Tutorial Week 6

Definition 1. *The longest common prefix of two sequences x and y is denoted by* $\mathrm{lcp}(x, y)$*. For some ordered alphabet A such that $a < b$, we say that a string u is smaller lexicographically than a string v, if either u is a prefix of v, or* $\mathrm{lcp}(u, v) = x$ *and* $u[|x|] < v[|x|]$.

Exercise 1. *Consider the following list of sequences: anchoritess, ancestral, ancestress, ancestor, ancestorial, ancestorially, anchored, anchoress, anchoritic, and, anchorite, anchor, anchorable, anchorage, ancestries, ancestry, anchorages, anchorate, ancestresses, anchoring, ancestors, anchoretical, anchorites, anchoretish, anchoretism.*

Consider a random distribution of this list.:
a) *Do a sorting of the list with bucket sort.*
b) *Calculate the* lcp *of consecutive elements.*
c) *Using the following Lemma compute the rest of the* lcp*'s.*

Lemma 2. *Let $L_0 \leq L_1 \leq \ldots \leq L_{n-1}$ and take the integers d, i, and f, with $-1 < d < i < f < n$. Then* $\mathrm{lcp}(L_d, L_f) = \min(\mathrm{lcp}(L_d, L_i), \mathrm{lcp}(L_i, L_f))$.

d) *Draw the binary tree of the* lcp*'s for the previous item.*

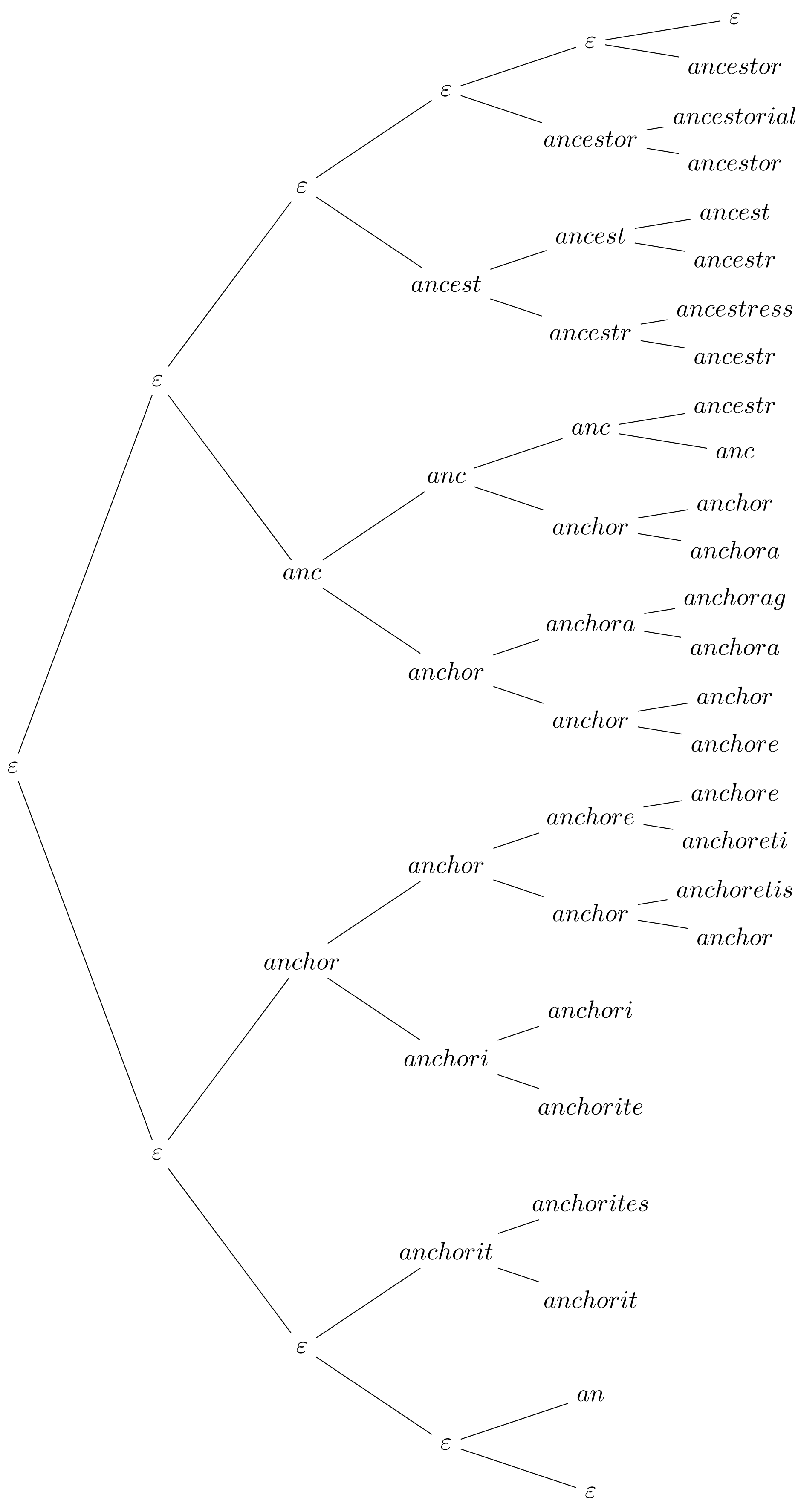
Solution: **a)** The ordered list of sequences is the following: *ancestor*, *ancestorial*, *ancestorially*, *ancestors*, *ancestral*, *ancestress*, *ancestresses*, *ancestries*, *ancestry*, *anchor*, *anchorable*, *anchorage*, *anchorages*, *anchorate*, *anchored*, *anchoress*, *anchoretical*, *anchoretish*, *anchoretism*, *anchoring*, *anchorite*, *anchorites*, *anchoritess*, *anchoritic*, *and*.

To obtain the list by bucket sort, we extend by *nill* positions all short sequences, such that all have the same length. Next, we arrange them by looking at the letters from right to left, keeping them in place (that is, if we compare the jth letter of some strings u and v, such that u comes before v in the text, then u will occur after v in the new list, only if $u[j] > v[j]$).

b) If the sequences are ordered lexicographically, we consider in the following list the element i, to be the ith element in the list (e.g., *ancestress* will be denoted by 6.) As there are 25 elements in our list, we will denote all these by the numbers $0, 1, \ldots, 24$. Please recall that a -1 has to be added at the beginning and end to ease the computation and follow the algorithm presented in the course.

The list of all the consecutive lcp's is the following:
lcp$(-1, 0) = \varepsilon$
lcp$(0, 1) = ancestor$
lcp$(1, 2) = ancestorial$
lcp$(2, 3) = ancestor$
lcp$(3, 4) = ancest$
lcp$(4, 5) = ancestr$
lcp$(5, 6) = ancestress$
lcp$(6, 7) = ancestr$
lcp$(7, 8) = ancestr$
lcp$(8, 9) = anc$
lcp$(9, 10) = anchor$
lcp$(10, 11) = anchora$
lcp$(11, 12) = anchorag$
lcp$(12, 13) = anchora$
lcp$(13, 14) = anchor$
lcp$(14, 15) = anchore$
lcp$(15, 16) = anchore$
lcp$(16, 17) = anchoreti$
lcp$(17, 18) = anchoretis$
lcp$(18, 19) = anchor$
lcp$(19, 20) = anchori$
lcp$(20, 21) = anchorite$
lcp$(21, 22) = anchorites$
lcp$(22, 23) = anchorit$
lcp$(23, 24) = an$
lcp$(24, 25) = \varepsilon$

c) and d) We fill directly the balanced binary tree representation of the lcp's with the corresponding values, using Lemma 2 . The presentation of the balanced binary tree is the following:


ε
ancestor
ancestorial
ancestor
ancest
ancestr
ancestress
ancestr
ancestr
anc
anchor
anchora
anchorag
anchora
anchor
anchore
anchore
anchoreti
anchoretis
anchor
anchori
anchorite
anchorites
anchorit
an
ε
ε
ε
ancestor
ancest
ancestr
ε
ancest
anc
anchor
anchora
anchor
anchore
anchor
anchori
anchorit
ε
ε
anc
anchor
ε
ε
ε