Alignment of whole genomes using MUMmer

Presentation in Algorithms in Bioinformatics (TÖ111F autumn 2014)

Hannes Pétur Eggertsson

November 18, 2014

Motivation

In 1999...

- the number of sequenced genomes was increasing rapidly
- when a new genome is sequenced, one could ask himself:
  - How does this genome align to the other genomes?

Problem:

- We have algorithms that were used for single gene sequences (up to 10,000 bp)
- But, they won't work well with whole genomes (millions of base pairs or more)
  - Take up way too much memory or
  - Have unacceptable computational time

Introduction

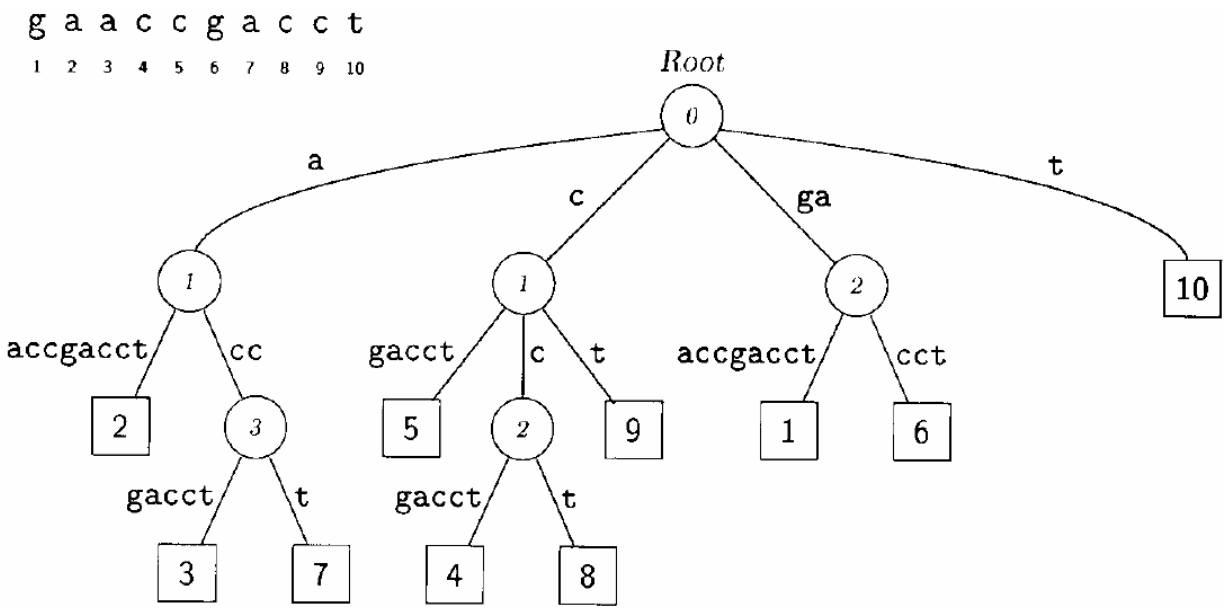
MUMmer:

- Is a system used to align whole genome sequences.
- Uses suffix trees as a data structure.
- Assumes that the two genomes are similar/related.
- The algorithm can be split into several steps that I'll talk about in detail

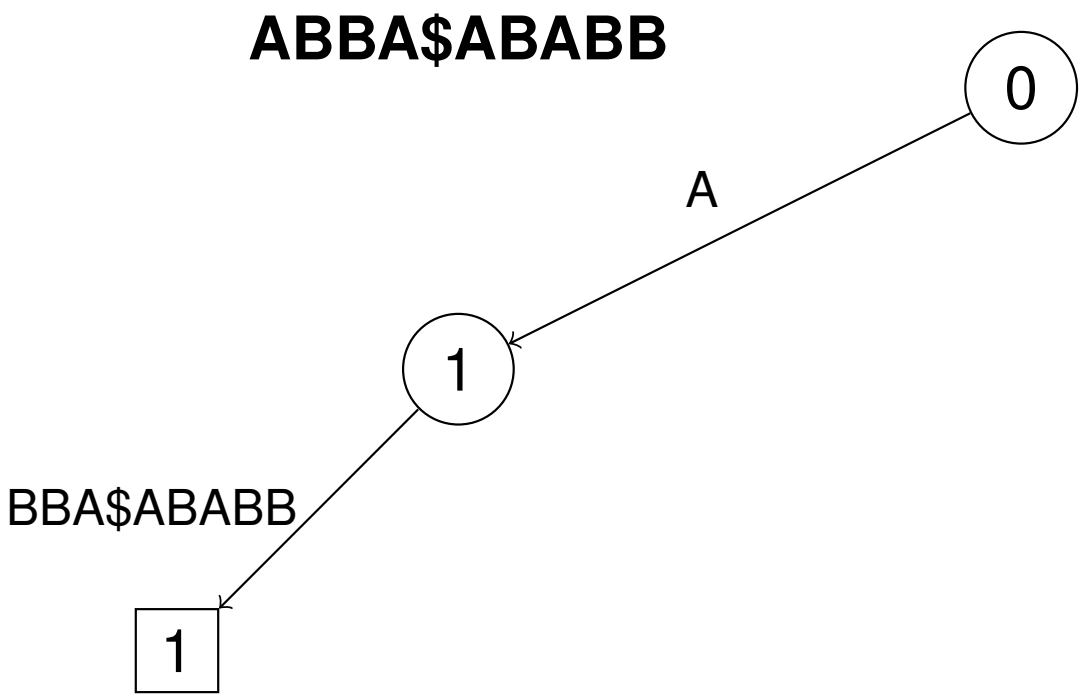
Step 1: Creating a suffix tree

A suffix tree is a compact representation that stores all possible suffixes of an input sequence.

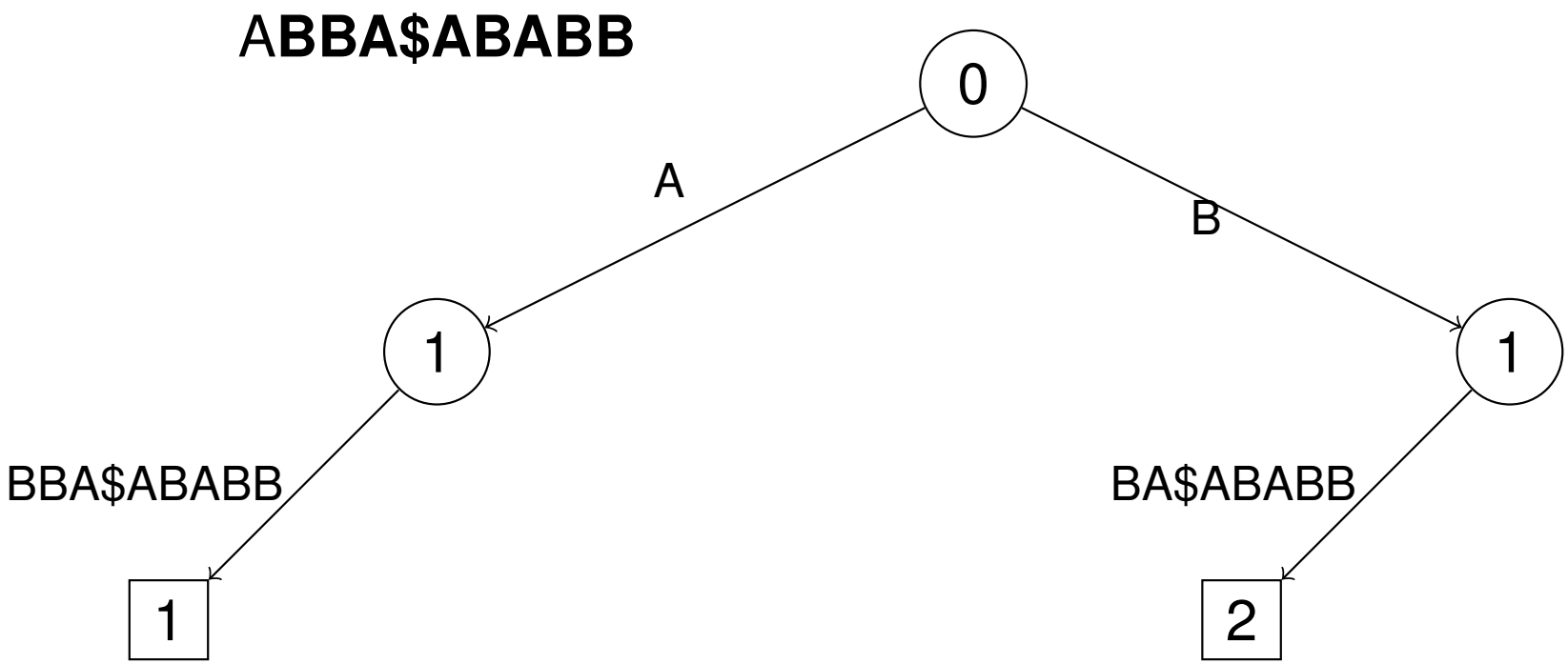
- Square nodes are leaves.
  - Store information about the starting position of the suffix.
- Circular nodes are internal nodes.
  - That means two or more sequences share the same prefix.
  - Store information about the length of the shared prefix.



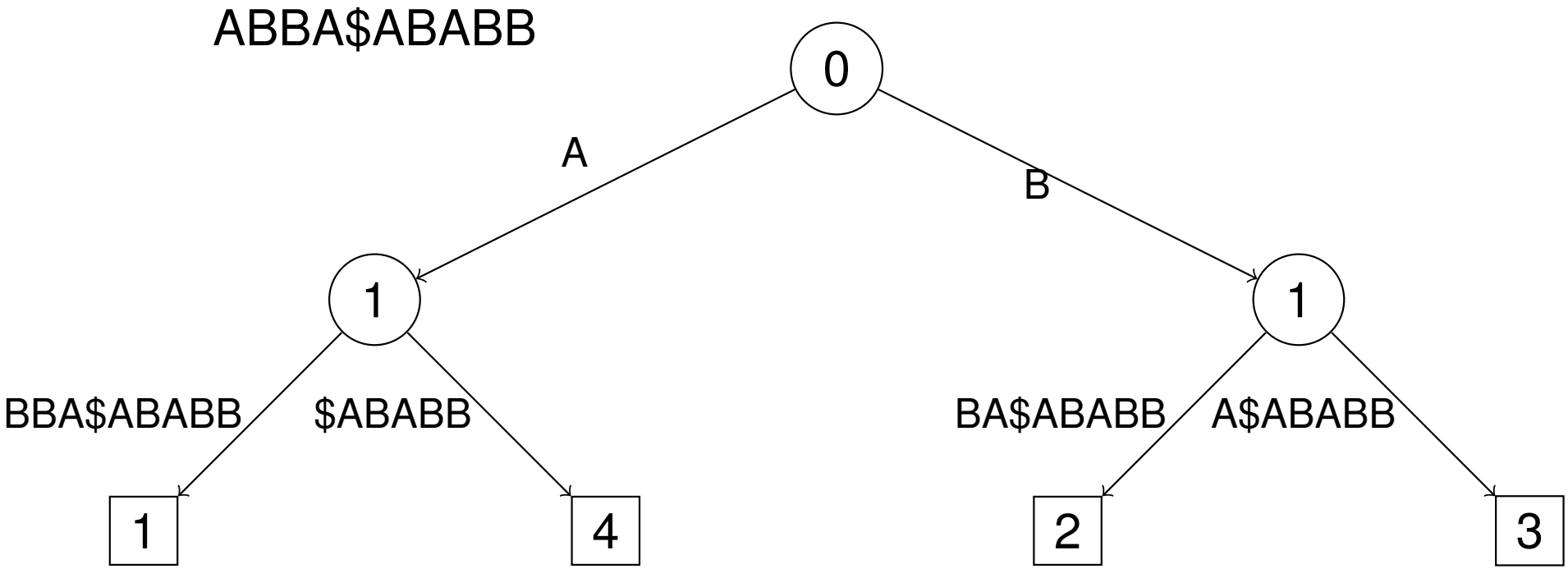
Step 1: Creating a suffix tree



Step 1: Creating a suffix tree



Step 1: Creating a suffix tree



Step 2: MUM decomposition

Let us first define what a MUM is

Definition

A subsequence is a MUM (Maximal Unique Matches) if and only if:

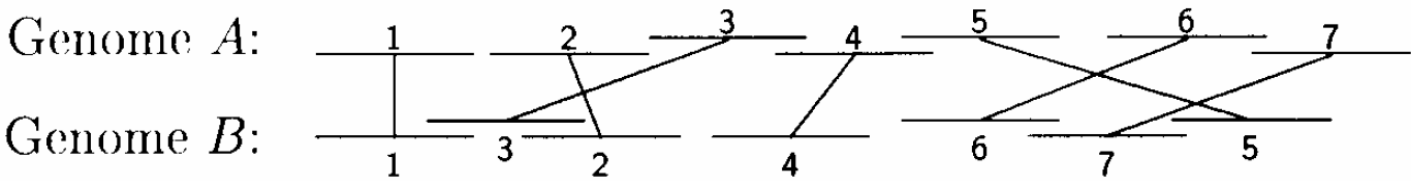
- The subsequence has a exact match on both genomes
- It is not a subsequence of another matched sequence
  - This means the sequence is surrounded be mismatches
- It is unique
  - It appears exactly once in both genomes

Example

```
tcgatcGACGATCGCGGCCGTAGATCGAATAACGAGAGAGCATAAcgactta
gcattaGACGATCGCGGCCGTAGATCGAATAACGAGAGAGCATAAtccagag
```

Step 3: Sorting the matches found in the MUM alignment

Once we have found the MUMs, we might end up with something like this:



Problem! We cannot align the two genomes using the MUMs because they aren't in correct order.

Step 4: Closing the gaps

MUMmer 3

Conclusion