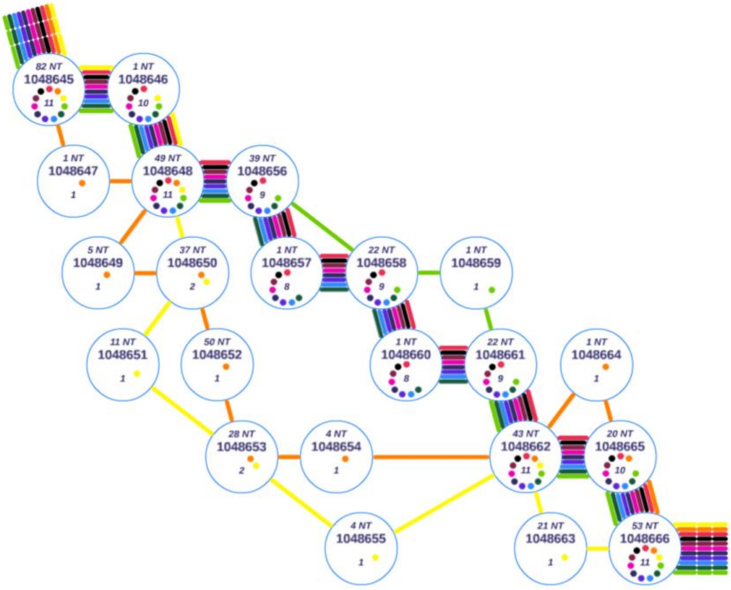
GENOME GRAPH SEQUENCE ALIGNMENT AND MISMATCH DETECTION

Responsible supervisor: Thomas Abeel – t.abeel@tudelft.nl - EEMCS Pattern Recognition and Bioinformatics

Background

Pan-genome references, also known as graph reference genomes, can represent multiple genome sequences in a data structure. While this is conceptually a great idea, there is a lack of tools to make them practically useable. In this project, we aim to tackle the two most important algorithms for DNA sequence analysis on this new data structure. Assuming we have a graph reference, we would like to align where billions of short strings align within this graph, what their mismatches and whether consensus mismatches can be found pointing towards novel DNA mutations.



The goal of this proposal is to develop algorithms for (short) DNA sequence alignment on genome graphs. On the one hand there is the challenge to do this for the billions of tiny fragments (<300 length) that get produced by DNA sequencing machines, which have relatively low error rate (<1%). While on the other hand new DNA measuring technologies provide another challenge with high error rates (~10%) with much larger lengths (>10kb).

These methods could be develop based on classical sequence analysis algorithms or based on machine learning approaches.