Implicit Transpositions in DCJ Scenarios

Pavel Avdeyev
The George Washington University
avdeyev@gwu.edu

Shuai Jiang
University of South Carolina

Max A. Alekseyev
The George Washington University

ABSTRACT

Background. Genome rearrangements are dramatic evolutionary events that change genome structures. The number of genome rearrangements between two genomes represents a good measure for their evolutionary closeness and is used as such in phylogenomic studies. The most common rearrangements are *reversals* that inverse contiguous segments of chromosomes, *translocations* that exchange tails of two chromosomes, and *fissions/fusions* that split/merge chromosomes. All these rearrangements can be conveniently modeled by Double-Cut-and-Join (DCJ) operations, which make up to 2 “cuts” in a genome and “glue” the resulting genomic fragments in a new order. *Transpositions* represent yet another type of genome rearrangements that relocate genomic segments across the genome. 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* and pose a question of how many transpositions can be *simultaneously* recovered from a given DCJ scenario by shuffling DCJs and replacing suitable pairs of consecutive DCJs with transpositions. We consider both *shortest* DCJ scenarios resulting from the maximum parsimony assumption, and more general *proper* DCJ scenarios based on certain realistic but less restrictive assumptions.

Methods. For genomes P and Q composed of the same set of genes, the *breakpoint graph* $G(P, Q)$ defined as a graph on the gene extremities as vertices and edges of two colors encoding genes adjacencies in genomes P and Q . It represents a collection of cycles and paths consisting of undirected edges alternating between the two colors. We distinguish the following types of cycles and paths with respect to their *length* ℓ (i.e., the number of edges in a cycle or path): trivial cycles and paths ($\ell = 2$), even paths (ℓ is even) and odd paths (ℓ is odd). We denote the number of cycles, trivial cycles, paths, trivial paths, even paths, and odd paths in $G(P, Q)$ as $c(P, Q)$, $c_2(P, Q)$, $p(P, Q)$, $p_2(P, Q)$, $p_{\text{even}}(P, Q)$, and $p_{\text{odd}}(P, Q)$, respectively.

A DCJ scenario transforming genome P into 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 and trivial paths. It is well known that the *DCJ distance* (i.e., the length of a shortest DCJ scenario) between genomes P and Q on n genes equals $d_{\text{DCJ}}(P, Q) = n - c(P, Q) - \frac{p_{\text{even}}(P, Q)}{2}$.

Simultaneously recovering the maximum number of transpositions m from a DCJ scenario t , we will obtain a scenario of length $|t| - m$ composed of m transpositions and $|t| - 2m$ DCJs. The proportion of transpositions in this scenario is $r(t) = \frac{m}{|t| - m}$, which

we refer to as the *rate of implicit transpositions* in t . Since there exist many different shortest/proper DCJ scenarios between two genomes, it is important to derive a lower bound for $r(t)$ that does not depend on t , but only on the given genomes.

Results and Evaluation. We prove that for any shortest DCJ scenario t between genomes P and Q ,

$$r(t) \geq \frac{\lceil E(P, Q)/4 \rceil}{d_{\text{DCJ}}(P, Q) - \lceil E(P, Q)/4 \rceil},$$

where $E(P, Q) = n - 2 \cdot c(P, Q) - p(P, Q) - \frac{p_{\text{even}}(P, Q)}{2} + c_2(P, Q) + p_2(P, Q)$. Similarly, for any proper DCJ scenario t between genomes P and Q ,

$$r(t) \geq \frac{1}{2} - \frac{3s(P, Q)}{8d_{\text{DCJ}}(P, Q) + 2s(P, Q)},$$

where $s(P, Q) = n + p_{\text{odd}}(P, Q) + \frac{p_{\text{even}}(P, Q)}{2} - c_2(P, Q) - p_2(P, Q)$. The obtained bounds imply that implicit appearance of transpositions in DCJ scenarios may be unavoidable or even abundant for some pairs of genomes.

We analyzed a set of three mammalian genomes: *rat* (R), *macaque* (M), and *human* (H) represented as sequences of 1,360 synteny blocks, and demonstrated that the obtained lower bound for $r(t)$ is consistent with the existing statistical estimation for the transposition rate:

Genome pair	DCJ distance	Bound for $r(t)$, t is proper	Bound for $r(t)$, t is shortest	Estimated rate of transpositions
H & M	106	0.06	0.10	0.25
H & R	707	0.11	0.17	0.26
M & R	701	0.10	0.17	0.28

We also analyzed a set of five yeast genomes represented as sequences of the same 710 synteny blocks, and observed that the rate of implicit transpositions in DCJ scenarios between these genomes is at least 0.06.

Conclusion. The present study provides a step towards better understanding of the properties of transpositions and how they may affect reconstruction of the evolutionary history. In the future work, we plan to extend our method to support other evolutionary events such as gene deletions/insertions and duplications. This will increase the accuracy and make the method applicable to genomes (such as plants) whose evolutionary history is rich in such events.

Acknowledgements. The work is supported by the National Science Foundation under the grant No. IIS-1462107 and published in *Front. Genet.* 8 (2017), 212. <https://doi.org/10.3389/fgene.2017.00212>

CCS CONCEPTS

• **Mathematics of computing** → **Graph theory**; • **Applied computing** → **Bioinformatics**; *Computational genomics*;

ACM Reference Format:

Pavel Avdeyev, Shuai Jiang, and Max A. Alekseyev. 2018. Implicit Transpositions in DCJ Scenarios. In *ACM-BCB'18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, August 29-September 1, 2018, Washington, DC, USA*. ACM, New York, NY, USA, 1 page. <https://doi.org/10.1145/3233547.3233724>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA

© 2018 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5794-4/18/08.

<https://doi.org/10.1145/3233547.3233724>