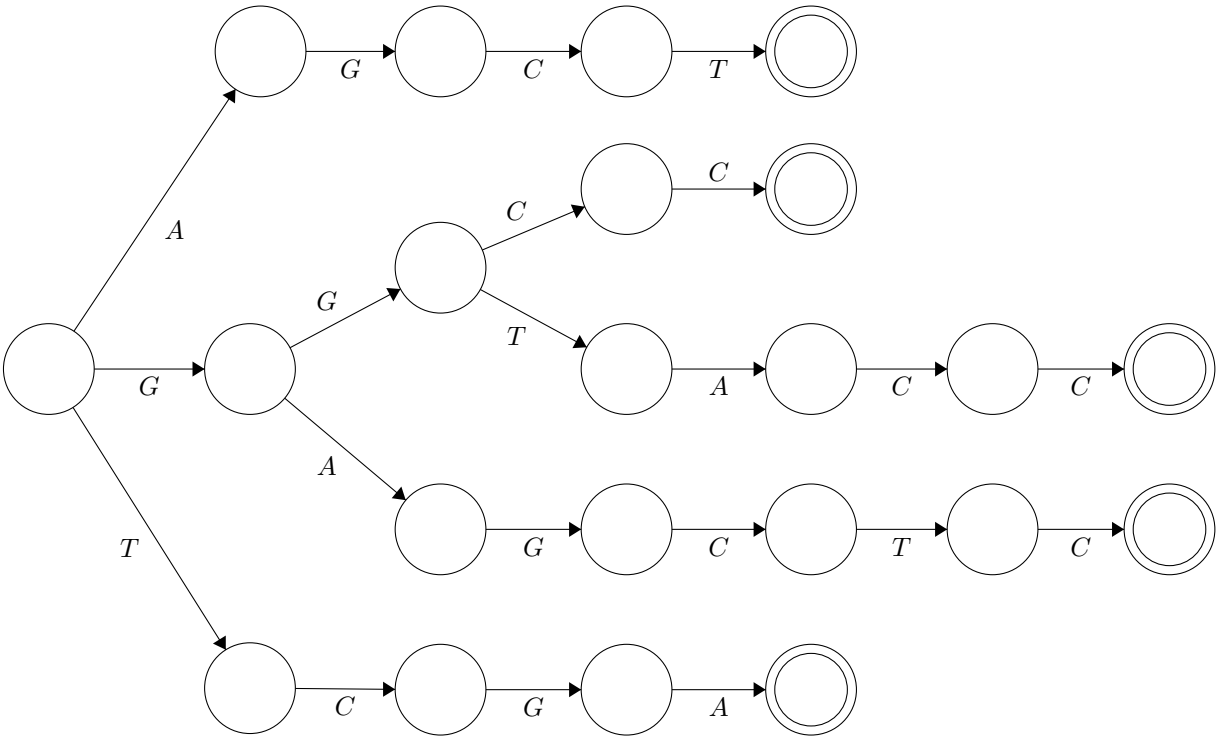
Sequence Bioinformatics
Assignment 05

Emil Paulitz, Raphael Olipitz

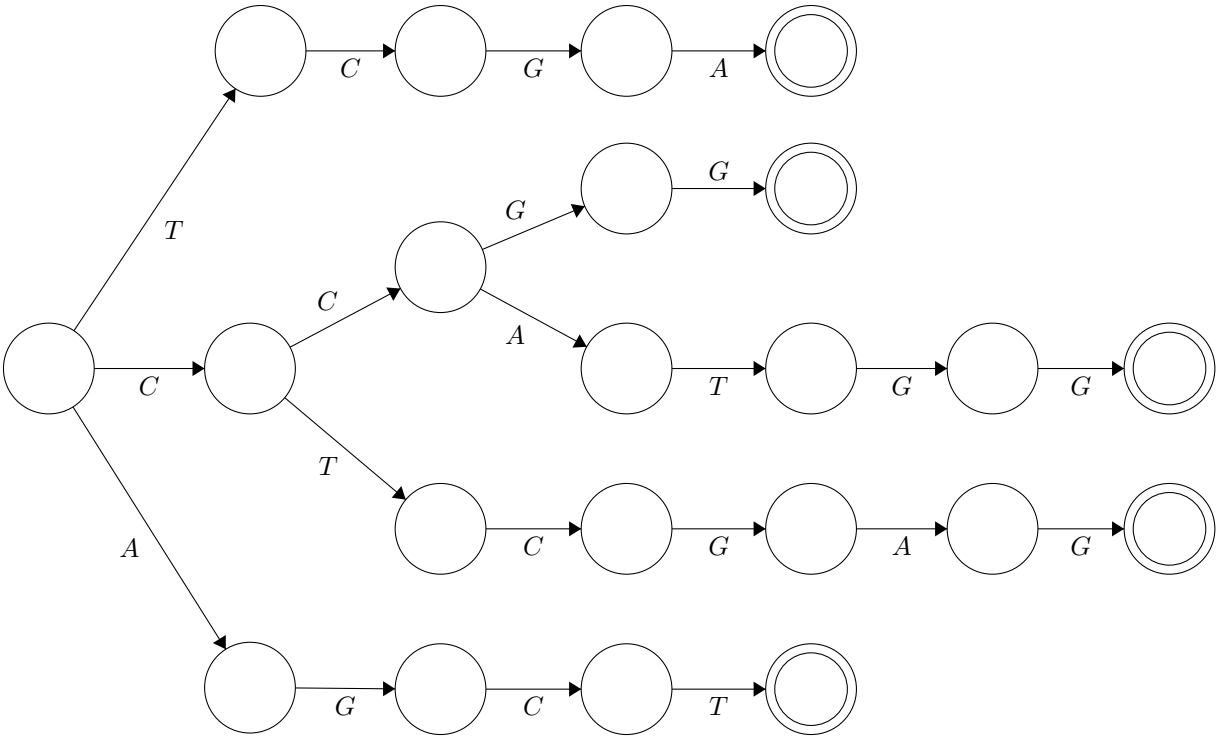
December 15, 2020

Task 1

Forward Trie:

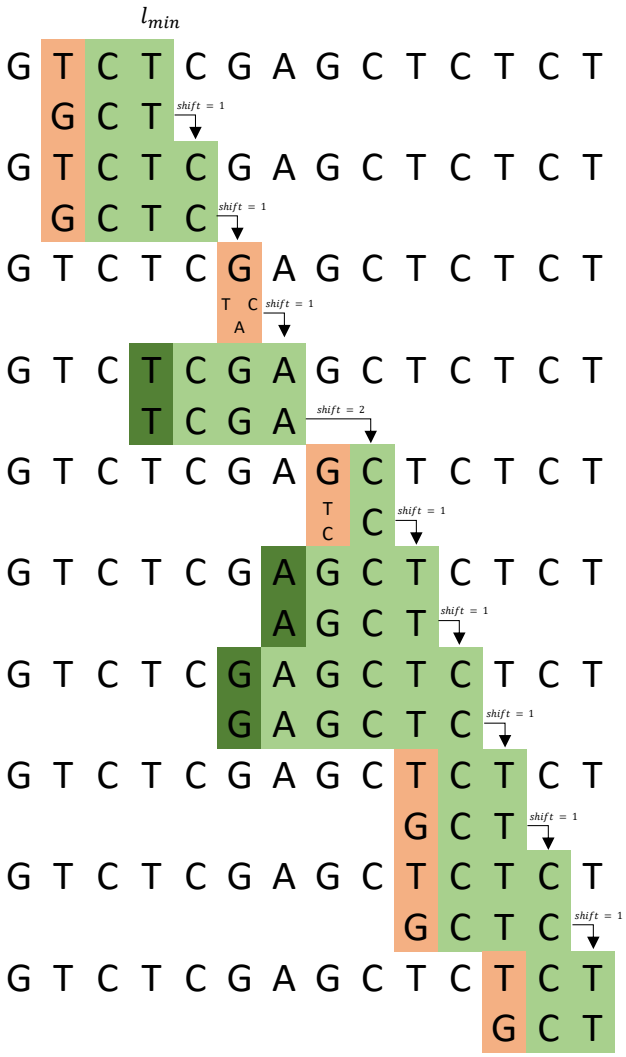


Reverse Trie:



$l_{\min} = 4$
$d(A) = 2$
$d(C) = 1$
$d(G) = 1$
$d(T) = 1$

Task 2

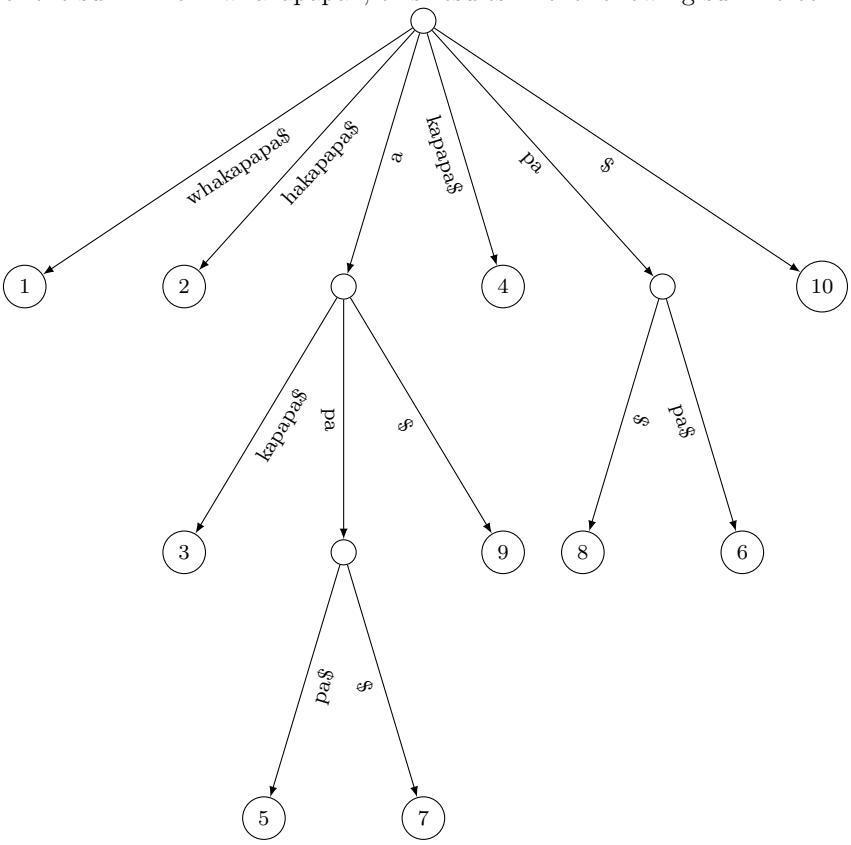


If the algorithm can check with a single lookup (using keys for example) whether there exists an edge with the desired character, it performs $3 + 4 + 1 + 4 + 2 + 4 + 6 + 3 + 4 + 3 = 34$ letter comparisons.

Task 3

One iterates over every suffix, starting with the longest one (the whole word). For every suffix, one starts at the root and tries to match the current suffix with

the entries on the existing edges (for the first suffix, no edges exist). If this fails, one starts a new branch at the position in the tree no match was found (for the first suffix, the root), which then ends in a leaf that holds the starting position of the suffix. For "whakapapa", this results in the following suffix tree:



Task 4

To find all occurrences of the search query 'pa', one would start from the root and

1. compare the first letter on every outgoing node with the first letter of the query. In this case, one would search for an outgoing branch that has a label beginning with 'p'.
2. If one is found, the other letters are compared with the branch's label, in our case 'a' matches.
3. Because the query is finished, one follows every path to a leaf and reports the starting point of this suffix, recorded at the leaf: 7 and 5.