Problem 1. Pattern Matching.

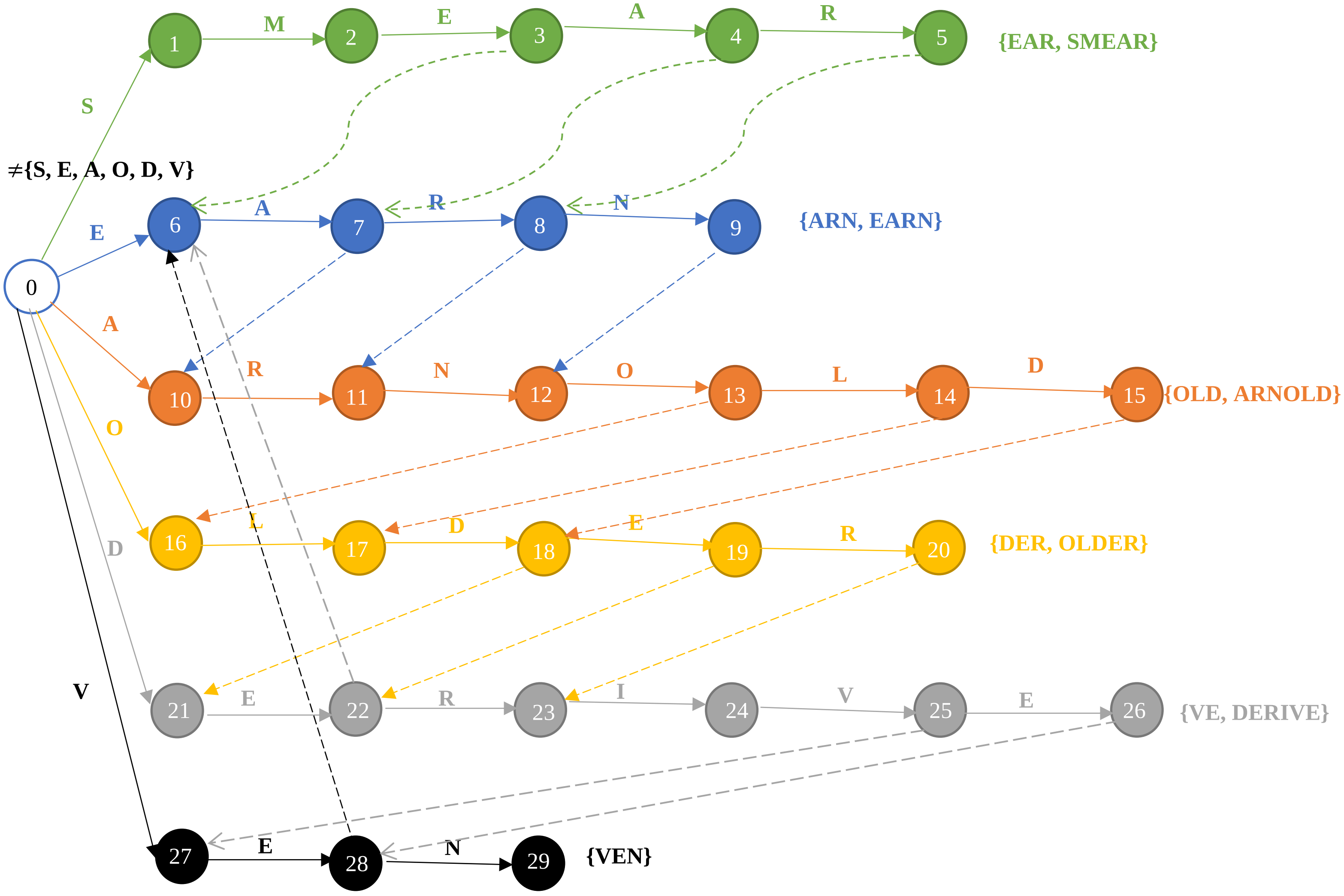
30 points.

Build a keyword tree, with fail edges (*as in the Aho-Corasick algorithm*) for the following dictionary. Mark with dashed arrows the "fail edges" that do not go back to the root. That is, you do not need to show fail edges going back to the root. You may but are not required to mark the nodes that correspond to pattern matches.

Count and note down the number of fail edges your keyword tree has. (Again, only the fail edges you marked, i.e., fail edges that do not go back to the root.)

Number of fail edges: 16

SMEAR
EARN
ARNOLD
OLDER
DERIVE
VEN

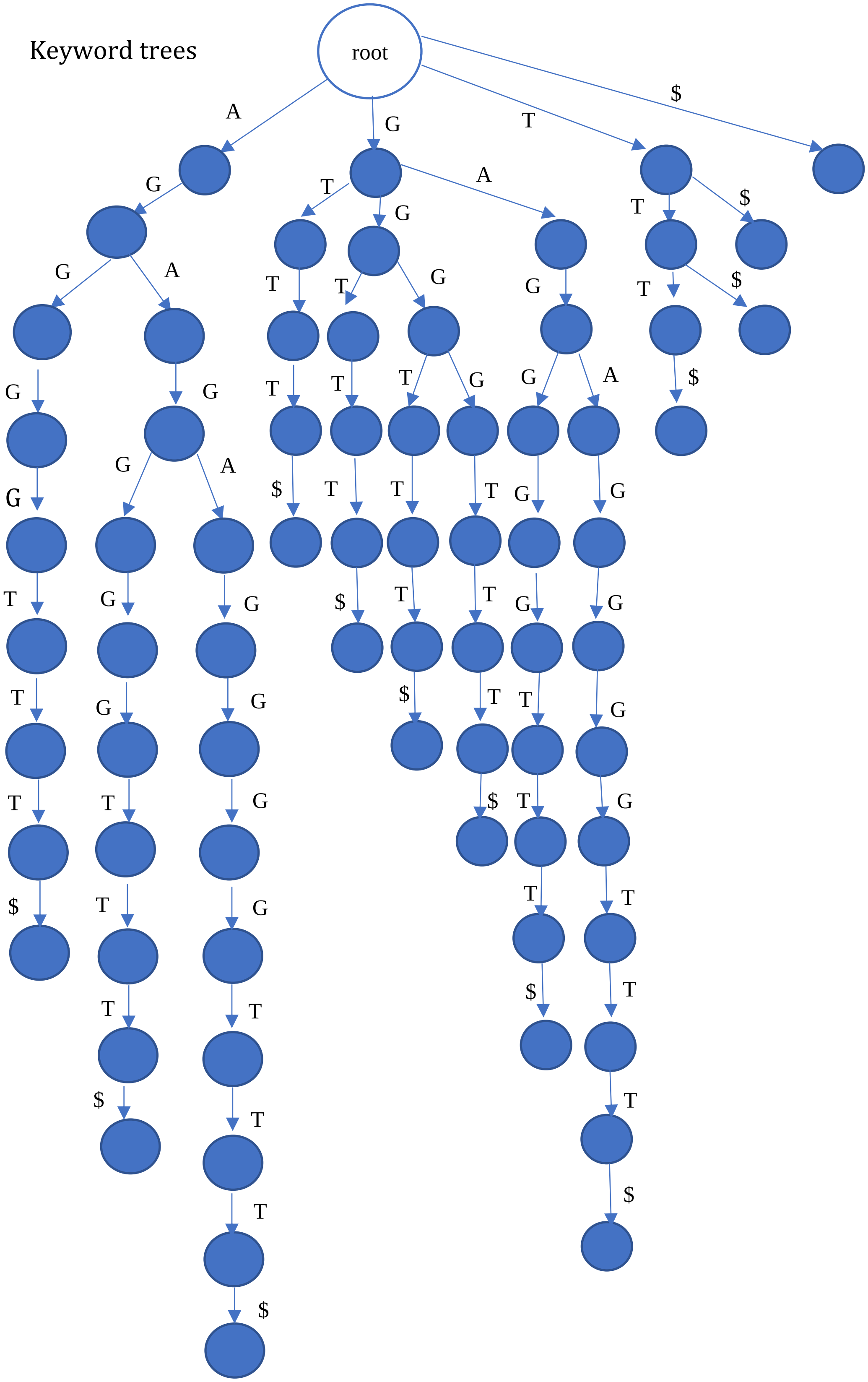


Problem 2. Suffix Tree
20pts

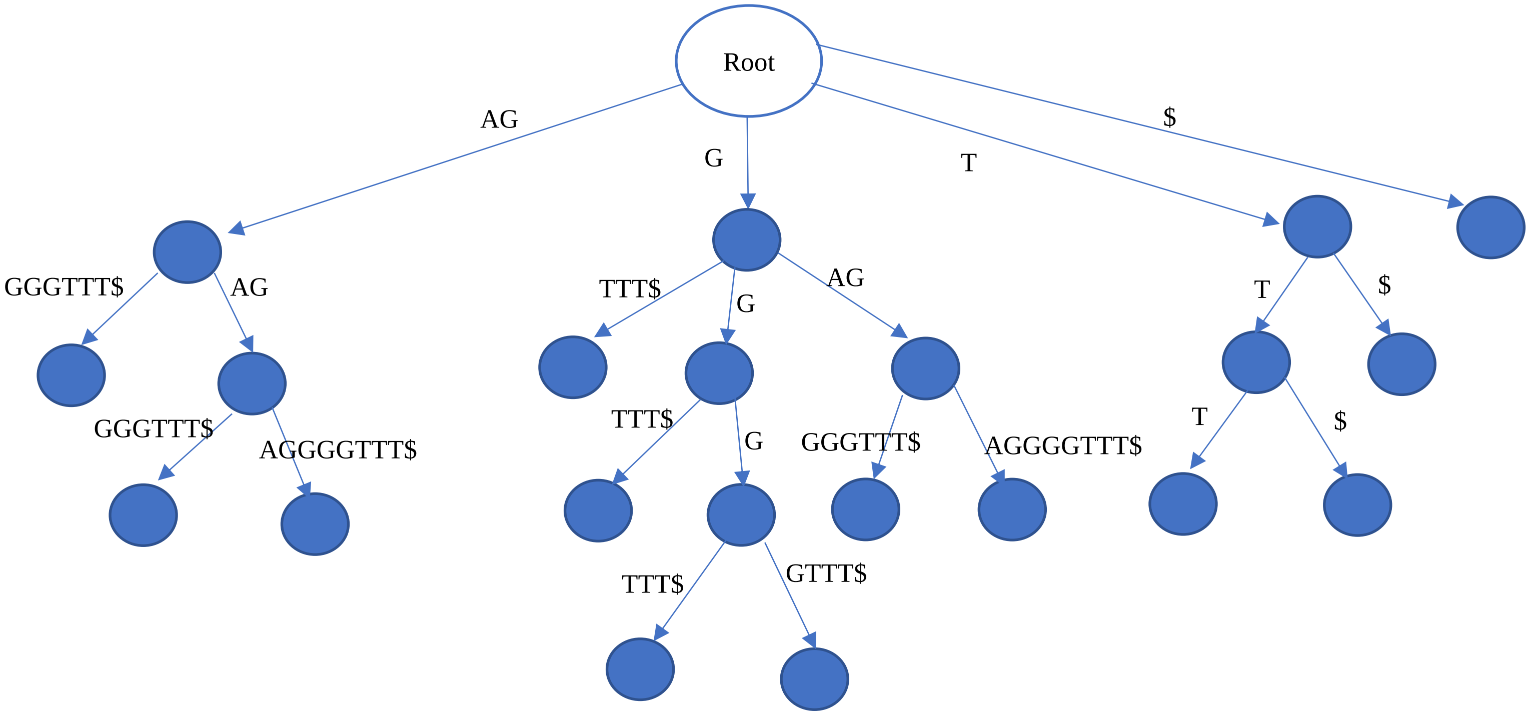
Build a suffix tree, with compressed edges, i.e., no internal nodes with only a single child node, from the following sequence.
AGAGAGGGGTTT

Suffixes:
AGAGAGGGGTTT
GAGAGGGGTTT
AGAGGGGTTT
GAGGGGTTT
AGGGGTTT
GGGGTTT
GGGTTT
GGTTT
GTTT
TTT
TT
T

Keyword trees



Suffix Tree


Root
AG
G
T
$
GGGTTT$
AG
TTT$
G
AG
T
$
GGGTTT$
AGGGGTTT$
TTT$
G
GGGTTT$
AGGGGTTT$
T
$
TTT$
GTTT$

Problem 3. Statistics revisited

20pts

Consider sequence S (of length L) evolved from an ancestral sequence that had the same length (L) but was composed of all 'A's. The sequence S has N_A 'A's and $L - N_A$ other nucleotides (C or G or T). Assume that:

- each position in the ancestral sequence evolved independently until today, which is 1000 generations later,
- in each generation, a nucleotide mutates with probability μ and stays the same with probability 1-μ,
- no nucleotide will have mutated twice in the 1000 generations (the chances are exceedingly low, and you may ignore the possibility).

a) Calculate the expected number of A's in S as a function of μ. (10 points)

b) Then, by equating this expectation to the observed count N_A, write down a formula for μ as a function of N_A. (6 points)

c) Using this formula for μ, compute its value when L = 10000 and N_A = 1000. (4 points)

(a) First generation: $E(A_1) = L - E$ (nucleotide mutates)

$$= L - \mu L$$

$$= L(1-\mu)$$

Second generation: $E(A_2) = L(1-\mu) - L(1-\mu)\mu$

$$= L - 2L\mu + L\mu^2$$

$$= L(1-\mu)^2$$

1000 generations: $E(A_{1000}) = L(1-\mu)^{1000}$

(b) $L(1-\mu)^{1000} = N_A$

$$(1-\mu)^{1000} = \frac{N_A}{L}$$

$$1-\mu = \sqrt[1000]{\frac{N_A}{L}}$$

$$\mu = 1 - \sqrt[1000]{\frac{N_A}{L}}$$

(c) $\mu = 1 - \sqrt[1000]{\frac{1000}{10000}}$

$$\mu = 1 - \sqrt[1000]{\frac{1}{10}}$$

$$\mu = 1 - 0.9977 = 0.0023$$

Therefore, the value of $\mu = 0.0023 \ll 1$