

# Calibrating Genomic Distance via Universal Operation

Richard Friedberg<sup>1</sup>, Oliver Attie<sup>2</sup> and Sophia Yancopoulos<sup>3</sup>

**Keywords:** genomic distance, translocation, inversion, reversal, transposition, sorting permutations

## 1 Introduction.

Large-scale sequencing projects increasingly enable comparison of species on the genome level. Phylogenetic tree analysis, traditionally based on individual genes, can currently be conducted on the basis of gene order rearrangements. Although a variety of processes are widely understood to contribute to genomic evolution, the challenge has been to find a consistent and biologically valid set of operations since no consensus has yet been reached on which set, if any, would be definitive.

The problem of transforming one genome to another via a given set of operations, e.g.: sorting by reversals (SBR) [1], is well established in the field of genome rearrangements, and a number of classical solutions have been obtained for sorting genomic permutations by a variety of operations.

## 2 The Double Cut and Join Operation.

Our elementary operation, “double cut and join” [DCJ], is a local operation on four points, which correspond to gene (or synteny block) ends, connected in pairs. It consists of cutting the connections between them and rejoining to form two new pairs; which is possible in two ways as shown in Fig. 1.

We give this operation weight 1, regardless of the gene end, whether the two initial pairs are on the same chromosome or not, or which of the two ways we rejoin.

Transformation of genome A to genome B by successive DCJ operations can, somewhat surprisingly, be recast as the sorting of a permutation by pair exchange. The number of elements to be sorted equals the number of breakpoints. A cycle length  $L$  in the permutation corresponds to a cycle of Hannenhalli and Pevzner [2] containing  $L$  black lines alternating with  $L$  gray lines. The resulting genomic distance is given by  $b-c$  since the problem is equivalent to sorting a permutation by pairs, and hurdles don't enter into our formulation.

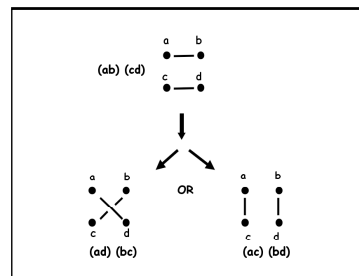


Figure 1. A general DCJ Operation.

<sup>1</sup> Department of Physics, Columbia University, NY, NY, USA. E-mail: rfriedberg1@nyc.rr.com

<sup>2</sup> Center for the Study of Gene Structure and Function, Hunter College, NY, NY, USA. E-mail: oattie@GENECTR.HUNTER.CUNY.EDU

<sup>3</sup> Institute for Medical Research, North Shore-LIJ Health System, Manhasset, NY, USA. E-mail: syancopo@nshs.edu

### 3 The Menu of Operations on Linear Chromosomes

Depending on the chromosomal context, we identify the DCJ operation with familiar operations. The cutting of a cycle in two places in single chromosome genomes admits two ways of rejoining: one reverses a segment of the chromosome and the other snips out a circular fragment. We identify this fragment with a transposon, and its creation and absorption as a *generalized transposition*, which we give a weight of 2. For linear chromosomes, we arrive at the following menu of operations:

- Translocations (including fission and fusion): weight 1.
- Inversions: weight 1.
- Creation (weight 1) and immediate absorption (weight 1) of a transposon.

### 4 An Optimal Algorithm to Obtain a Rearrangement Scenario.

We perform the different kinds of operations in a definite order, decreasing  $b - c$  by eliminating at least one breakpoint at each step:

- At the outset identify particular end-blocks in genome A with certain end-blocks in genome B, introducing null chromosomes if necessary in either genome, whose end-blocks are identified with particular end-blocks in the other genome. Each end-block appears once in each genome, and each cycle contains no more than two end-blocks. The choice of identification is optimal.
- Perform fissions and fusions on genome A until each end-block is connected either to the same gene end (we use "gene" for simplicity, could also be a synteny block) or end-block as in genome B.
- Perform translocations until every cycle is confined to a single chromosome in each genome.
- Perform reversals according to the "score" prescription of Bergeron [3] until all genes have the same direction as in genome B. (What remains is a permutation with only "positive" terms.)
- Complete the unscrambling by transposon creation and absorption (or *block-interchange* [4].)

We compare our values for genomic distances to those obtained by Bourque *et al.* [5] using macrorearrangement synteny blocks greater than 300 kb for human, mouse and rat.

### References

- [3] Bergeron, A. 2005. A Very Elementary Presentation of the Hannenhalli-Pevzner Theory. *Discrete Applied Mathematics* **146** (2):134-145.
- [5] Bourque, G., Pevzner, P. and Tesler, G. 2004. Reconstructing the genomic architecture of ancestral mammals: Lessons from human mouse, and rat genomes. *Genome Research* **14** (4) pp. 507-516
- [1] Caprara, A. 1997. Sorting by reversals is difficult. In: *Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 97)*, New York: ACM. pp. 75-83.
- [4] Christie, D.A. 1996. Sorting permutations by block interchanges. *Information Processing Letters*, 60:165-169
- [2] Pevzner, P.A. 2000. *Computational Molecular Biology, An Algorithmic Approach*, MIT Press. Chapter 10.