound in Proposition 3 is less incisive, it would be of interest to explore similar phylogenetic applications for other models of genome evolution in which  $\lambda_2$  is independent of  $n$ , such as those involving breakpoint distance under reversals of fixed length.

## 4 Concluding comments

Gene order provides a promising tool for inferring evolutionary history, by providing a means to estimate the taxonomic distance between extant species. However, because the particular sequence of genome rearrangements in the past is generally unknown, the question of how sensitive such estimates are to perturbations in the sequence arises. We have investigated some classes of genome rearrangement where single-element perturbations either have a provably small impact, or potentially large impact on the resulting genomes (and thereby on the estimates of taxonomic distance).

It would be interesting to extend the analysis in this paper to other types of genome rearrangement events of biological relevance, such as such as translocations (Labarre 2006) and block-interchanges (Chin et al. 2007).

Also, we have concentrated on worst-case properties of perturbation in this paper, which is useful for applying the Azuma inequality in the previous section. However, it would also be of interest undertake an *average* case analysis of the expected effect of perturbations on the distance between genomes. In particular, what is the expected impact of a small change in a sequence of genome rearrangements? A related task would be to investigate the *variance* of the distance between genomes under the stochastic model, which is related to the average case analysis by a standard inequality (Steele 1986). Simulation studies would provide some initial insights into these questions.

**Acknowledgments** We thank Marston Conder, Eamonn O'Brien and Li San Wang for some helpful comments. We also thank the two anonymous reviewers for several helpful suggestions.

## References

- Alon N, Spencer J (1992) The probabilistic method. Wiley, New York
- Bafna V, Pevzner PA (1996) Genome rearrangements and sorting by reversals. *SIAM J Comput* 25(2):272–289
- Bergeron A, Mixtacki J, Stoye J (2009) A new linear time algorithm to compute the genomic distance via the double cut and join distance. *Theor Comput Sci* 410(51):5300–5316
- Chen T, Skiena S (1996) Sorting with fixed-length reversals. *Discret Appl Math* 71:269–295
- Chin LL, Ying CL, Yen LH, Chuan YT (2007) Analysis of genome rearrangement by block-interchanges. *Methods Mol Biol* 396:121–134
- Daskalakis C, Mossel E, Roch S (2010) Evolutionary trees and the Ising model on the Bethe lattice: a proof of Steel's conjecture. *Probab Theor Relat Fields* 149:149–189
- Eppstein DBA (1992) Word processing in groups. A K Peters/CRC Press, New York
- Erdős PL, Steel MA, Székely LA, Warnow T (1999) A few logs suffice to build (almost) all trees (part 1). *Rand Struct Alg* 14(2):153–184
- Evans SN, Speed TP (1993) Invariants of some probability models used in phylogenetic inference. *Ann Stat* 21:355–377
- Fertin G, Labarre A, Rusu I, Tannier E, Vialette S (2009) Combinatorics of genome rearrangements. The MIT Press, Cambridge
- Gronau I, Moran S, Snir S (2008) Fast and reliable reconstruction of phylogenetic trees with very short edges. In: SODA: ACM-SIAM symposium on discrete algorithms. Society for Industrial and Applied Mathematics, Philadelphia, pp. 379–388
- Hannenhalli S, Pevzner PA (1999) Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations via reversals. *J Assoc Comput Mach* 46(1):1–27
- Hilborn RC (2004) Sea gulls, butterflies, and grasshoppers: a brief history of the butterfly effect in nonlinear dynamics. *Am J Phys* 72(4):425–427

- Holmgren R (1994) A first course in discrete dynamical systems, 2nd edn. Springer, New York
- Kececioğlu JD, Sankoff D (1995) Exact and approximate algorithms for sorting by reversals with application to genome rearrangement. *Algorithmica* 13:180–210
- Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. *Proc Natl Acad Sci USA* 78:454–458
- Kostantinova E (2008) Some problems on Cayley graphs. *Linear Algebra Appl.* 429:2754–2769
- Kunkle D, Cooperman G (2009) Harnessing parallel disks to solve Rubik's cube. *J Symb Comput* 44(7):872–890
- Labarre L (2006) New bounds and tractable instances for the transposition distance. *IEEE/ACM Trans Comput Biol Bioinf* 3(4):380–394
- Mossel E, Steel M (2005) How much can evolved characters tell us about the tree that generated them? In: Gascuel O (ed) *Mathematics of evolution and phylogeny*. Oxford University Press, Oxford, pp 384–412
- Pevzner P (2000) *Computational molecular biology*. MIT Press, Cambridge
- Rotman JJ (1995) *An introduction to the theory of groups*. Springer, New York
- 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, Blanchette M (1997) The median problem for breakpoints in comparative genomics. *Computing and Combinatorics, Shanghai*, pp 251–263
- Sankoff D, Blanchette M (1998) Multiple genome rearrangement and breakpoint phylogeny. *J Comput Biol* 5:555–570
- Setubal J, Meidanis M (1997) *Introduction to computational molecular biology*. PWS Publishing Company, Boston
- Semple C, Steel M (2003) *Phylogenetics*. Oxford University Press, Oxford
- Sinha A, Meller J (2008) Sensitivity analysis for reversal distance and breakpoint re-use in genome rearrangements. *Pac J Biocomput* 13:37–48
- Steele J.M. (1986) An Efron-Stein inequality for nonsymmetric statistics. *Ann Stat* 14(2):753–758
- Trifonov V, Rabadan R (2010) Frequency analysis techniques for identification of viral genetic data. *mBio* 1(3):e00156-10
- Wang L-S (2002) Genome rearrangement phylogeny using weighbor. In: *Lecture Notes for Computer Sciences* No. 2452. Proceedings for the second workshop on algorithms in bioinformatics (WABI'02), Rome, pp 112–125
- Wang L-S, Warnow T (2005) Distance-based genome rearrangement phylogeny. In: Gascuel O (ed) *Mathematics of evolution and phylogeny*. Oxford University Press, Oxford, pp 353–380