A Unifying View of Genome Rearrangements

Anne Bergeron¹, Julia Mixtacki², and Jens Stoye³

¹ Comparative Genomics Laboratory, Université du Québec à Montréal, Canada
bergeron.anne@uqam.ca
² International NRW Graduate School in Bioinformatics and Genome Research, Universität Bielefeld, Germany
julia.mixtacki@uni-bielefeld.de
³ Technische Fakultät, Universität Bielefeld, Germany
stoye@techfak.uni-bielefeld.de

Abstract. Genome rearrangements have been modeled by a variety of operations such as inversions, translocations, fissions, fusions, transpositions and block interchanges. The double cut and join operation, introduced by Yancopoulos et al., allows to model all the classical operations while simplifying the algorithms. In this paper we show a simple way to apply this operation to the most general type of genomes with a mixed collection of linear and circular chromosomes. We also describe a graph structure that allows simplifying the theory and distance computation considerably, as neither capping nor concatenation of the linear chromosomes are necessary.

1 Introduction

The problem of sorting multichromosomal genomes can be stated as: Given two genomes A and B, the goal is to find a shortest sequence of rearrangement operations that transforms A into B. The length of such a shortest sequence is called the distance between A and B. Clearly, the solutions depend on what kind of rearrangement operations are allowed.

Given their prevalence in eukaryotic genomes [1], the usual choices of operations include translocations, fissions, fusions and inversions. However, there are some indications that transpositions should also be included in the set of operations [2], but the lack of theoretical results showing how to include transpositions in the models led to algorithms that simulate transpositions as sequences of inversions.

In [3], the authors describe a general framework in which circular and linear chromosomes can coexist throughout evolving genomes. They model inversions, translocations, fissions, fusions, transpositions and block interchanges with a single operation, called the double cut and join operation. This general model accounts for the genomic evidence of the coexistence of both linear and circular chromosomes or plasmids in many genomes [4,5].

In this paper, we present a simplified formalization of genomes with coexisting circular and linear chromosomes, and a formal treatment of sorting such genomes by the double cut and join operation. We introduce a very simple data structure,
the adjacency graph, that is symmetric with respect to the two genomes under study and is closely related to the visual picture of the genomes themselves. We also show how the algebraic simplicity of the double cut and join operation yields efficient sorting algorithms that can be tailored to optimize the use of certain types of operations.

2 Notes on Graphs with Vertices of Degree One or Two

An essential ingredient in genome rearrangement studies are graphs where each vertex has degree one or two. Here we recall some of their properties.

Let $G$ be a graph where each vertex has degree one or two. We call a vertex of degree one external and a vertex of degree two internal. An internal vertex connecting edges $p$ and $q$ is denoted by the unordered multiset $\{p, q\}$ and an external vertex incident to an edge $p$ by the singleton set $\{p\}$.

It follows immediately from the definition of $G$ that any connected component of $G$ is either circular, consisting only of internal vertices, or it is linear, consisting of internal vertices bounded by two external vertices, one at each end. We denote circular components as cycles and linear components as paths. A cycle or path is even if it has an even number of edges, otherwise it is odd.

Example 1. The following graph has four vertices of degree one and six vertices of degree two. It has two cycles and two paths, one of which is even and one of which is odd.

![Graph with vertices of degree one or two](image)

Definition 1. The double cut and join (DCJ) operation acts on two vertices $u$ and $v$ of a graph with vertices of degree one or two in one of the following three ways:

(a) If both $u = \{p, q\}$ and $v = \{r, s\}$ are internal vertices, these are replaced by the two vertices $\{p, r\}$ and $\{s, q\}$ or by the two vertices $\{p, s\}$ and $\{q, r\}$.

(b) If $u = \{p, q\}$ is internal and $v = \{r\}$ is external, these are replaced by $\{p, r\}$ and $\{q\}$ or by $\{q, r\}$ and $\{p\}$.

(c) If both $u = \{q\}$ and $v = \{r\}$ are external, these are replaced by $\{q, r\}$.

In addition, as an inverse of case (c), a single internal vertex $\{q, r\}$ can be replaced by two external vertices $\{q\}$ and $\{r\}$.

Figure 1 illustrates the definition.

The DCJ operation, although defined locally on a pair of vertices, has global effects on the connected components of the graph. In order to describe these, we use a terminology essentially borrowed from biology.

First, consider Figure 2. If the two vertices are contained in two different paths and at least one of them is internal, then these paths exchange their ends, which