

# Implicit Transpositions in Shortest DCJ Scenarios

Shuai Jiang and Max A. Alekseyev<sup>(✉)</sup>

George Washington University, Washington, DC, USA  
maxal@gwu.edu

**Abstract.** Genome rearrangements are large-scale evolutionary events that shuffle genomic architectures. The minimal number of such events between two genomes is often used in phylogenomic studies to measure the evolutionary distance between the genomes. Double-Cut-and-Join (DCJ) operations represent a convenient model of most common genome rearrangements (reversals, translocations, fissions, and fusions), while other genome rearrangements, such as transpositions, can be modeled by pairs of DCJs. Since the DCJ model does not directly account for transpositions, their impact on DCJ scenarios is unclear.

In the current work, we study implicit appearance of transpositions (as pairs of DCJs) in shortest DCJ scenarios and prove uniform lower and upper bounds for their proportion. Our results imply that implicit transpositions may be unavoidable and even appear in a significant proportion for some genomes. We estimate that in mammalian evolution transpositions constitute at least 17 % of genome rearrangements.

**Keywords:** Genome rearrangements · Transpositions · DCJ

## 1 Introduction

Genome rearrangements are dramatic evolutionary events that change genome structures. Since large-scale rearrangements are rare, it is natural to assume that the evolution history between two genomes corresponds to a shortest rearrangement scenario between them. The most common rearrangements are *reversals* that inverse contiguous segments of chromosomes, *translocations* that exchange tails of two chromosomes, and *fissions/fusions* that split/glue chromosomes. All these rearrangements can be conveniently modeled by the Double-Cut-and-Join (DCJ) operations [15], also known as 2-breaks [2], which make 2 “cuts” in a genome and “glues” the resulting genomic fragments in a new order.

*Transpositions* represent yet another type of genome rearrangements that relocate genomic segments across the genome. In contrast to reversal-like rearrangements modeled by DCJs (2-breaks), transpositions correspond to 3-breaks [2], which make 3 cuts and 3 gluings in a genome. Transpositions are more “powerful” than reversal-like rearrangements and in the model that includes both types of rearrangements (as 3-breaks and DCJs), transpositions tend to appear

in shortest scenarios in a large proportion. However, in reality transpositions happen more rarely than reversals and typically appear in a small proportion in the course of evolution (e.g., in *Drosophila* evolution transpositions are estimated to constitute less than 10% of genome rearrangements [13]). Earlier we showed that even the most promising model of *weighted genomic distance* [4, 6, 7] (where transpositions are assigned a higher weight) cannot bound the proportion of transpositions in the resulting rearrangement scenarios to biologically reasonable value [8]. This result emphasizes the need for a biologically adequate model for analysis of transpositions among other types of genome rearrangements.

While a transposition cannot be directly modeled by a DCJ, it can be modeled by a pair of DCJs. We refer to such pair of DCJs as an *implicit transposition*. We remark that DCJs forming an implicit transposition may not necessarily appear consecutively in a DCJ scenario. Furthermore, two implicit transpositions may share a DCJ and thus correspond to at most one actual transposition. We study appearance of implicit transpositions in shortest DCJ scenarios and derive a lower and an upper bounds for their proportion. Our results imply that implicit transpositions may be unavoidable in shortest DCJ scenarios between some genomes and even appear in a large proportion. In particular, we describe an extreme case where shortest DCJ scenarios entirely consist of implicit transpositions.

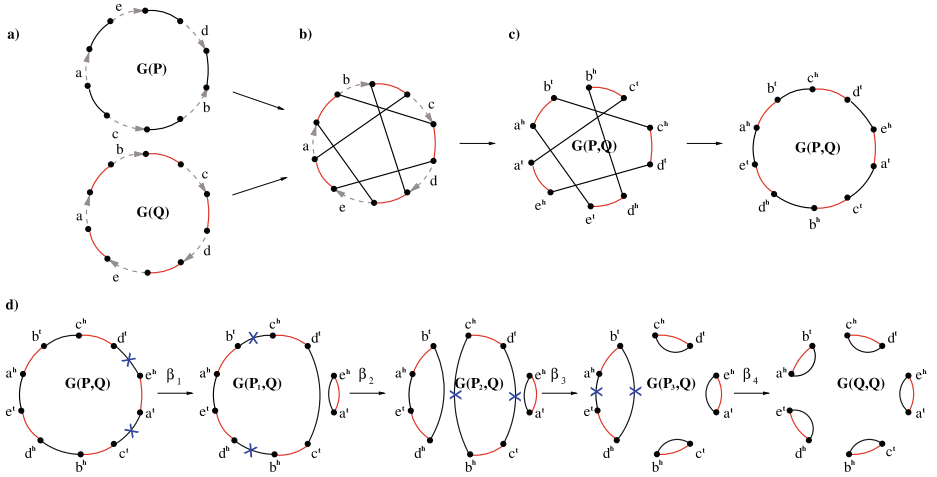
The paper is organized as follows. We describe graph-theoretical representation of genomes, rearrangement models (DCJs and  $k$ -breaks), and rearrangement scenarios in Sect. 2, and length-preserving modifications of DCJ scenarios along with dependency graphs capturing their combinatorial structure in Sect. 3. In Sect. 4, we study the appearance of implicit transpositions in shortest DCJ scenarios between two genomes and prove the uniform lower and upper bounds for their proportion (rate). In Sect. 5, we apply the obtained results for estimating the rate transpositions in mammalian evolution. We conclude the paper with discussion in Sect. 6.

## 2 Breakpoint Graphs and Rearrangement Scenarios

In the current study, we restrict our analysis to genomes with circular chromosomes (analysis of genomes with linear chromosomes will be published elsewhere). We represent a circular chromosome in genome  $P$  consisting of  $n$  genes as a cycle with  $n$  directed edges (encoding genes and their strands) alternating with  $n$  undirected edges connecting extremities of adjacent genes. A genome graph  $G(P)$  is a collection of such cycles (Fig. 1a).

A DCJ [15] (also called a 2-break [2]) in genome  $P$  corresponds to a replacement of a pair of undirected edges with a different pair of undirected edges on the same set of four vertices in  $G(P)$ . Similarly, a 3-break [2] in genome  $P$  corresponds to a replacement of a triple of undirected edges with different triple of undirected edges on the same set of six vertices in  $G(P)$ .

For genomes  $P$  and  $Q$  composed of the same set of genes, the *breakpoint graph*  $G(P, Q)$  is defined as the superposition of individual genome graphs  $G(P)$  and  $G(Q)$ , and can be constructed by “gluing” the identically labeled directed edges in the graphs (Fig. 1b, c). From now on, we will ignore directed edges and assume that  $G(P, Q)$  consists only of undirected edges, where the edges from genome  $P$  ( $P$ -edges) are colored black and the edges from genome  $Q$  ( $Q$ -edges) are colored red. Then the breakpoint graph  $G(P, Q)$  represents a collection of cycles consisting of undirected edges alternating between black and red colors. We distinguish the following types of cycles with respect to their *length*  $\ell$  (i.e., the number of black edges in a cycle): trivial cycles ( $\ell = 1$ ) and odd cycles ( $\ell$  is odd). We denote the number of cycles, trivial cycles, and odd cycles in  $G(P, Q)$  as  $c(P, Q)$ ,  $c_1(P, Q)$ , and  $c_{odd}(P, Q)$ , respectively.



**Fig. 1.** (a) Genome graphs  $G(P)$  and  $G(Q)$  for unichromosomal circular genomes  $P = (+a + e + d - b - c)$  and  $Q = (+a + b + c + d + e)$ , where undirected  $P$ -edges and  $Q$ -edges are colored black and red, respectively. (b) The superposition of genome graphs  $G(P)$  and  $G(Q)$ . (c) The breakpoint graph  $G(P, Q)$  is obtained from the superposition of  $G(P)$  and  $G(Q)$  with removal of directed edges. The graph  $G(P, Q)$  is formed by a single black-red cycle, i.e.,  $c(P, Q) = 1$ . (d) A transformation of the breakpoint graph  $G(P, Q)$  into  $G(Q, Q)$ , which corresponds to a shortest DCJ scenario (of length  $d_{DCJ}(P, Q) = 4$ ) between genomes  $P$  and  $Q$ .

A DCJ scenario from genome  $P$  to genome  $Q$  corresponds to a transformation of the breakpoint graph  $G(P, Q)$  into the breakpoint graph  $G(Q, Q)$ , which consists of trivial cycles (Fig. 1d).

**Lemma 1** ([2, 15]). *In a shortest DCJ scenario between two genomes, each DCJ splits some cycle in the corresponding breakpoint graph into two and thus increases the number of cycles by one.*

From Lemma 1, one can immediately get a formula for the DCJ distance:

**Theorem 1** ([2, 15]). *The DCJ distance between genomes  $P$  and  $Q$  on  $n$  genes is*

$$d_{\text{DCJ}}(P, Q) = n - c(P, Q).$$

We remark that a DCJ (2-break) represents a particular case of a 3-break. A 3-break that is not a DCJ is called *complete*. The following lemma describes the effect of DCJs and complete 3-breaks in shortest 3-break scenarios.

**Lemma 2** ([2, 8]). *In a shortest 3-break scenario between two genomes, each DCJ (2-break) splits an even cycle in the corresponding breakpoint graph into two odd cycles, while each complete 3-break splits an odd cycle into three odd cycles. Therefore, each rearrangement in such a scenario increases the number of odd cycles in the breakpoint graph by two.*

A 3-break scenario from genome  $P$  to genome  $Q$  increases the number of odd cycles from  $c_{\text{odd}}(P, Q)$  in  $G(P, Q)$  to  $c_{\text{odd}}(Q, Q)$ , which is the number of genes in  $Q$ . From Lemma 2, the following formula for the 3-break distance emerges:

**Theorem 2** ([2]). *The 3-break distance between genomes  $P$  and  $Q$  on  $n$  genes is*

$$d_3(P, Q) = \frac{n - c_{\text{odd}}(P, Q)}{2}.$$

We will also need the following theorem that easily follows from Lemma 1.

**Theorem 3.** *Along a transformation of the breakpoint graphs from  $G(P, Q)$  to  $G(Q, Q)$  corresponding to a shortest DCJ scenario between genomes  $P$  and  $Q$ , any edge once removed is never recreated.*

*Proof.* Lemma 1 implies that after an edge  $(u, v)$  is removed by a DCJ from the breakpoint graph, the vertices  $u$  and  $v$  start to belong to distinct cycles and can never belong to the same cycle again (which would be the case if the edge  $(u, v)$  is ever re-created).  $\square$

### 3 Length-Preserving Operations and Dependency Graphs

We call rearrangement scenarios (in particular, single DCJs or pairs of DCJs) between the same two genomes *equivalent*.

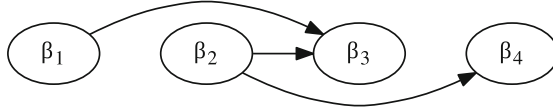
Since each DCJ removes and adds some edges in a breakpoint graph, two adjacent DCJs  $\alpha$  and  $\beta$  in a DCJ scenario are called *independent* if  $\beta$  removes edges that were not created by  $\alpha$ . Otherwise, if  $\beta$  removes some edge(s) created by  $\alpha$ , then  $\beta$  *depends* on  $\alpha$ . Furthermore, let  $k \in \{1, 2\}$  be the number of edges created by  $\alpha$  and removed by  $\beta$ . We say that  $\beta$  *strongly depends* on  $\alpha$  if  $k = 2$  and *weakly depends* on  $\alpha$  if  $k = 1$ . We remark that adjacent pair of strongly dependent DCJs may not appear in shortest DCJ scenarios, since such pair can be replaced by an equivalent single DCJ, decreasing the scenario length.

As we mentioned above, we can change the order of two adjacent independent DCJs and obtain an equivalent scenario. For a pair of adjacent weakly dependent DCJs, there exist exactly two other equivalent pairs of weakly dependent DCJs [5, 9]. We therefore consider the following two types of length-preserving operations, which can be applied to a pair of adjacent DCJs  $(\alpha, \beta)$  in a DCJ scenario:

- (T1) If  $\alpha$  and  $\beta$  are independent, we replace  $(\alpha, \beta)$  with  $(\beta, \alpha)$ .
- (T2) If  $\alpha$  and  $\beta$  are (weakly) dependent, we replace  $(\alpha, \beta)$  with an equivalent pair of weakly dependent DCJs.

It was shown [5] that any shortest DCJ scenario can be obtained from any other shortest DCJ scenario between the same two genomes using only operations of types (T1) and (T2).

To better capture and analyze the combinatorial structure of DCJs in a shortest DCJ scenario  $t$ , we construct the *dependency digraph*  $DG(t)$  (also known as *overlap graph* [11, 12]), whose vertices are labeled with DCJs from  $t$  and there is an arc  $(\alpha, \beta)$  whenever  $\beta$  depends on  $\alpha$  (Fig. 2).



**Fig. 2.** The dependency graph  $DG(t)$  for DCJ scenario  $t$  defined in Fig. 1d.

**Theorem 4.** Let  $t$  be a shortest DCJ scenario between genomes  $P$  and  $Q$  composed of the same  $n$  genes. Then

- (i) the number of arcs in  $DG(t)$  is  $n - 2 \cdot c(P, Q) + c_1(P, Q)$ ;
- (ii) both indegree and outdegree of each vertex in  $DG(t)$  are at most 2;
- (iii)  $t$  represents a topological ordering of  $DG(t)$ ;
- (iv)  $DG(t)$  is acyclic.

*Proof.* An arc  $(\alpha, \beta)$  in  $DG(t)$  corresponds in the breakpoint graph transformation  $t$  to an edge that is created by DCJ  $\alpha$  and removed by DCJ  $\beta$ . By Theorem 3 the removed edges are never recreated, implying that this correspondence is one-to-one.

There are  $n - c_1(P, Q)$   $P$ -edges in the non-trivial cycles in  $G(P, Q)$  and they have to be removed by DCJs from  $t$  in order to form trivial cycles. The other edges removed by DCJs from  $t$  must have been created by earlier DCJs. Since the total number of removed edges by DCJs in  $t$  is  $2 \cdot |t| = 2 \cdot d_{DCJ}(P, Q) = 2n - 2c(P, Q)$  (by Theorem 1), the number of such earlier created and then removed edges is  $2n - 2c(P, Q) - (n - c_1(P, Q)) = n - 2c(P, Q) + c_1(P, Q)$  and this gives the number of arcs in  $DG(t)$ .

Since by Theorem 3 the edges in the breakpoint graph transformation are not recreated, any DCJ in  $t$  (which removes two edges and creates two edges) depends on at most two other DCJs and may have at most two dependent DCJs. That is, both indegree and outdegree of any vertex in  $DG(t)$  are bounded by 2.

If  $(\alpha, \beta)$  is an arc in  $DG(t)$ , then a DCJ  $\beta$  removes some edge  $e$  created by a DCJ  $\alpha$ . By Theorem 3 no other DCJ besides  $\alpha$  can create  $e$ , and thus  $\beta$  must follow  $\alpha$  in  $t$ . So  $t$  represents a topological ordering for  $DG(t)$  and therefore  $DG(t)$  is acyclic.  $\square$

The following theorem refines the result of [5] with respect to sorting shortest DCJ scenarios with operations (T1) only.

**Theorem 5.** *Let  $t_1$  and  $t_2$  be shortest DCJ scenarios between the same two genomes. Scenario  $t_1$  can be obtained from scenario  $t_2$  with operations (T1) if and only if  $DG(t_1) = DG(t_2)$ .*

*Proof.* Suppose that  $t_1$  and  $t_2$  correspond to the same dependency graph, i.e.,  $DG(t_1) = DG(t_2) = G$ , then by Theorem 4 they represent topological orderings of  $G$ . We will show that  $t_1$  and  $t_2$  can be obtained from each other with operations (T1). Let

$$t_1 = (\alpha_1, \alpha_2, \dots, \alpha_k, \gamma, \dots) \text{ and} \\ t_2 = (\alpha_1, \alpha_2, \dots, \alpha_k, \beta_1, \beta_2, \dots, \beta_m, \gamma, \dots),$$

where  $\gamma \neq \beta_1$  are the first different DCJs in the two scenarios. We will show that  $\gamma$  in  $t_2$  can be moved to  $(k+1)$ -st position (i.e., its position in  $t_1$ ) with operations (T1). Since  $\beta_m$  follows  $\gamma$  in  $t_1$  but precedes  $\gamma$  in  $t_2$ , these vertices are not connected with an arc in  $G$  and we can apply operation (T1) to  $t_2$  to obtain  $(\alpha_1, \alpha_2, \dots, \alpha_k, \beta_1, \beta_2, \dots, \gamma, \beta_m, \dots)$ . After  $m$  such operations we get  $(\alpha_1, \alpha_2, \dots, \alpha_k, \gamma, \beta_1, \beta_2, \dots, \beta_m, \dots)$ , where  $\gamma$  is at the same position as in  $t_1$ . Using induction on  $k$ , we conclude that  $t_1$  can be obtained from  $t_2$  with operations (T1), and vice versa.

Now, suppose that DCJ scenarios  $t_1$  and  $t_2$  can be obtained from each other with operations (T1). Since operations (T1) changes only the order of DCJs in the scenario but keeps the DCJs themselves intact, the dependency graph is not affected by such operations either. Therefore,  $DG(t_1) = DG(t_2)$ .  $\square$

For a directed graph  $G$ , we define  $\overline{G}$  as the undirected graph obtained from  $G$  by making all arcs undirected. While Theorem 4 claims that  $DG(t)$  is acyclic, from the results of Shao *et al.* [14]<sup>1</sup> it follows that  $\overline{DG(t)}$  is a forest (i.e., a graph with no cycles):

**Theorem 6** ([14]). *Let  $t$  be a shortest DCJ scenario between two genomes composed of the same set of genes. Then the graph  $\overline{DG(t)}$  represents a forest.*

<sup>1</sup> Shao *et al.* [14] studies more general *trajectory graphs*, from which the dependency graphs can be obtained by contraction of edges.

## 4 Implicit Transpositions in Shortest DCJ Scenarios

While DCJs mimic most of common genome rearrangements (reversals, translocations, fissions, fusions), more complex rearrangements such as transpositions cannot be modeled by a single DCJ. A transposition, which cuts off a segment of a chromosome and inserts it into some other place in the genome, can be modeled by a pair of weakly dependent DCJs, replacing three undirected edges with three other undirected edges on the same six vertices in the genome graph. We remark that this operation is also known as a 3-break rearrangement [2].

Below we study how transpositions appearing in the course of evolution between two genomes may affect shortest DCJ scenarios between them. While a transposition constitutes a pair of consecutive DCJs, their positions in a DCJ scenario may not always be reconstructed correctly. In particular, the two DCJs forming a transposition may be interweaved with other independent DCJs that precede or follow the transposition in the evolutionary scenario, which inspires the following definition.

In a DCJ scenario  $t = (\alpha_1, \alpha_2, \dots, \alpha_n)$ , a pair of DCJs  $(\alpha_i, \alpha_j)$  forms an *implicit transposition* if they can be made adjacent by applying a number of operations (T1) and form a pair of weakly dependent DCJs. When these DCJs become adjacent, they can be replaced by a single transposition. We refer to such a transposition as *recovered* from the DCJ scenario  $t$ . This poses us a question of how many transpositions can be *simultaneously* recovered from a given shortest DCJ scenario  $t$ .

Since two distinct implicit transpositions in a shortest DCJ scenario  $t$  may share a DCJ, the maximum number of transpositions that can be recovered from  $t$  may be smaller than the number of implicit transpositions in  $t$ . We therefore are interested in (pairwise) *disjoint* implicit transpositions, which do not share any DCJs between them. Furthermore, it is not immediately clear if existence of a set of  $m$  disjoint implicit transpositions in  $t$  implies that  $m$  transpositions can be simultaneously recovered from  $t$ , but we will prove below that this is indeed the case. We therefore define  $\text{DIT}(t)$  as the maximum number of disjoint implicit transpositions in  $t$ , which will be shown also equal to the maximum number of transpositions that can be simultaneously recovered from  $t$ .

### 4.1 Disjoint Implicit Transpositions as Matchings

It can be easily seen that an implicit transposition formed by a pair of DCJs  $(\alpha, \beta)$  in a shortest DCJ scenario  $t$  corresponds to an arc in the dependency graph  $\text{DG}(t)$ . However, it is not immediately clear if every arc  $(x, y)$  in  $\text{DG}(t)$  represents an implicit transposition, i.e., if DCJs  $x$  and  $y$  in  $t$  can be made adjacent with operations (T1). In this section, we prove that any matching  $M$  in  $\text{DG}(t)$  forms a disjoint collection of implicit transpositions that can be simultaneously recovered from  $t$ .

We call a graph  $G$  a *directed forest* if  $\overline{G}$  is a forest. By Theorem 6,  $\text{DG}(t)$  represents a directed forest for any shortest DCJ scenario  $t$ .

**Lemma 3.** *Let  $G$  be a directed forest. Then for any arc  $(\alpha_1, \alpha_2)$  in  $G$ , there exists a topological ordering of  $G$  in which  $\alpha_1$  and  $\alpha_2$  are adjacent.*

*Proof.* Let  $G'$  be a graph obtained from  $G$  by removing the arc  $(\alpha_1, \alpha_2)$  and gluing vertices  $\alpha_1, \alpha_2$  into a new single vertex  $\beta$ . That is, for any arc  $(\alpha_i, \gamma)$  with  $\gamma \neq \alpha_2$  in  $G$ , there is an arc  $(\beta, \gamma)$  in  $G'$ ; and for any arc  $(\gamma, \alpha_i)$  with  $\gamma \neq \alpha_1$  in  $G$ , there is an arc  $(\gamma, \beta)$  in  $G'$ .

We claim that  $G'$  is a directed forest. Indeed, the  $\overline{G'}$  can be viewed as the result of contraction of the edge  $(\alpha_1, \alpha_2)$  in  $\overline{G}$  into a single vertex  $\beta$ . Such contraction cannot create a cycle, i.e.,  $\overline{G'}$  remains to be a forest.

Let  $t'$  be a topological ordering of  $G'$ . By replacing the vertex  $\beta$  in  $t'$  with pair of adjacent vertices  $\alpha_1, \alpha_2$ , we obtain the required topological ordering of  $G$ .  $\square$

**Theorem 7.** *Let  $G$  be a directed forest. Then for any matching  $M$  in  $G$ , there exists a topological ordering  $t$  of  $G$  such that for any arc  $(\alpha_1, \alpha_2) \in M$ , DCJs  $\alpha_1$  and  $\alpha_2$  are adjacent in  $t$ .*

*Proof.* We prove the theorem statement by induction on  $|M|$ . For the base case  $|M| = 1$ , the statement follows from Lemma 3. Assume now that the statement holds for  $|M| = m$ .

For  $|M| = m + 1$ , let  $(\alpha_1, \alpha_2)$  be an arc in  $M$ . We construct the graph  $G'$  as described in the proof of Lemma 3 and let  $M' = M \setminus \{(\alpha_1, \alpha_2)\}$ . Since  $G'$  is a directed forest and  $|M'| = m$ , by the induction assumption there is a topological ordering  $t'$  of  $G'$  such that for any arc  $(\alpha_1, \alpha_2) \in M'$ , the DCJs  $\alpha_1$  and  $\alpha_2$  are adjacent in  $t'$ .

We obtain  $t$  from  $t'$  by replacing the vertex  $\beta$  with the ordered pair of vertices  $\alpha_1, \alpha_2$ . It is easy to see that such  $t$  represents the required topological ordering for  $G$ .  $\square$

## 4.2 Bounds for the Rate of Implicit Transpositions

**Theorem 8.** *Let  $t$  be a shortest DCJ scenario between genomes  $P$  and  $Q$  composed of the same  $n$  genes. Then  $\text{DIT}(t) \geq T_L(P, Q)$ , where*

$$T_L(P, Q) = \left\lceil \frac{n - 2 \cdot c(P, Q) + c_1(P, Q)}{4} \right\rceil.$$

*Proof.* We know that the graph  $\overline{\text{DG}}(t)$  is a forest (Theorem 6), where degree of each vertex is bounded by 4 (Theorem 4). Let us construct a matching  $M$  in  $\overline{\text{DG}}(t)$  iteratively. Initially we let  $G = \overline{\text{DG}}(t)$  and  $M = \emptyset$ .

If  $G$  contains at least one edge, it also contains a leaf (i.e., vertex of degree 1)  $\alpha$ . We add its only incident edge  $(\alpha, \beta)$  to  $M$  and remove from  $G$  all edges incident to the vertex  $\beta$ . Clearly, at most four such edges are deleted. We repeat this procedure until all edges of  $G$  are removed. By Theorem 4, the graph  $\overline{\text{DG}}(t)$  contains  $n - 2 \cdot c(P, Q) + c_1(P, Q)$  edges and thus we perform at least  $\left\lceil \frac{n - 2 \cdot c(P, Q) + c_1(P, Q)}{4} \right\rceil = T_L(P, Q)$  iterations, implying that  $|M| \geq T_L(P, Q)$ .



By construction, it is clear that  $M$  forms a matching in  $\overline{\text{DG}(t)}$  and thus under a suitable orientation of the edges in  $M$ , it also forms a matching in  $\text{DG}(t)$ . By Theorem 7, there exists a topological ordering  $t'$  of  $\text{DG}(t)$  such that the endpoints of all edges in  $M$  are adjacent in  $t'$ . By Theorem 5, topological ordering  $t'$  can be obtained from  $t$  with operations (T1), implying that we can simultaneously recover from  $t$  all elements of  $M$ . Therefore,  $\text{DIT}(t) \geq |M| \geq T_L(P, Q)$ .  $\square$

**Theorem 9.** *Let  $t$  be a shortest DCJ scenario between genomes  $P$  and  $Q$  composed of the same  $n$  genes. Then  $\text{DIT}(t) \leq T_U(P, Q)$ , where*

$$T_U(P, Q) = \frac{n - 2 \cdot c(P, Q) + c_{\text{odd}}(P, Q)}{2}.$$

*Proof.* There exist  $\text{DIT}(t)$  pairwise disjoint implicit transpositions in  $t$ , which after a number of operations (T1) can be made adjacent. We replace each pair of DCJs forming an implicit transposition with a 3-break to obtain a 3-break scenario  $\ell$ . The length of  $\ell$  is  $|\ell| = |t| - \text{DIT}(t) = n - c(P, Q) - \text{DIT}(t)$ . We remark that  $|\ell|$  is no smaller than the 3-break distance between genomes  $P$  and  $Q$ :

$$n - c(P, Q) - \text{DIT}(t) \geq d_3(P, Q) = \frac{n - c_{\text{odd}}(P, Q)}{2},$$

implying the required upper bound for  $\text{DIT}(t)$ .  $\square$

From the perspective of the upper bound in Theorem 9, an extreme case for genomes  $P$  and  $Q$  on  $n$  genes is  $c(P, Q) = c_{\text{odd}}(P, Q) = 1$ , in which by Theorem 2 there exists a shortest 3-break scenario between  $P$  and  $Q$  of length  $\frac{n-1}{2}$ . According to Lemma 2, such scenario contains no DCJs but only complete 3-breaks. By replacing each 3-break in this scenario with an equivalent pair of DCJs (forming an implicit transposition), we can get a DCJ scenario  $t$  between  $P$  and  $Q$  with  $2 \cdot \text{DIT}(t) = n - 1$  DCJs. It is easy to see that in this case we have  $\text{DIT}(t) = T_U(P, Q) = \frac{n-1}{2}$ , i.e., the upper bound for  $\text{DIT}(t)$  is tight.

Let  $t$  be a shortest DCJ scenario between genomes  $P$  and  $Q$ . There exist  $\text{DIT}(t)$  pairwise disjoint implicit transpositions in  $t$ . The pair of DCJs in each of these implicit transpositions can be made adjacent with operations (T1). If we replace each adjacent pair of DCJs forming an implicit transposition in the resulting scenario with an actual transposition (complete 3-break), then we obtain a scenario  $t'$  of length  $d_{\text{DCJ}}(P, Q) - \text{DIT}(t)$  composed of  $d_{\text{DCJ}}(P, Q) - 2 \cdot \text{DIT}(t)$  DCJs and  $\text{DIT}(t)$  transpositions. Thus the proportion of transpositions in  $t'$  is  $\frac{\text{DIT}(t)}{d_{\text{DCJ}}(P, Q) - \text{DIT}(t)}$ . We refer to this proportion as the *rate of implicit transpositions* in  $t$  and denote it by  $r(t)$ . The following theorem gives uniform bounds for  $r(t)$  that do not depend on a particular scenario  $t$ .

**Theorem 10.** *Let  $t$  be any shortest DCJ scenario between genomes  $P$  and  $Q$ . Then*

$$\frac{T_L(P, Q)}{d_{\text{DCJ}}(P, Q) - T_L(P, Q)} \leq r(t) \leq \frac{T_U(P, Q)}{d_{\text{DCJ}}(P, Q) - T_U(P, Q)}.$$

*Proof.* Since  $r(t) = \frac{\text{DIT}(t)}{d_{\text{DCJ}}(P,Q) - \text{DIT}(t)} = \frac{d_{\text{DCJ}}(P,Q)}{d_{\text{DCJ}}(P,Q) - \text{DIT}(t)} - 1$ , the value of  $r(t)$  monotonically increases as  $\text{DIT}(t)$  grows. The stated bounds for  $r(t)$  immediately follow from Theorems 8 and 9.  $\square$

## 5 Implicit Transpositions in Mammalian Evolution

Below we estimate the rate of implicit transpositions recovered from pair-wise DCJ scenarios between 6 mammalian species: mouse, rat, dog, macaque, human, and chimpanzee. Their gene orders and pairwise orthology relationship were obtained from Ensembl BioMart tool [10] on the following genomes: *Mus musculus* (GRCm38.p1), *Rattus norvegicus* (Rnor.5.0), *Canis familiaris* (CanFam3.1), *Homo sapiens* (GRCh37.p12), *Macaca mulatta* (MMUL.1.0), and *Pan troglodytes* (CHIMP2.1.4). Since our approach is currently limited to circular chromosomes, we artificially circularized chromosomes in each genome (such circularization is expected to have a minor impact [1]). For each pair of genomes, we represented them as sequences of shared genes present in one copy in each of these genomes<sup>2</sup> and used Theorem 10 to compute the lower and upper bounds for the rate of implicit transpositions between them. The results are given in Table 1.

**Table 1.** Estimation for the rate of implicit transpositions between pairs of genomes among mouse (M), rat (R), dog (D), macaque (Q), human (H), and chimpanzee (C).

Genome pairs	Shared genes	DCJ distance	Lower bound	Upper bound
M & R	14312	832	0.18	0.79
M & D	13852	1131	0.20	0.86
M & H	14119	1173	0.20	0.86
M & C	12608	1004	0.21	0.85
M & Q	13537	947	0.20	0.84
R & D	13312	1310	0.20	0.83
R & H	13352	1110	0.20	0.81
R & C	11942	961	0.21	0.80
R & Q	13064	1181	0.20	0.81
D & H	13808	1123	0.18	0.85
D & C	12372	984	0.19	0.85
D & Q	13449	1139	0.18	0.86
H & C	15646	929	0.17	0.93
H & Q	14408	973	0.17	0.91
C & Q	12912	884	0.18	0.91

<sup>2</sup> We remark that since the set of shared genes varies across genome pairs, the DCJ distances between genomes in different pairs are incomparable.

Table 1 demonstrates that the rate of implicit transpositions in mammalian evolution is at least 0.17. This is consistent and close to the estimate of the transposition rate in mammalian evolution as 0.26 recently obtained with statistical methods [3]. While we showed that the upper bound may be tight for some pairs of genomes, we believe that this not the case for mammalian genomes and the upper bound values in Table 1 appear to be superfluous.

## 6 Discussion

We continue our study of the combinatorial structure of DCJ scenarios from the perspective of simple length-preserving transformations, each affecting only a pair of consecutive DCJs (first introduced in [5]). Earlier we showed [9] that any shortest DCJ scenario between a genome with  $m \geq 1$  circular chromosomes and a linear genome (consisting of linear chromosomes) can be transformed into a shortest DCJ scenario, where circular chromosomes are eliminated by the first  $m$  DCJs and the rest represents a scenario between linear genomes. We further used this construction to obtain an approximate solution for the linear genome median problem.

In the current work, we study how evolutionary transpositions may implicitly appear in shortest DCJ scenarios and prove uniform bounds on their rate. Since transpositions are rather powerful rearrangements, it is not surprising that they may appear in a significant proportion that cannot be easily bounded in rearrangement scenarios between some genomes. Even though we do not yet have a recipe for limiting the effect of transpositions in the combined DCJ (2-break) and 3-break model (for which we earlier proved failure of the weighting approach [8]), our current study provides a step towards better understanding of the properties of transpositions and how they may affect reconstruction of the evolutionary history.

Our analysis of mammalian genomes demonstrates that the lower bound for the implicit transposition rate is close to the estimation obtained with statistical methods [3]. It is interesting to notice that the rate attains extreme values at pairs of primate genomes, which cannot be directly explained by the number of shared genes or DCJ distance between them and thus may indicate higher complexity of primate genomes with respect to the transposition analysis. While our approach is currently limited to circular chromosomes and applied for mammalian genomes with artificially circularized chromosomes, such circularization is expected to have a minor impact on the resulting estimates [1]. Extension of our approach to linear genomes will be published elsewhere.

**Acknowledgments.** The work was supported by the National Science Foundation under the grant No. IIS-1462107.

## References

1. Alekseyev, M.A.: Multi-break rearrangements and breakpoint re-uses: from circular to linear genomes. *J. Comput. Biol.* **15**(8), 1117–1131 (2008)

2. Alekseyev, M.A., Pevzner, P.A.: Multi-break rearrangements and chromosomal evolution. *Theor. Comput. Sci.* **395**(2), 193–202 (2008)
3. Alexeev, N., Aidagulov, R., Alekseyev, M.A.: A computational method for the rate estimation of evolutionary transpositions. In: Ortuño, F., Rojas, I. (eds.) *IWBBIO 2015, Part I. LNCS*, vol. 9043, pp. 471–480. Springer, Heidelberg (2015)
4. Bader, M., Ohlebusch, E.: Sorting by weighted reversals, transpositions, and inverted transpositions. *J. Comput. Biol.* **14**(5), 615–636 (2007)
5. Braga, M.D., Stoye, J.: The solution space of sorting by DCJ. *J. Comput. Biol.* **17**(9), 1145–1165 (2010)
6. Eriksen, N.:  $(1 + \epsilon)$ -approximation of sorting by reversals and transpositions. In: Gascuel, O., Moret, B.M. (eds.) *WABI 2001. LNCS*, vol. 2149, pp. 227–237. Springer, Heidelberg (2001)
7. Fertin, G., Labarre, A., Rusu, I., Tannier, E., Vialette, S.: *Combinatorics of Genome Rearrangements*. MIT Press, Cambridge (2009)
8. Jiang, S., Alekseyev, M.A.: Weighted genomic distance can hardly impose a bound on the proportion of transpositions. In: Bafna, V., Sahinalp, S.C. (eds.) *RECOMB 2011. LNCS*, vol. 6577, pp. 124–133. Springer, Heidelberg (2011)
9. Jiang, S., Alekseyev, M.A.: Linearization of median genomes under DCJ. In: Brown, D., Morgenstern, B. (eds.) *WABI 2014. LNCS*, vol. 8701, pp. 97–106. Springer, Heidelberg (2014)
10. Kasprzyk, A.: BioMart: driving a paradigm change in biological data management. *Database* 2011, bar049 (2011)
11. Ouangraoua, A., Bergeron, A.: Combinatorial structure of genome rearrangements scenarios. *J. Comput. Biol.* **17**(9), 1129–1144 (2010)
12. Ozery-Flato, M., Shamir, R.: Sorting by translocations via reversals theory. In: Bourque, G., El-Mabrouk, N. (eds.) *RECOMB-CG 2006. LNCS (LNBI)*, vol. 4205, pp. 87–98. Springer, Heidelberg (2006)
13. Ranz, J., González, J., Casals, F., Ruiz, A.: Low occurrence of gene transposition events during the evolution of the genus *Drosophila*. *Evolution* **57**(6), 1325–1335 (2003)
14. Shao, M., Lin, Y., Moret, B.: Sorting genomes with rearrangements and segmental duplications through trajectory graphs. *BMC Bioinform.* **14**(15), 1–8 (2013)
15. Yancopoulos, S., Attie, O., Friedberg, R.: Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics* **21**(16), 3340–3346 (2005)

Algorithms for Computational Biology

Second International Conference, AlCoB 2015, Mexico

City, Mexico, August 4-5, 2015, Proceedings

Dediu, A.-H.; Hernández-Quiroz, F.; Martin-Vide, C.;

Rosenblueth, D.A. (Eds.)

2015, X, 155 p. 49 illus., 2 illus. in color., Softcover

ISBN: 978-3-319-21232-6