Universität Bielefeld
Technische Fakultät

AG Genominformatik
Prof. Dr. Jens Stoye
Dr. habil. Marília D. V. Braga

Algorithms in Genome Research

Winter 2025/2026

Exercises

Number 9, Discussion: 2026-January-16

1. Pangenome openness.

   (a) What is an *open* pangenome and what is a *closed* pangenome?

   (b) Schematically, how do an open and a closed pangenome look like

   - as a Venn diagram in gene-based pangenomics?
   - as a pangenome graph (e.g. variation graph or colored de Bruijn graph) in genome-based pangenomics?

   (c) Why is is better to speak only of the *openness* of a pangenome?

2. Construct the positional Burrows Wheeler Transformation (pBWT) after processing the following six binary strings (representing genomic haplotypes):

$$\begin{aligned}
s_1 &= \mathtt{10101010000101011}\\
s_2 &= \mathtt{01101101000110001}\\
s_3 &= \mathtt{01001101000101011}\\
s_4 &= \mathtt{10101111000111001}\\
s_5 &= \mathtt{01101001000101011}\\
s_6 &= \mathtt{10001010000101011}
\end{aligned}$$

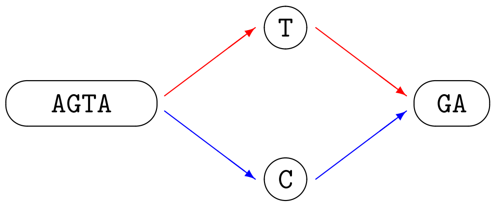
   (a) Is there a pronounced recombination site visible?

   (b) What is the largest haplotype block ending at the end of this genomic region?

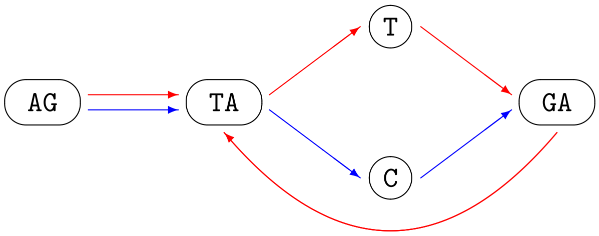
3. Two popular data structures to represent a genome-based pangenome are the variation graph and the colored de Bruijn graph.

   (a) Given the following two variation graphs, find compacted colored de Bruijn graphs of dimension $k = 3$ that contain the same sets of strings.

(a.1)



(a.2)



(b) Given the following three “genome” sequences. Construct their compacted colored de Bruijn graph of dimension $k = 4$.

CAGGATCAGAACGGC
GGACCCAGGATAGA
AGGACCCATAGAACGGC

Find a variation graph that represents the same set of strings.

4. Develop the details of an algorithm that takes as input a variation graph G and a query sequence S, and finds a position in G where an optimal (unit-cost) semi-global alignment of S and any (sub)string represented in G ends.

   Note: First consider that G is a directed acyclic graph (DAG). Then generalize your algorithm to the case where G may contain cycles.