Introduction to string indexing

UNIVERSITY OF MARYLAND

Tries

A trie (pronounced "try") is a rooted tree representing a collection of strings with one node per common prefix

Smallest tree such that:

Each edge is labeled with a character $c \in \Sigma$

A node has at most one outgoing edge labeled c, for $c \in \Sigma$

Each key is "spelled out" along some path starting at the root

Natural way to represent either a *set* or a *map* where keys are strings

This structure is also known as a Σ-tree

Tries: example

Represent this map with a trie:

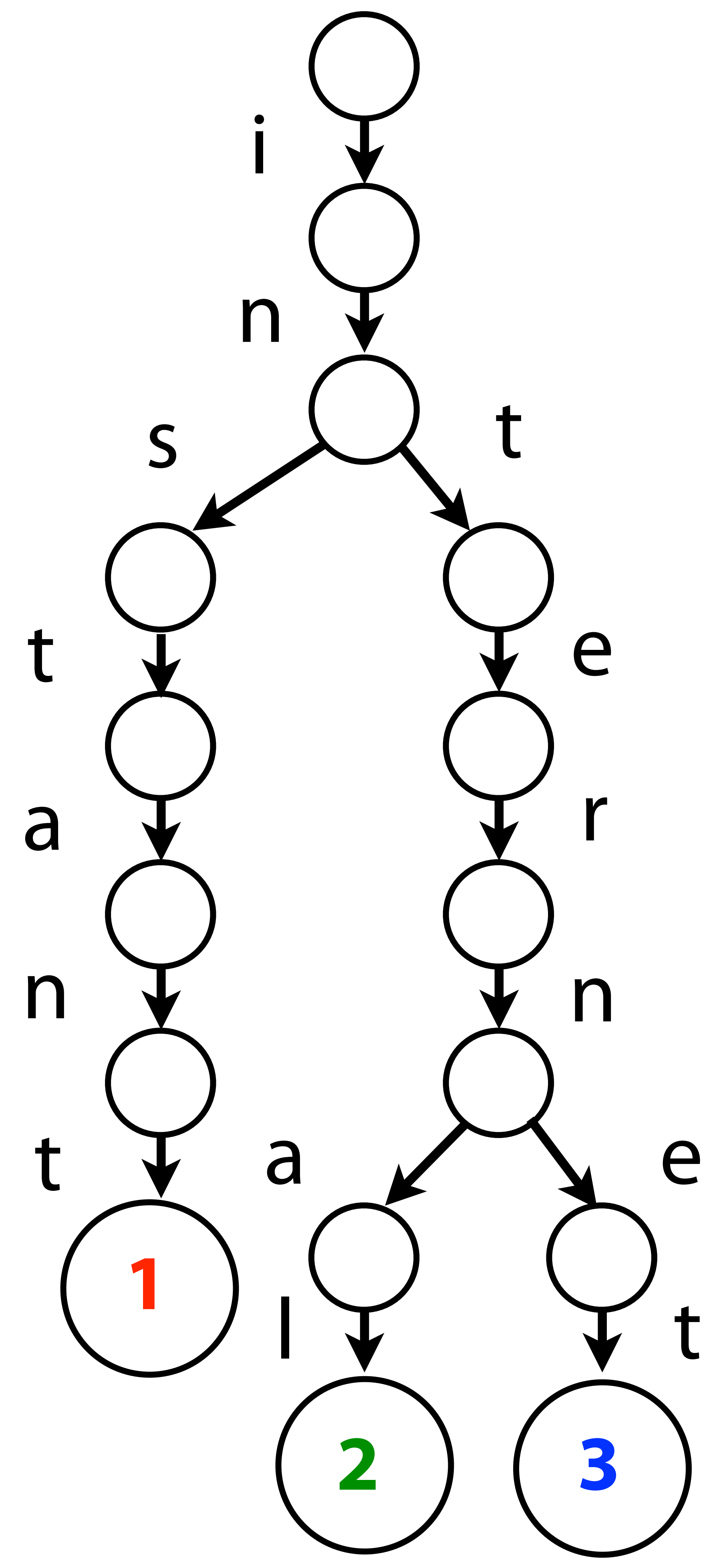
Key	Value
instant	1
internal	2
internet	3

The smallest tree such that:

Each edge is labeled with a character $c \in \Sigma$

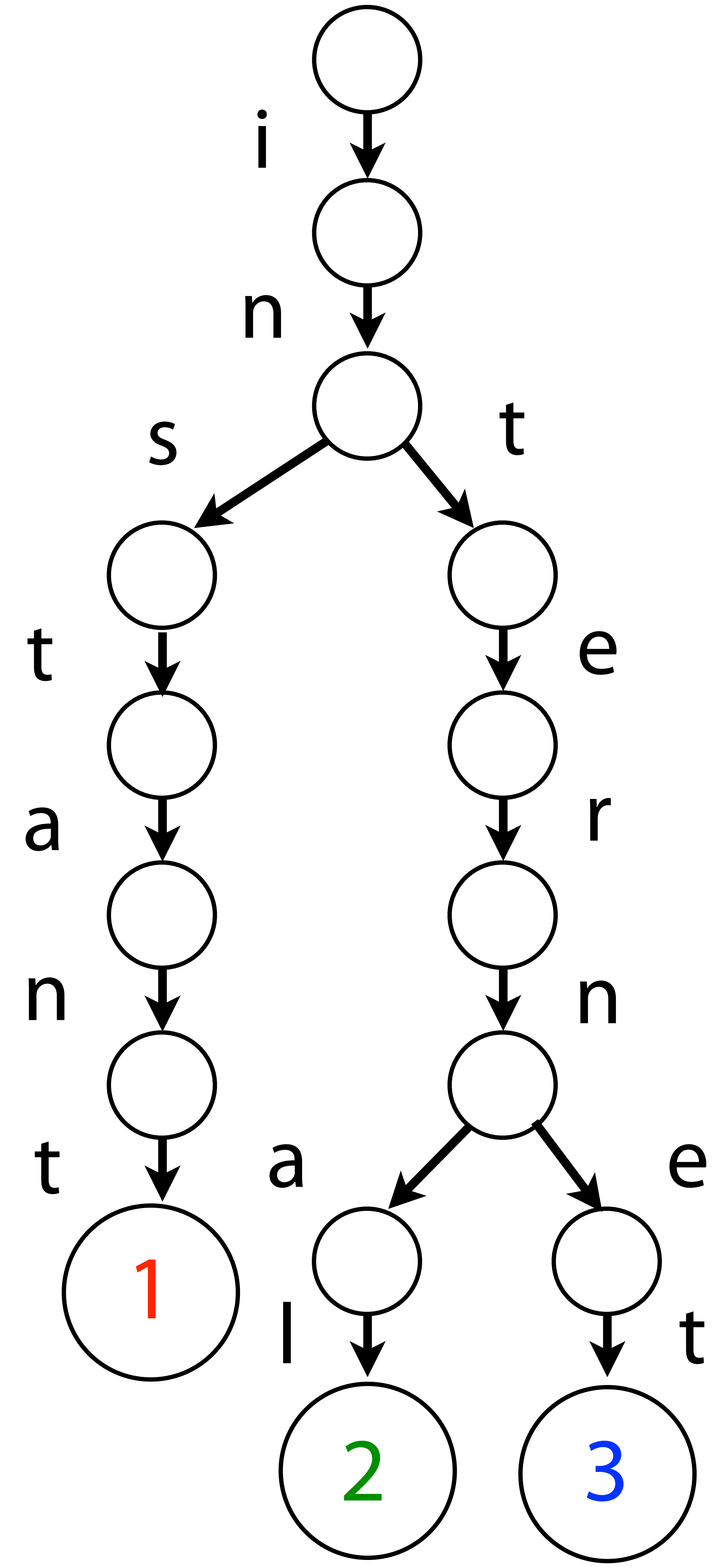
A node has at most one outgoing edge labeled c, for $c \in \Sigma$

Each key is "spelled out" along some path starting at the root



Tries

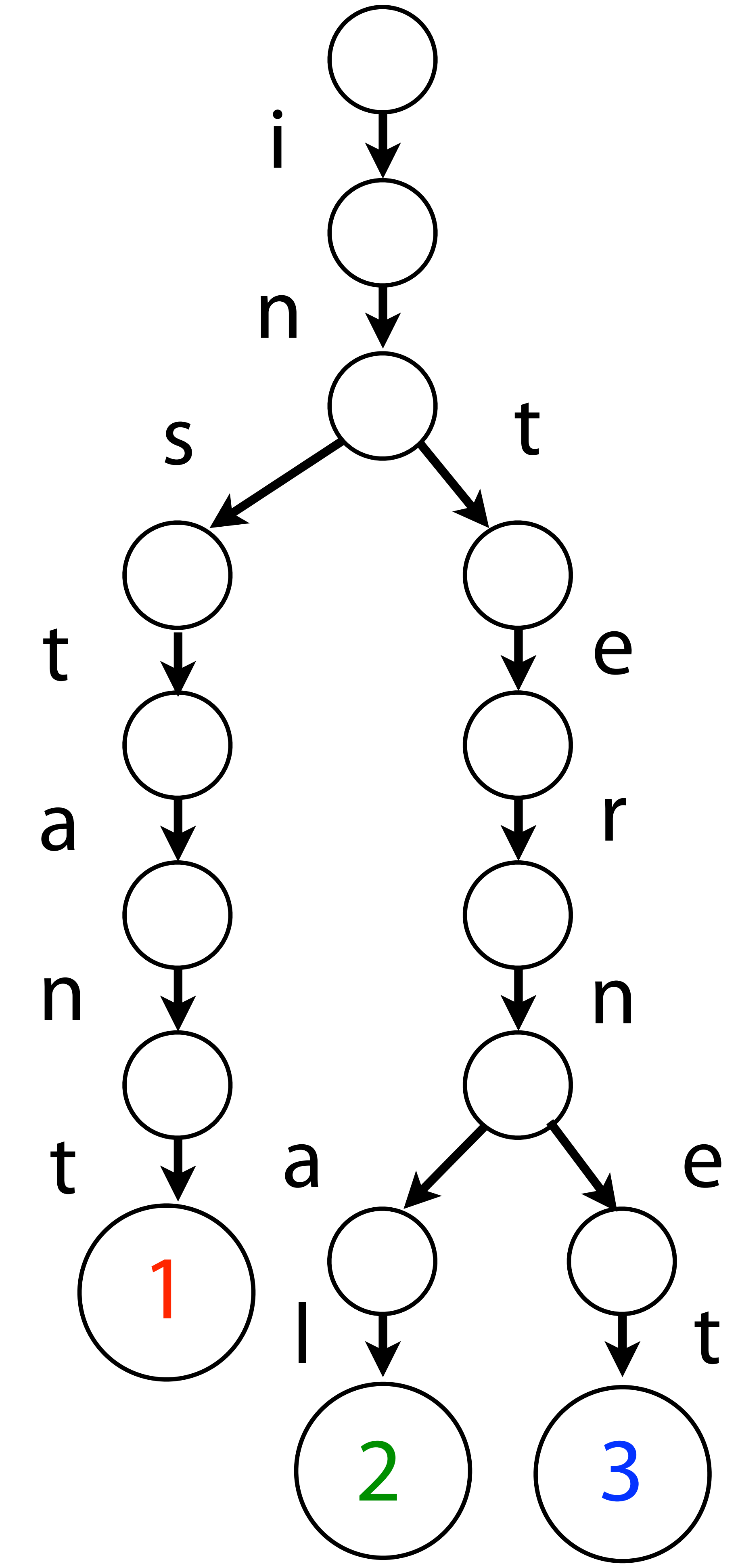
How do we check whether "infer" is in the trie?



Tries

How do we check whether "infer" is in the trie?

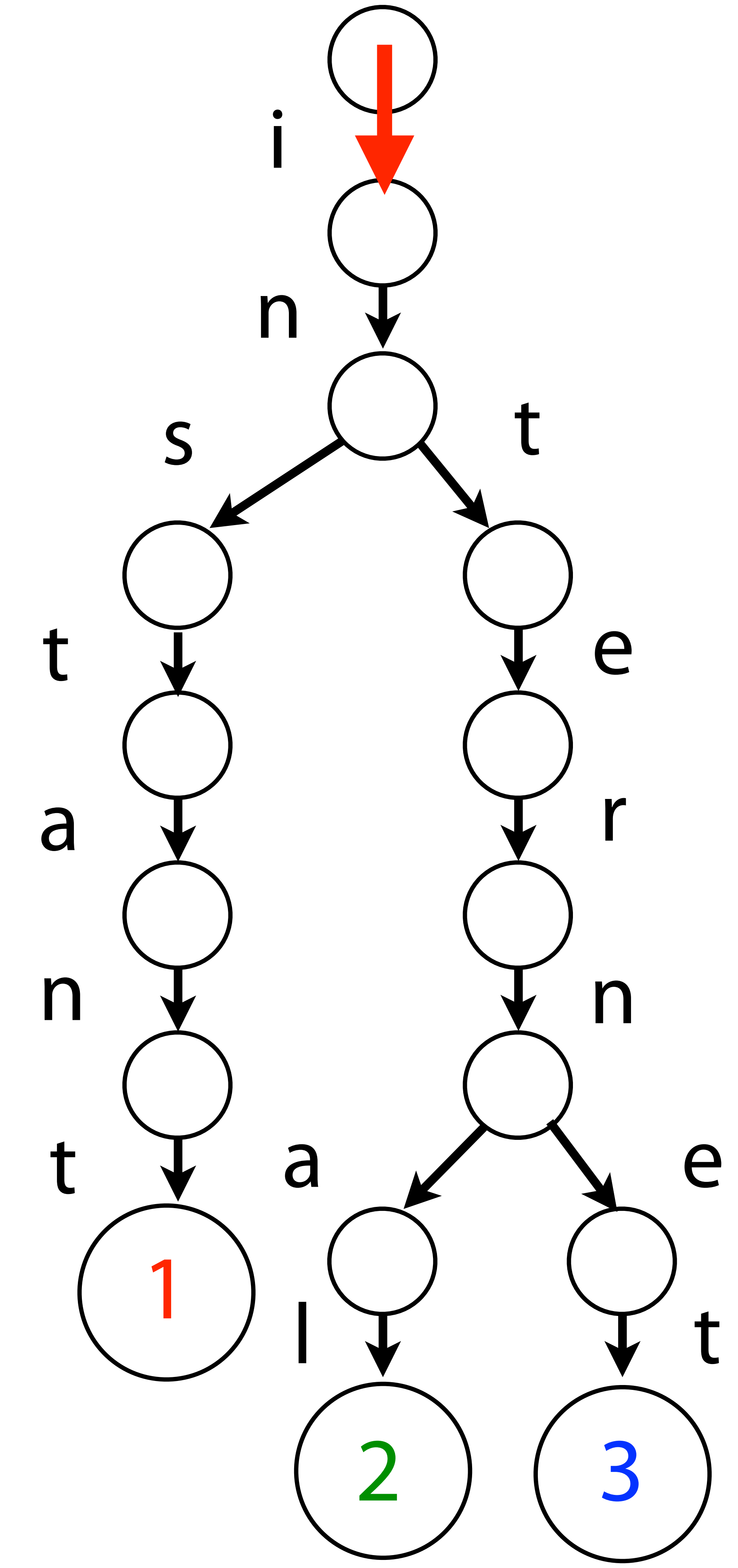
Start at root and try to match successive characters of "infer" to edges in trie



Tries

How do we check whether "infer" is in the trie?

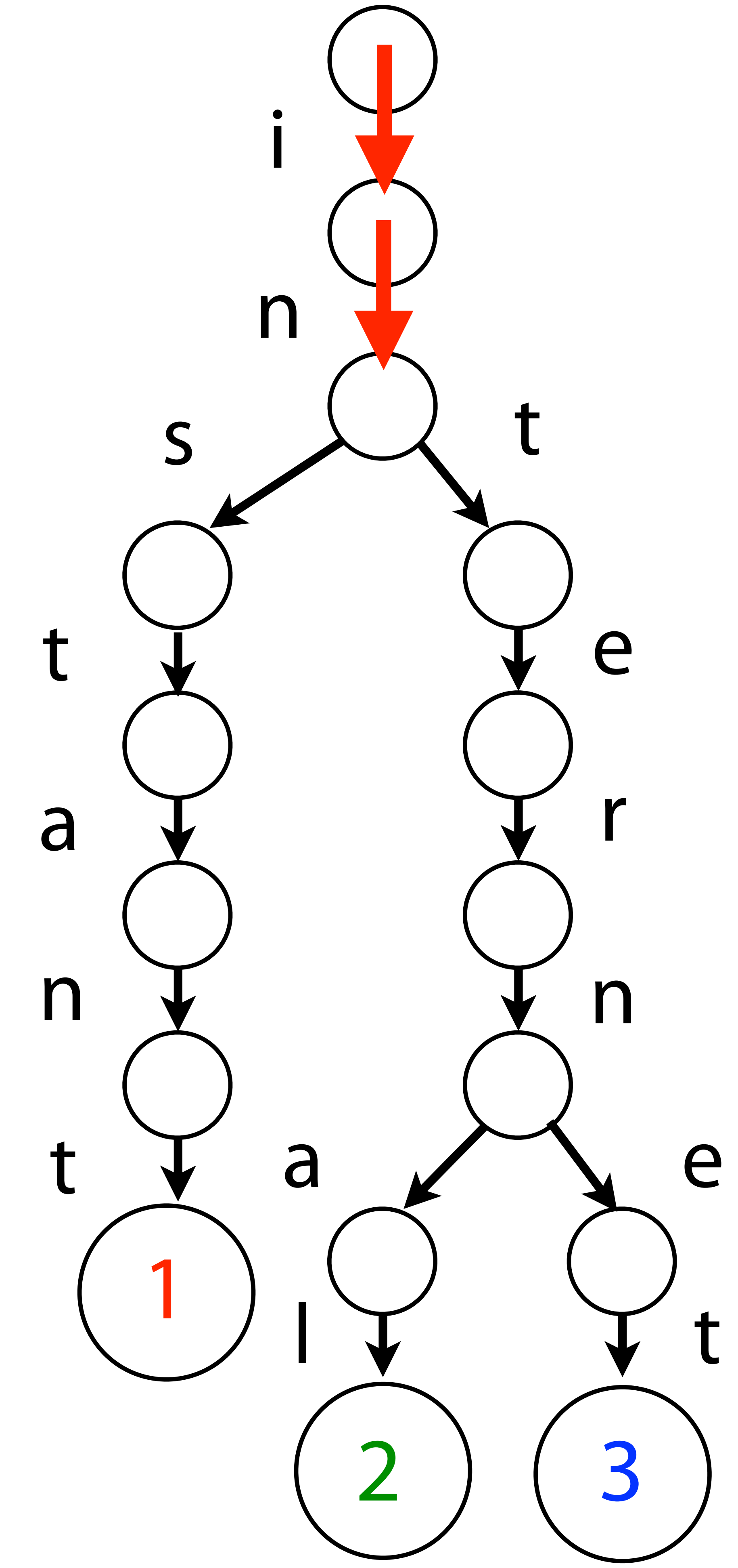
Start at root and try to match successive characters of "infer" to edges in trie



Tries

How do we check whether “infer” is in the trie?

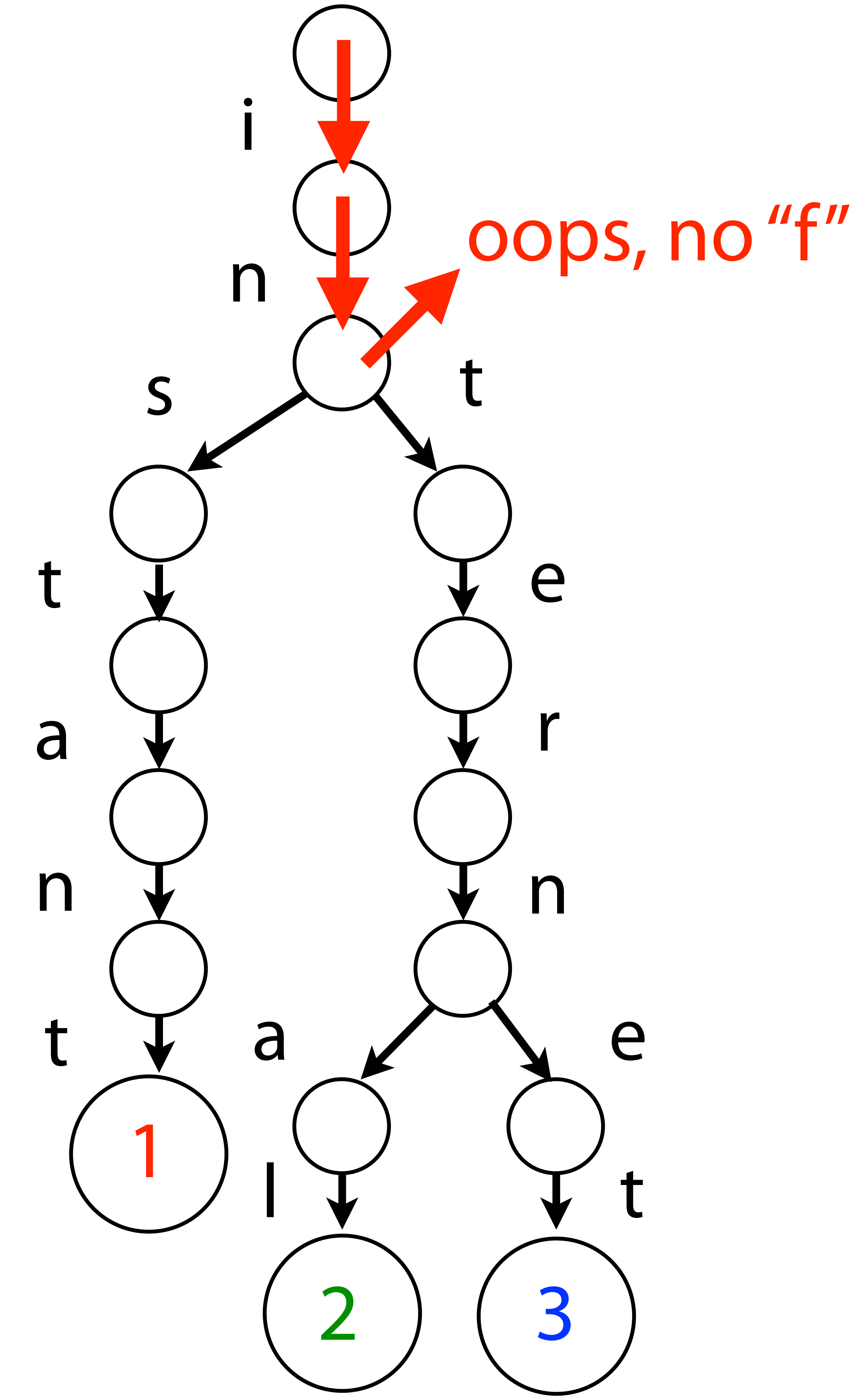
Start at root and try to match successive characters of “infer” to edges in trie



Tries

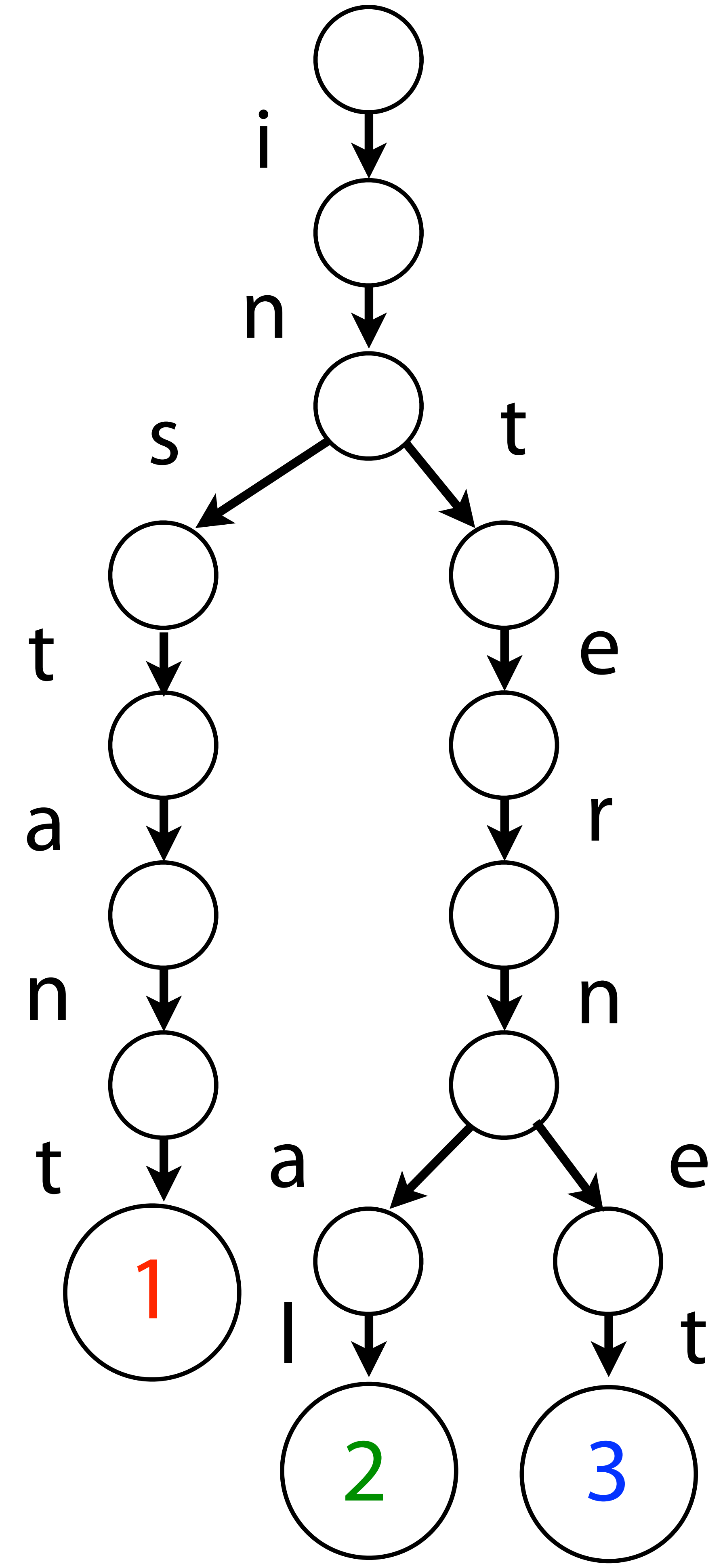
How do we check whether “infer” is in the trie?

Start at root and try to match successive characters of “infer” to edges in trie



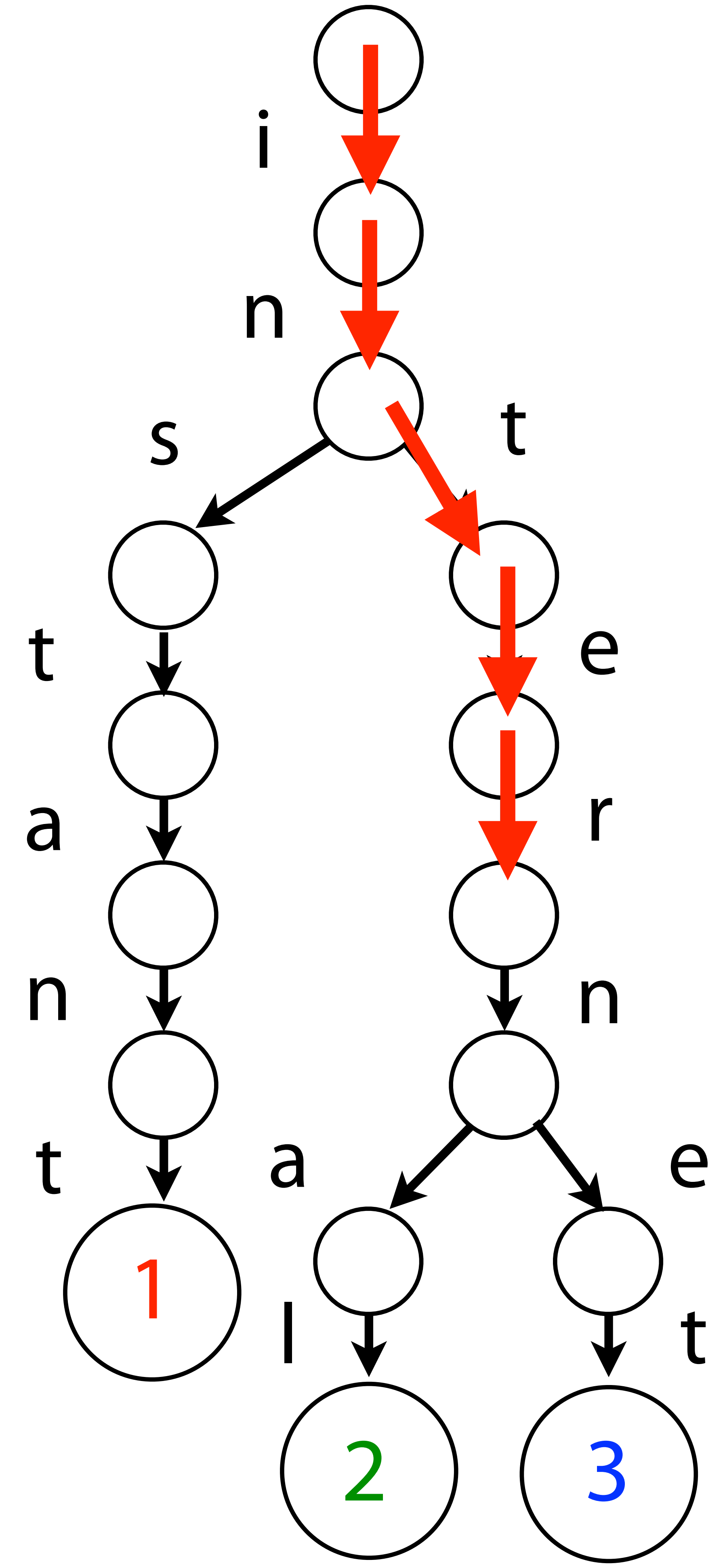
Tries

Matching “interesting”



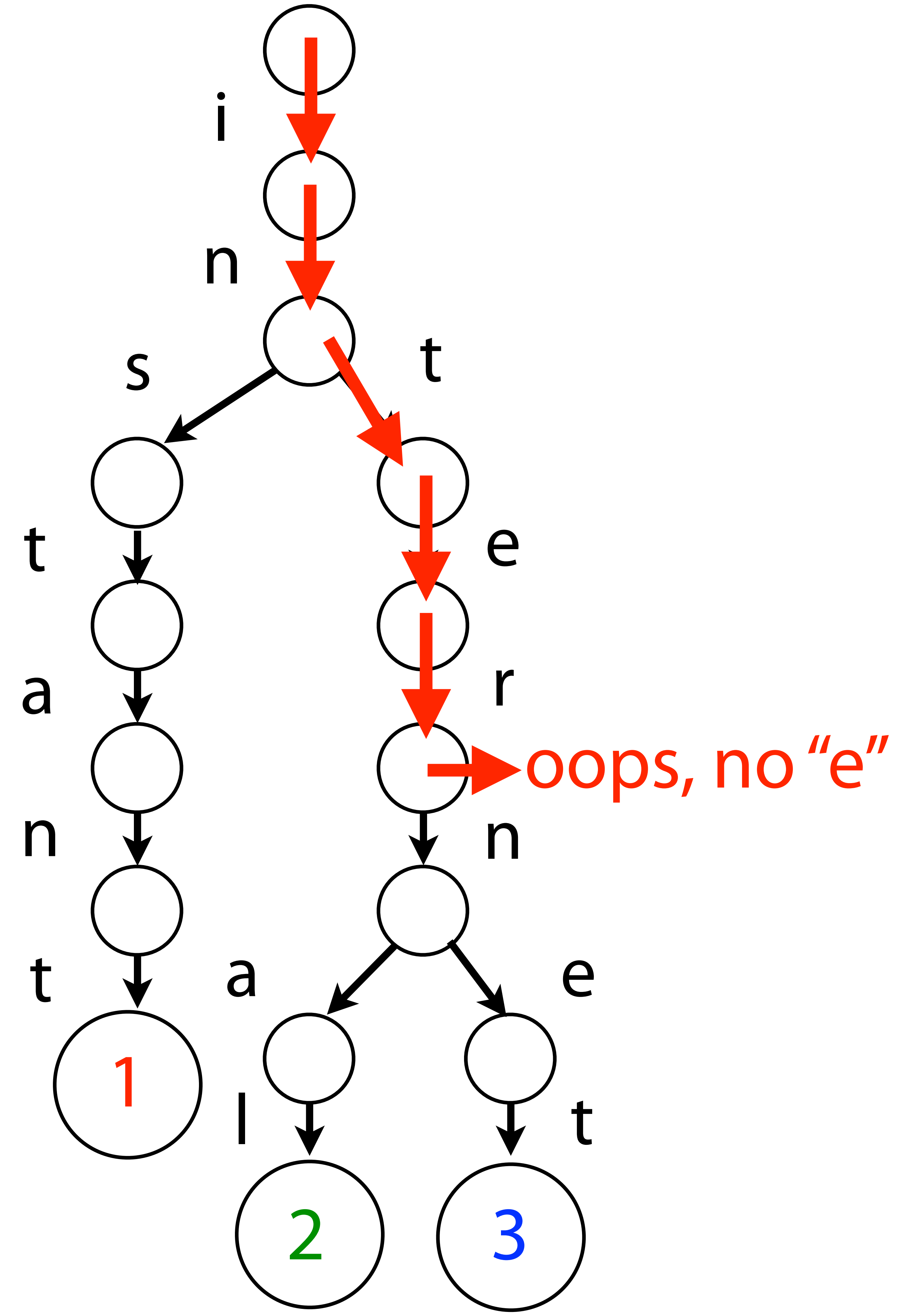
Tries

Matching “interesting”



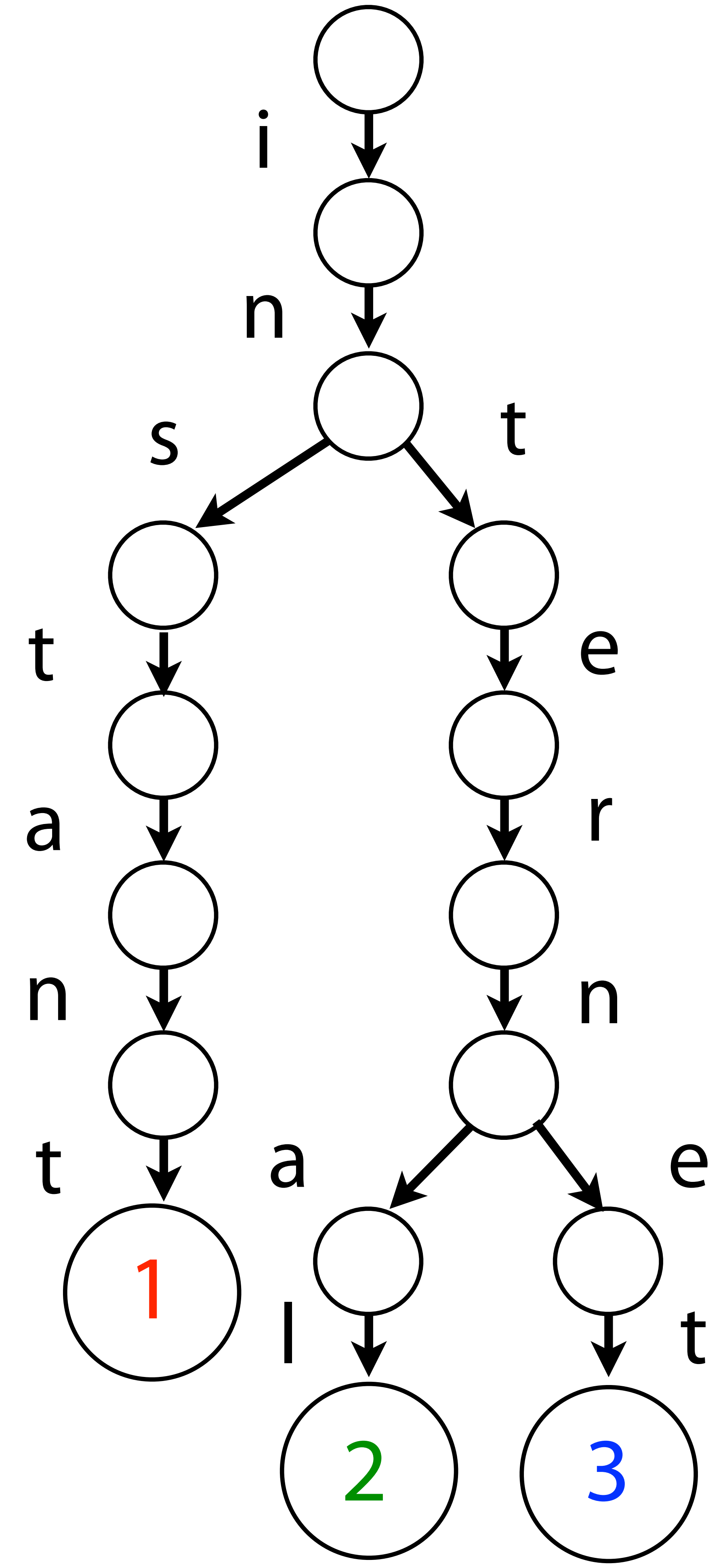
Tries

Matching “interesting”



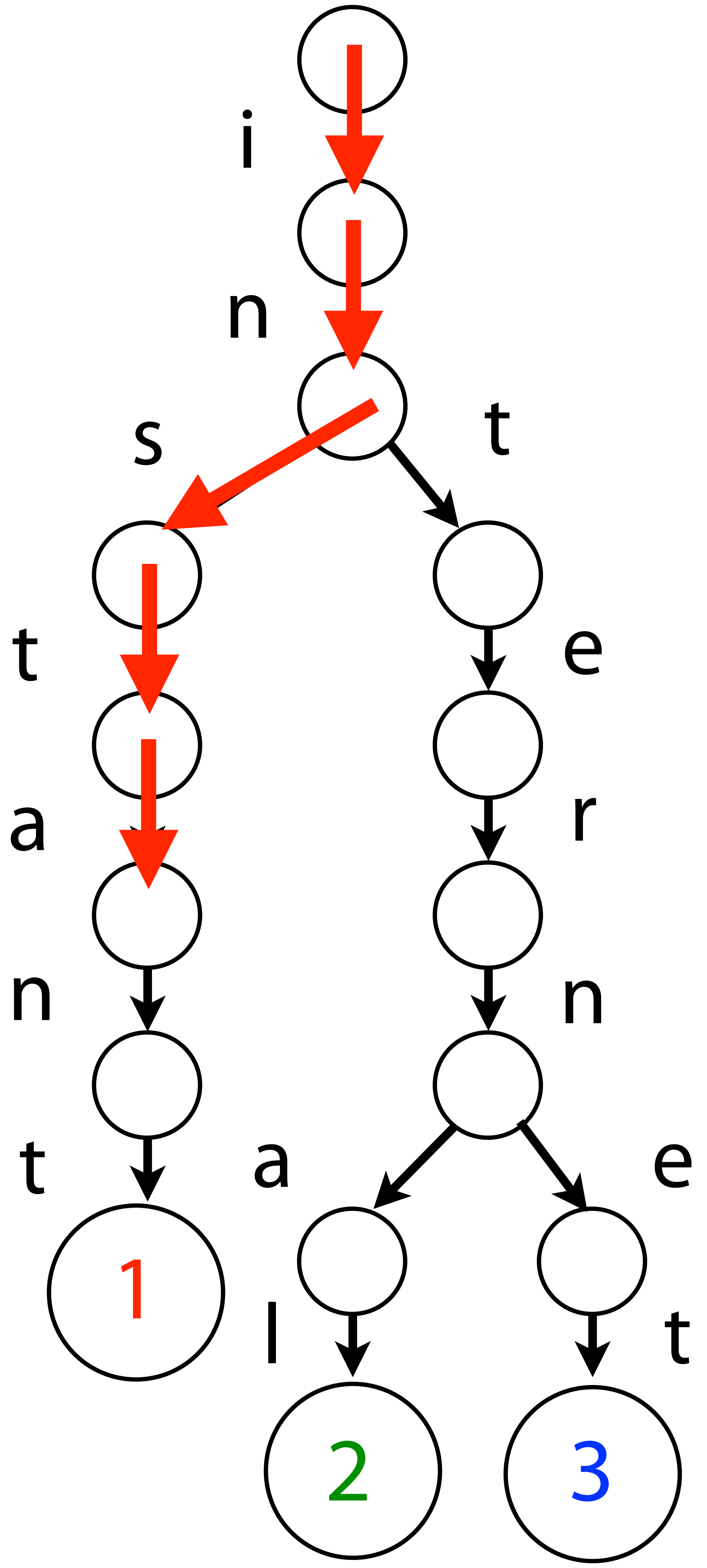
Tries

Matching “insta”



Tries

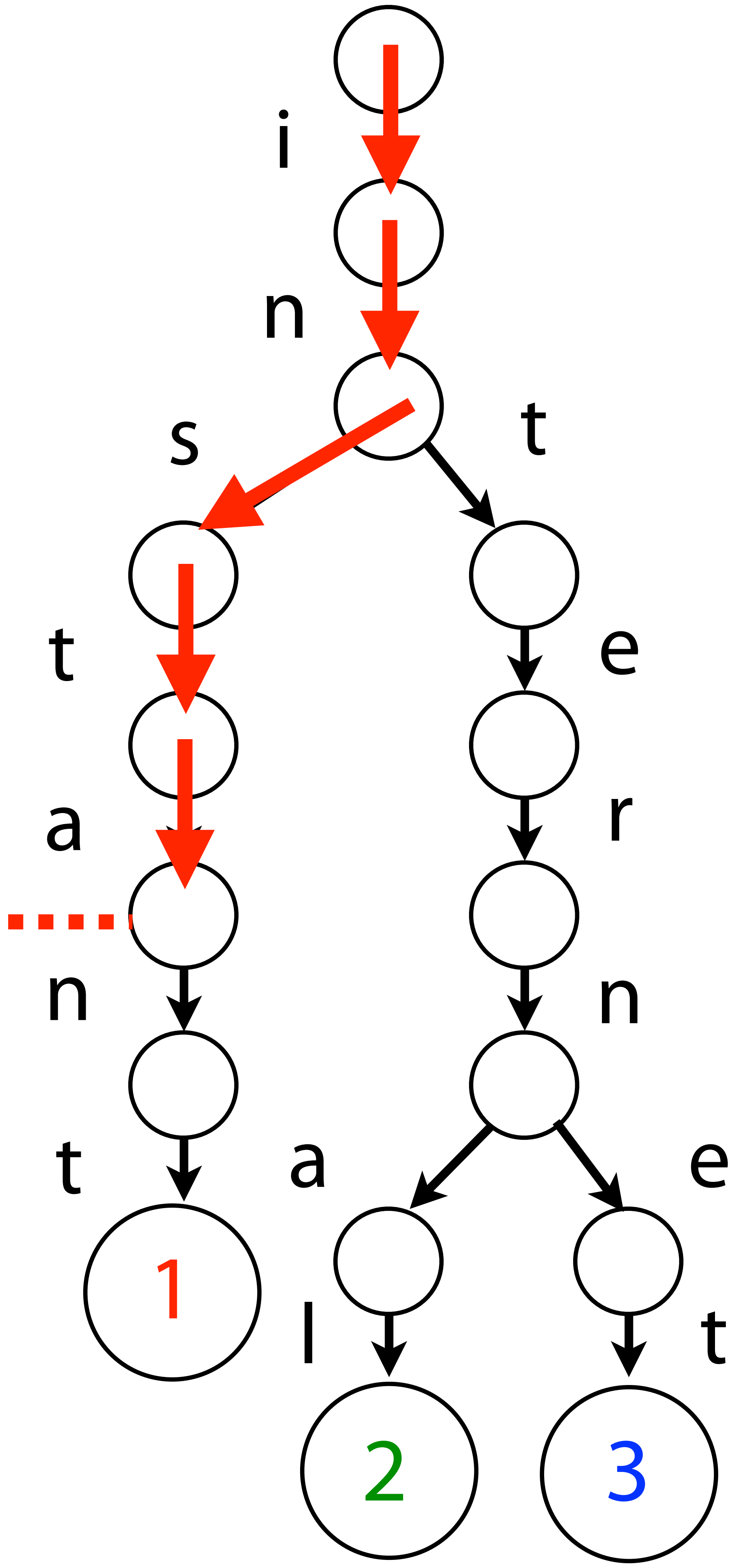
Matching “insta”



Tries

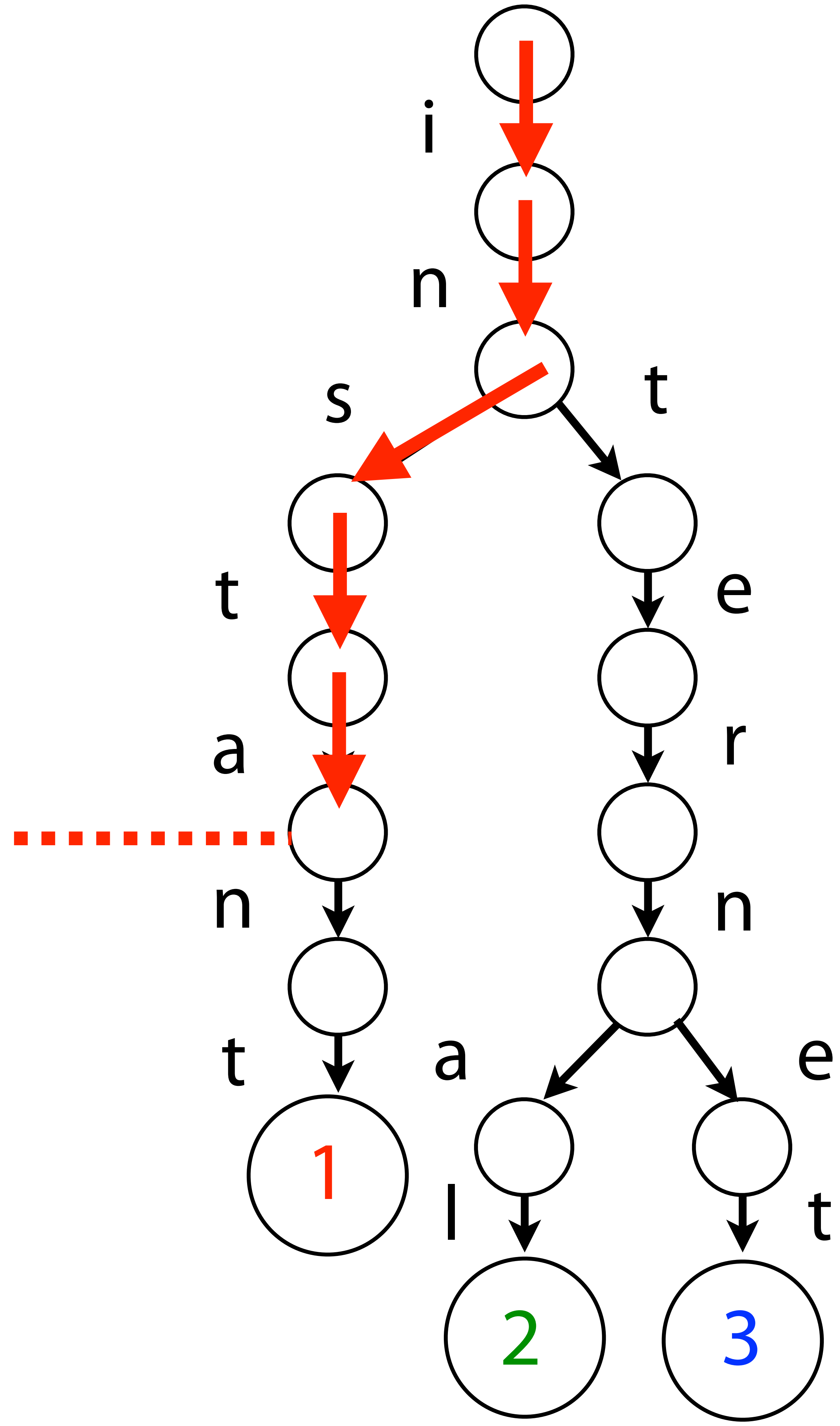
Matching “insta”

No value associated with node, so “insta” wasn’t a key



Tries

Matching “insta”

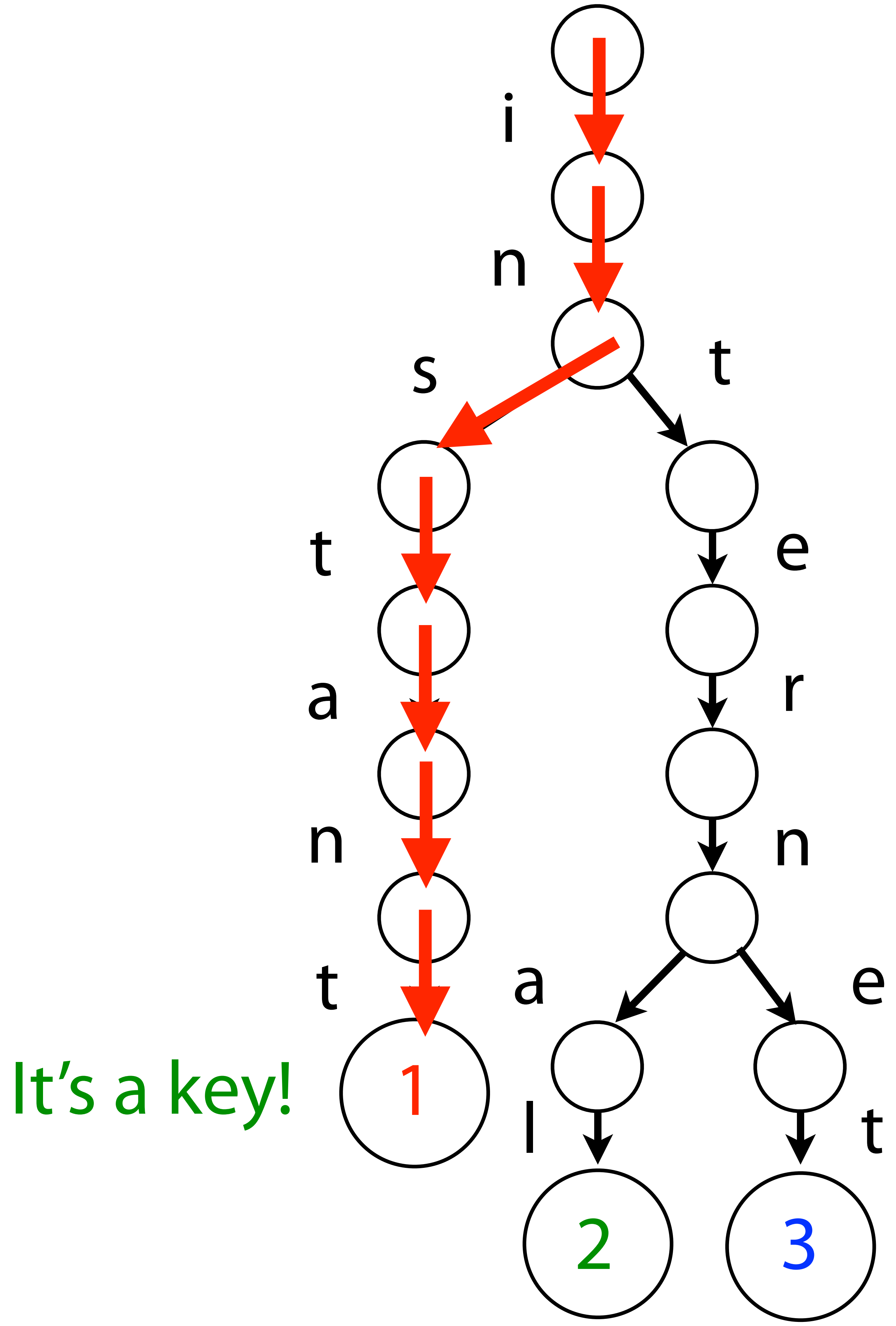


No value associated with node, so “insta” wasn’t a key

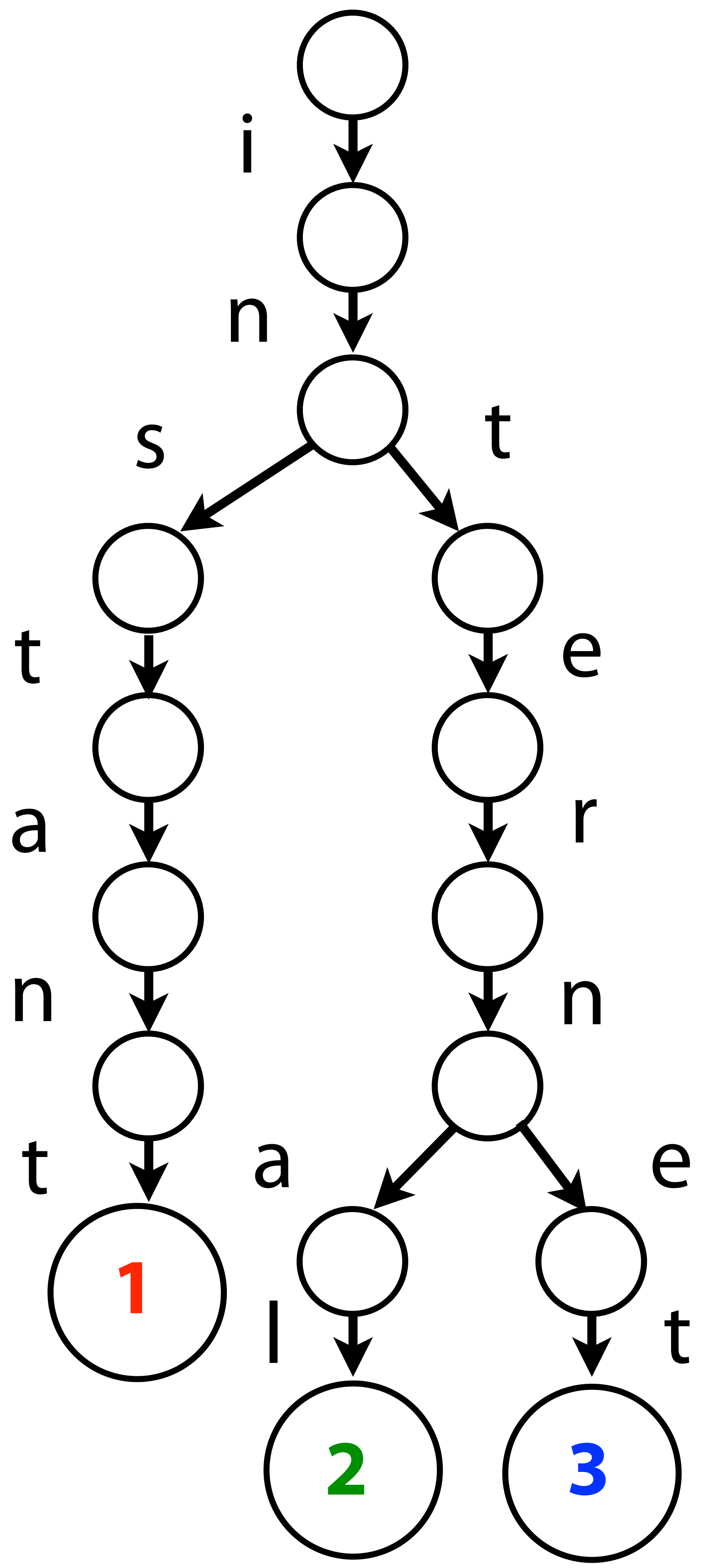
But it was a prefix of some key

Tries

Matching "instant"



Tries: example



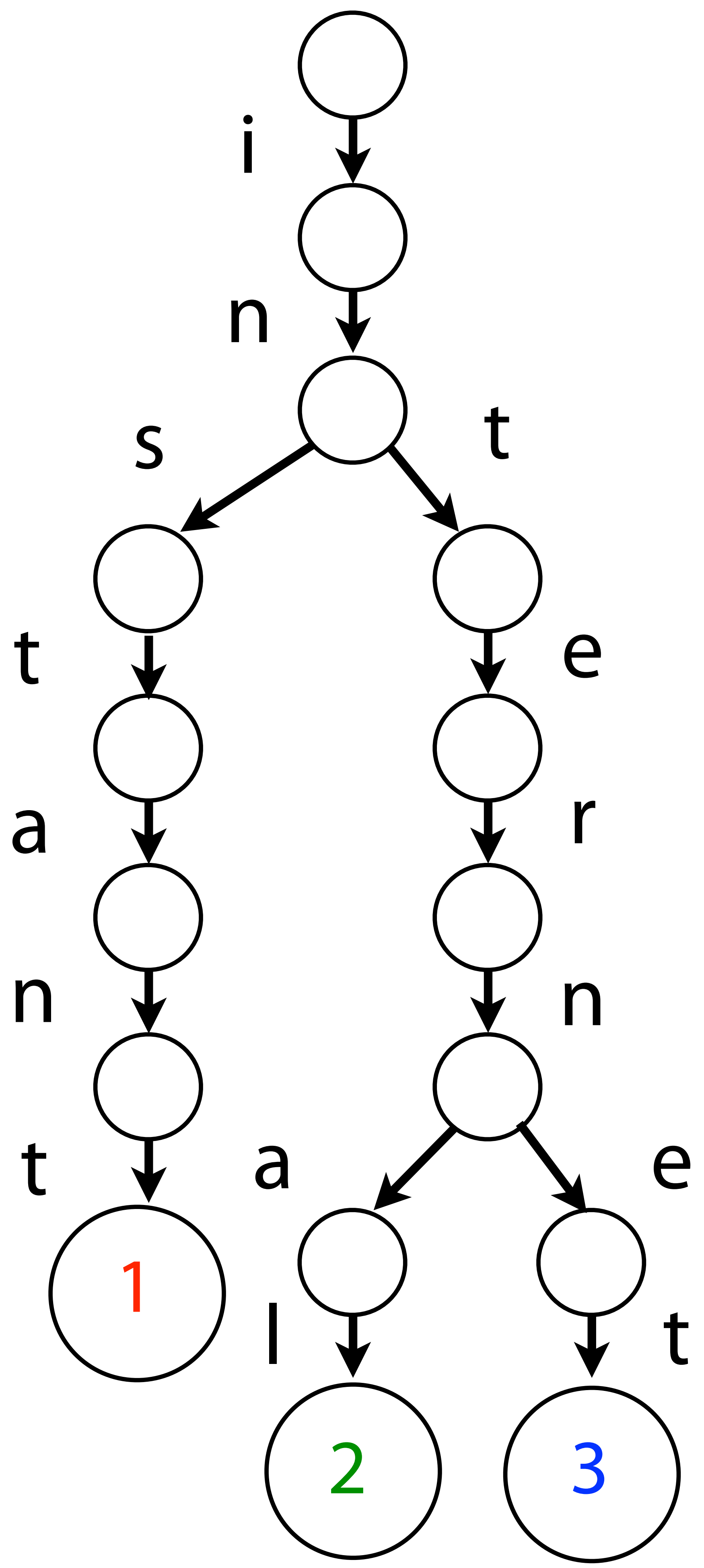
Checking for presence of a key P, where $n = |P|$, is $\boldsymbol{O(n)}$ time

If total length of all keys is N, trie has $\boldsymbol{O(N)}$ nodes

What about $|\Sigma|$?

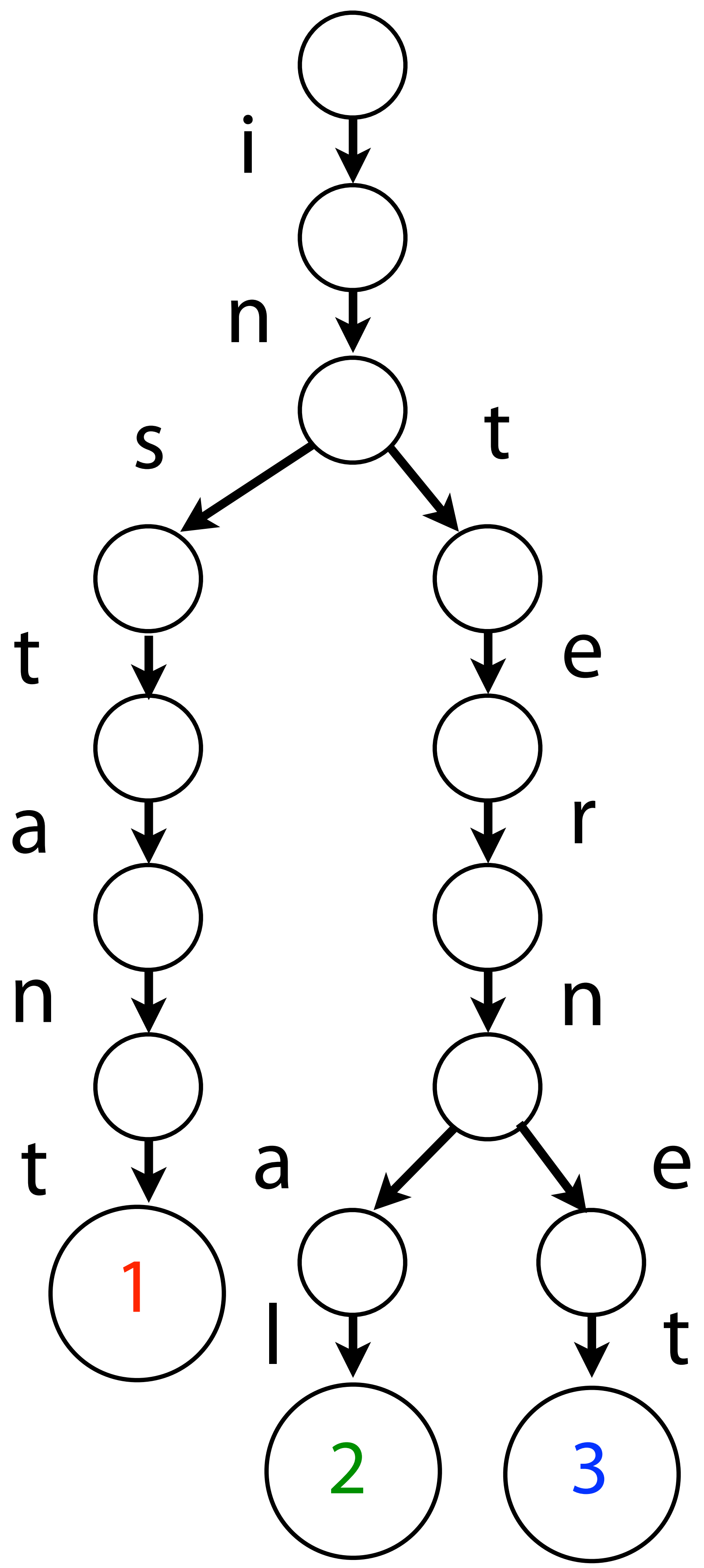
Depends how we represent outgoing edges. If we don't assume $|\Sigma|$ is a small constant, it shows up in one or both bounds.

Tries



How to represent edges between a node and its children?

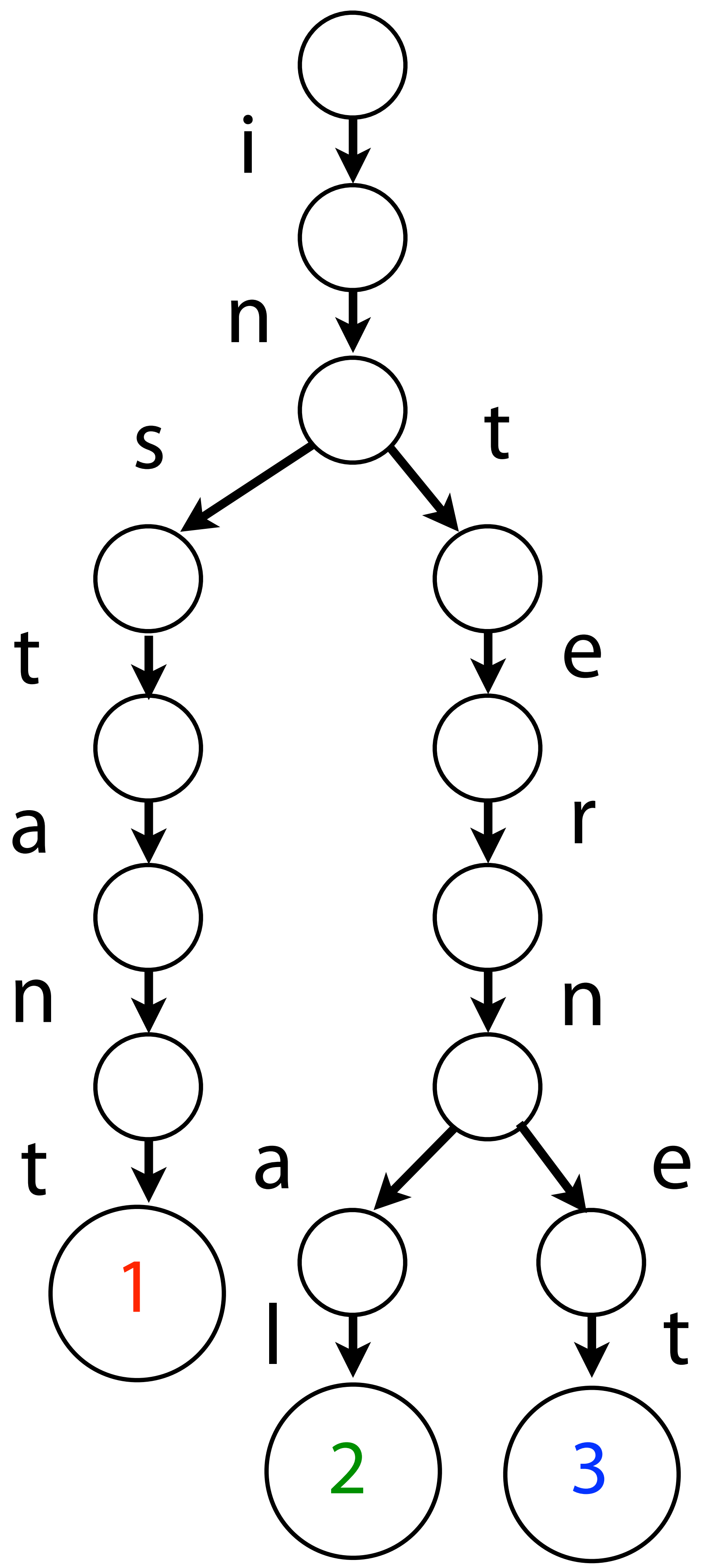
Tries



How to represent edges between a node and its children?

Map (from characters to child nodes)

Tries

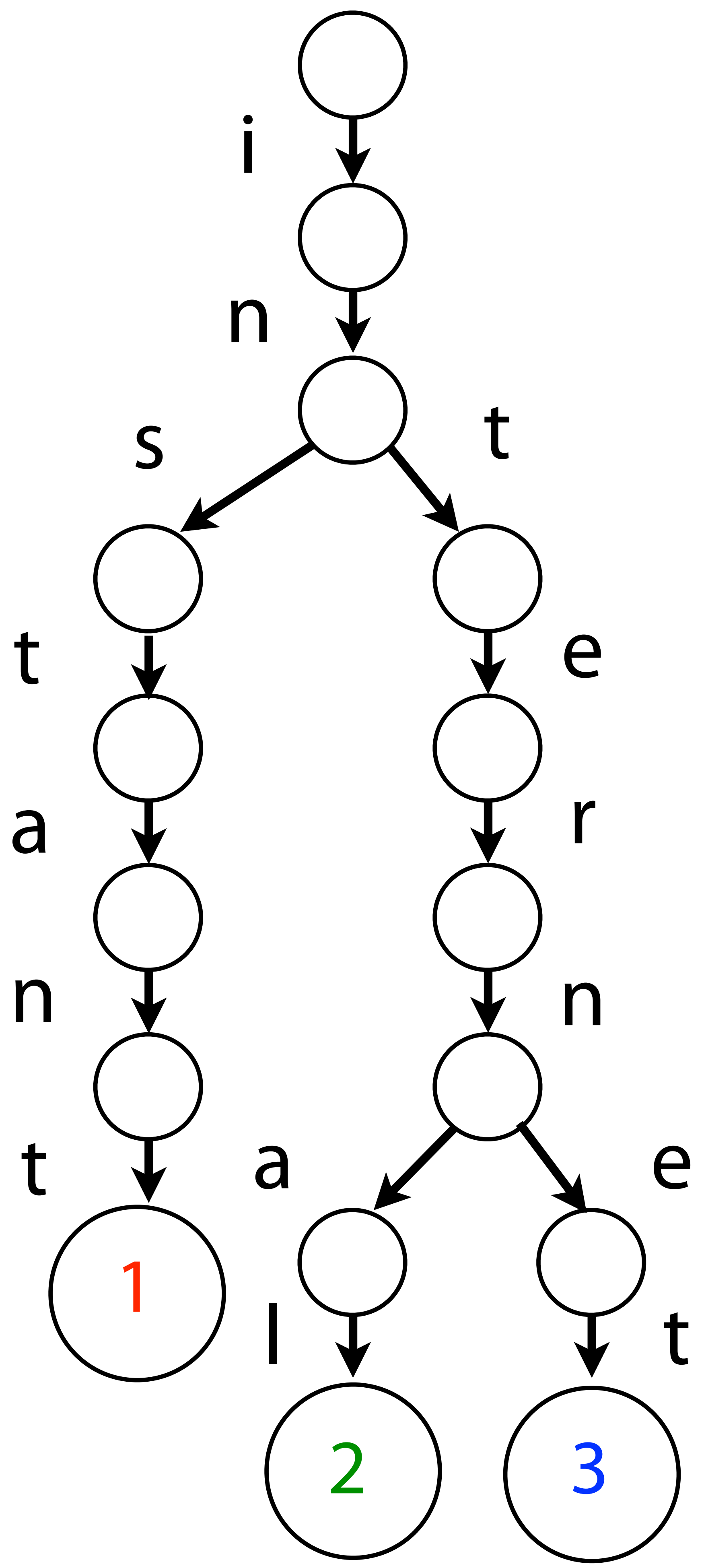


How to represent edges between a node and its children?

Map (from characters to child nodes)

Idea 1: Hash table

Tries



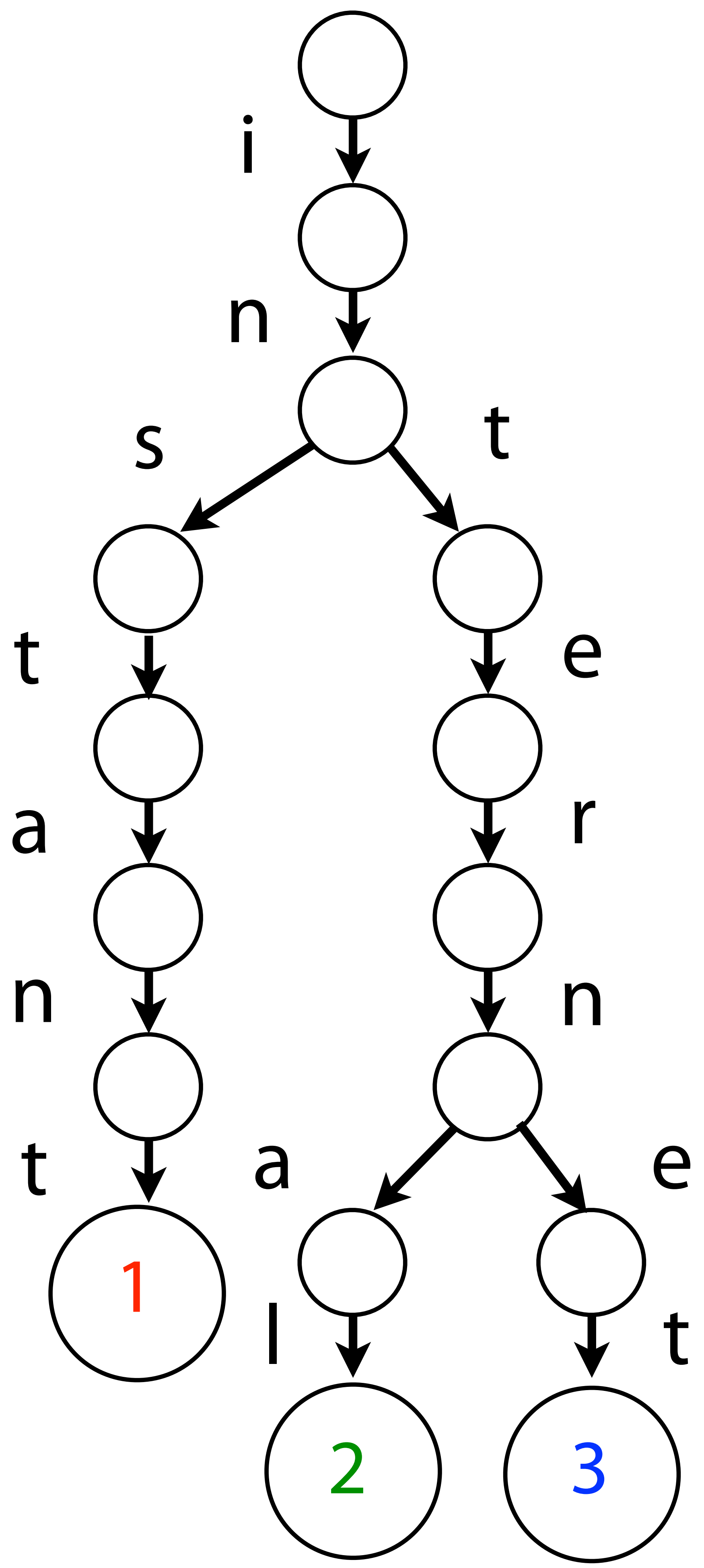
How to represent edges between a node and its children?

Map (from characters to child nodes)

Idea 1: Hash table

Idea 2: Sorted lists

Tries



How to represent edges between a node and its children?

Map (from characters to child nodes)

Idea 1: Hash table

Idea 2: Sorted lists

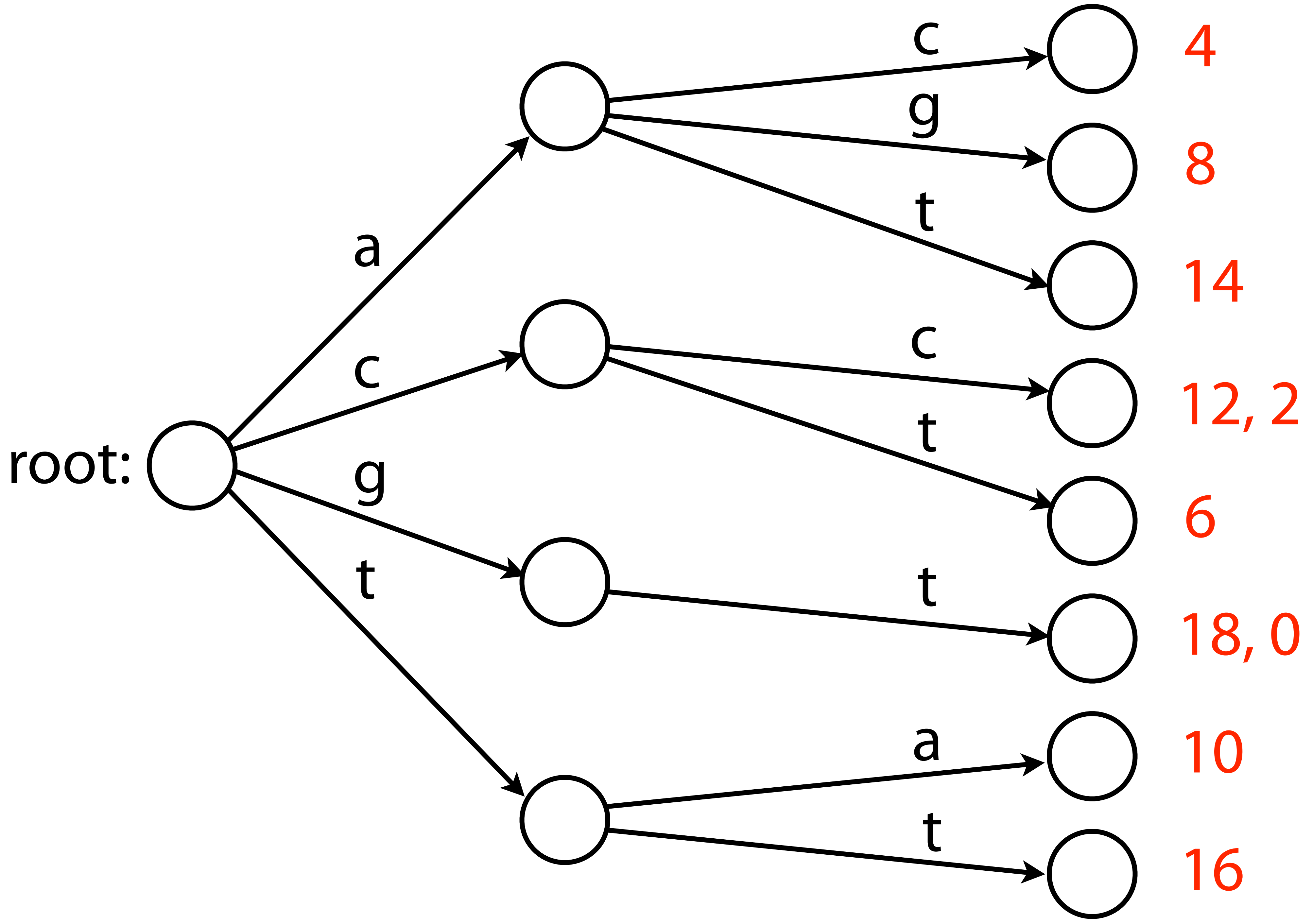
Assuming hash table, it's reasonable to say querying with P, $|P| = n$, is $O(n)$ time

Tries: another example

We can index *T* with a trie. The trie maps substrings to offsets where they occur

ac	4
ag	8
at	14
cc	12
cc	2
ct	6
gt	18
gt	0
ta	10
tt	16

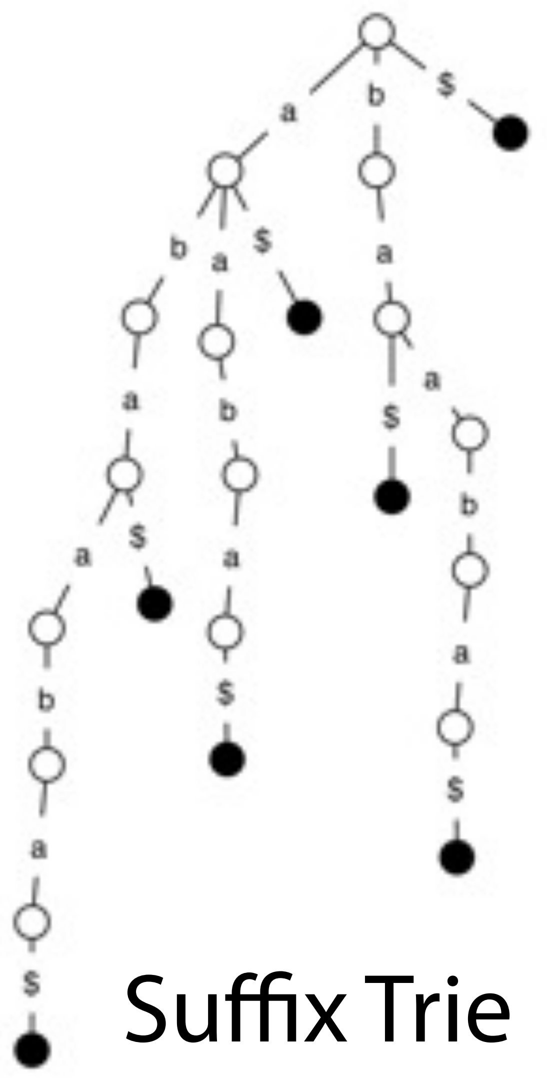
Index



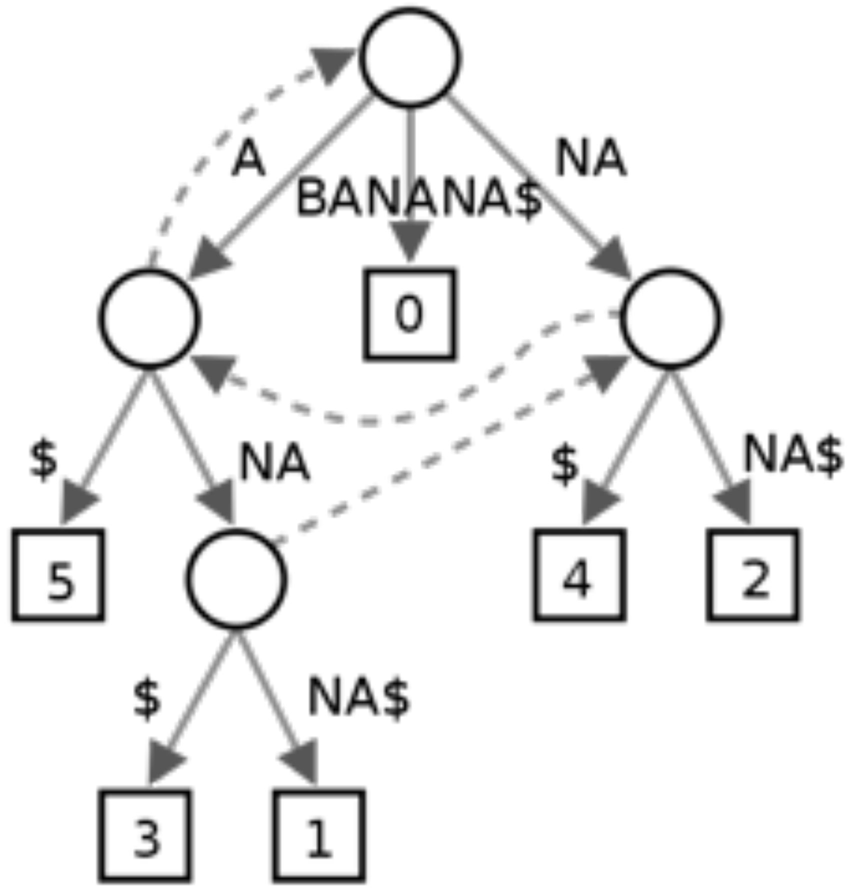
Indexing with suffixes

Some indices (e.g. the inverted index) are based on extracting substrings from *T*

A very different approach is to extract *suffixes* from *T*. This will lead us to some interesting and practical index data structures:



Suffix Trie



Suffix Tree

6	$
5	A$
3	ANA$
1	ANANA$
0	BANANA$
4	NA$
2	NANA$

Suffix Array

$ BANAN**A**
A $BANA**N**
A NA$BA**N**
A NANA$**B**
B ANANA**$**
N A$BAN**A**
N ANA$B**A**

FM Index

Suffix trie

Build a **trie** containing all **suffixes** of a text *T*

T: GTTATAGCTGATCGCGGCGTAGCGG

GTTATAGCTGATCGCGGCGTAGCGG
TTATAGCTGATCGCGGCGTAGCGG
TATAGCTGATCGCGGCGTAGCGG
ATAGCTGATCGCGGCGTAGCGG
TAGCTGATCGCGGCGTAGCGG
AGCTGATCGCGGCGTAGCGG
GCTGATCGCGGCGTAGCGG
CTGATCGCGGCGTAGCGG
TGATCGCGGCGTAGCGG
GATCGCGGCGTAGCGG
ATCGCGGCGTAGCGG
TCGCGGCGTAGCGG
CGCGGCGTAGCGG
GCGGCGTAGCGG
CGGCGTAGCGG
GGCGTAGCGG
GCGTAGCGG
CGTAGCGG
GTAGCGG
TAGCGG
AGCGG
GCGG
CGG
GG
G

$m(m+1)/2$ chars

Suffix trie

First add special *terminal character* **$** to the end of *T*

$ is a character that does not appear elsewhere in *T*, and we define it to be less than other characters (for DNA: **$** < **A** < **C** < **G** < **T**)

$** enforces a rule we're all used to using: e.g. "as" comes before "ash" in the dictionary. **$ <u>also guarantees no suffix is a prefix of any other suffix</u>.

```
T: G T T A T A G C T G A T C G C G G C G T A G C G G $
   G T T A T A G C T G A T C G C G G C G T A G C G G $
     T T A T A G C T G A T C G C G G C G T A G C G G $
       T A T A G C T G A T C G C G G C G T A G C G G $
         A T A G C T G A T C G C G G C G T A G C G G $
           T A G C T G A T C G C G G C G T A G C G G $
             A G C T G A T C G C G G C G T A G C G G $
               G C T G A T C G C G G C G T A G C G G $
                 C T G A T C G C G G C G T A G C G G $
                   T G A T C G C G G C G T A G C G G $
                     G A T C G C G G C G T A G C G G $
                       A T C G C G G C G T A G C G G $
                         T C G C G G C G T A G C G G $
                           C G C G G C G T A G C G G $
                             G C G G C G T A G C G G $
                               C G G C G T A G C G G $
                                 G G C G T A G C G G $
                                   G C G T A G C G G $
```

Suffix trie

T: aba

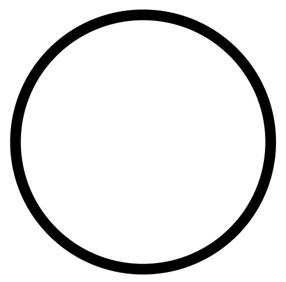
What's the suffix trie?

Suffix trie

T: aba$

What's the suffix trie?

Suffix trie



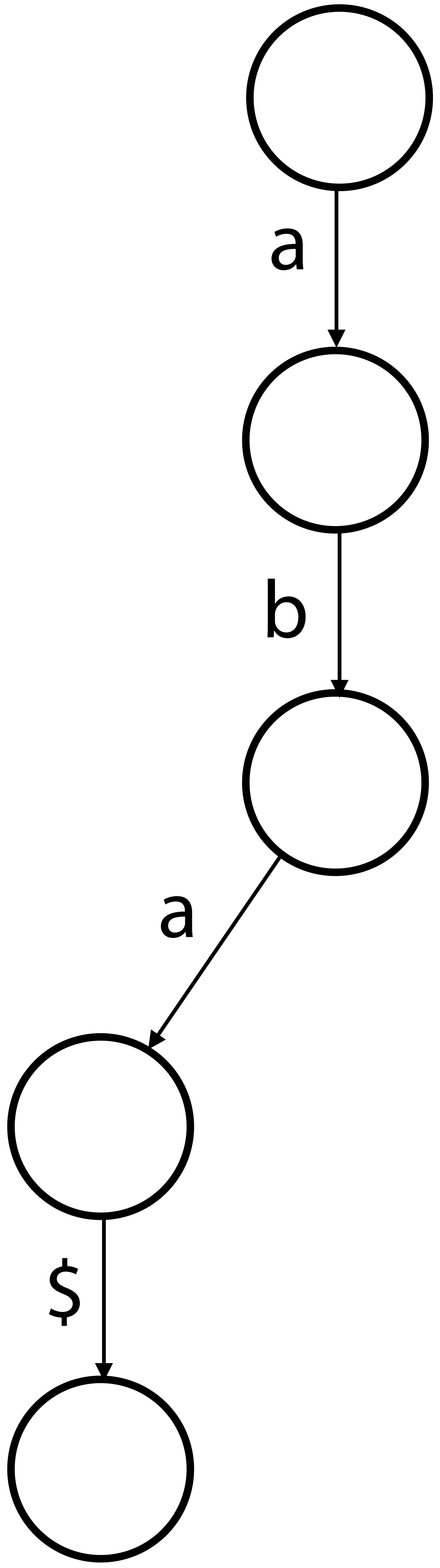
T: aba$

What's the suffix trie?

Suffix trie

T: aba$

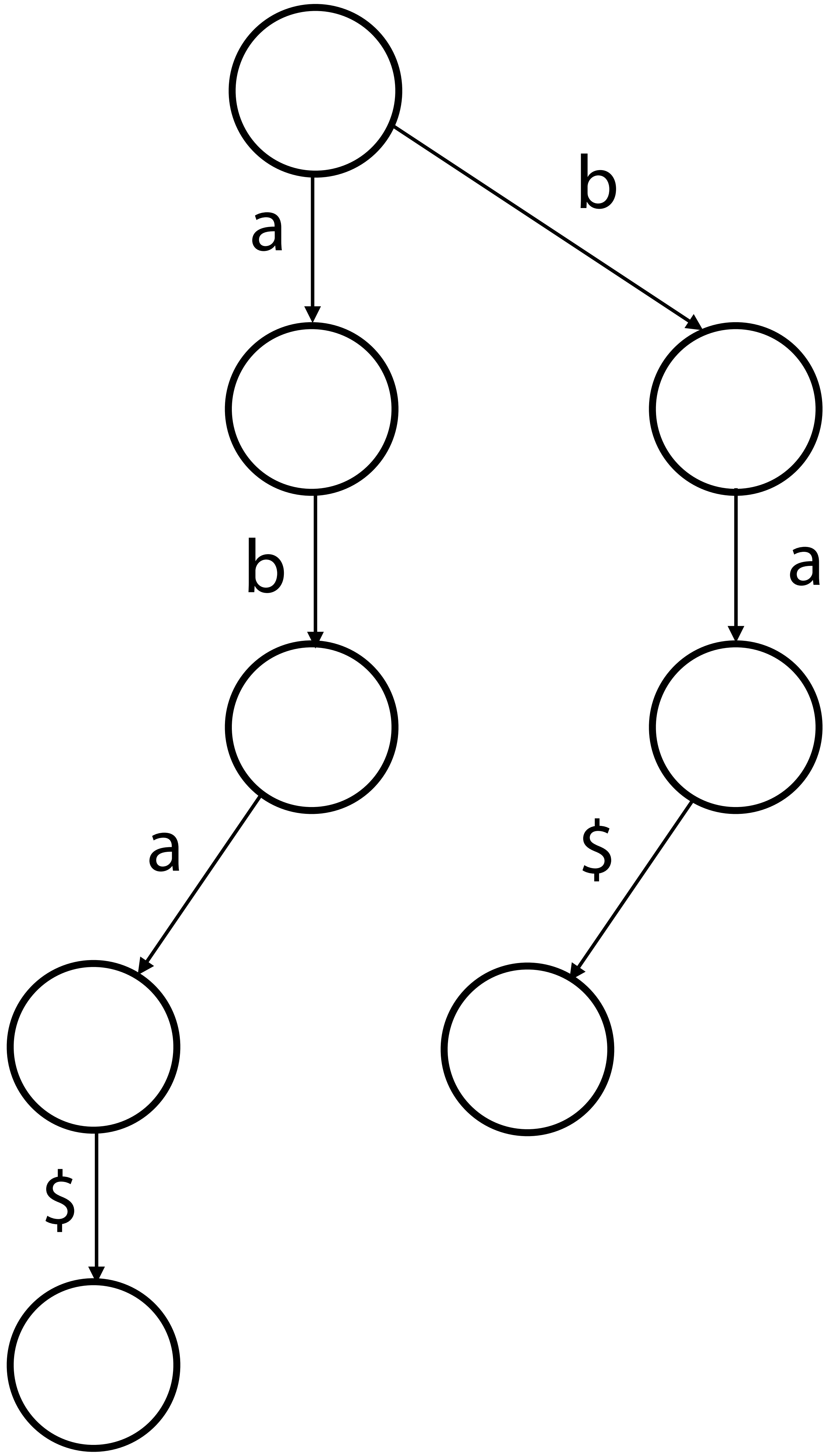
What's the suffix trie?



Suffix trie

T: aba$

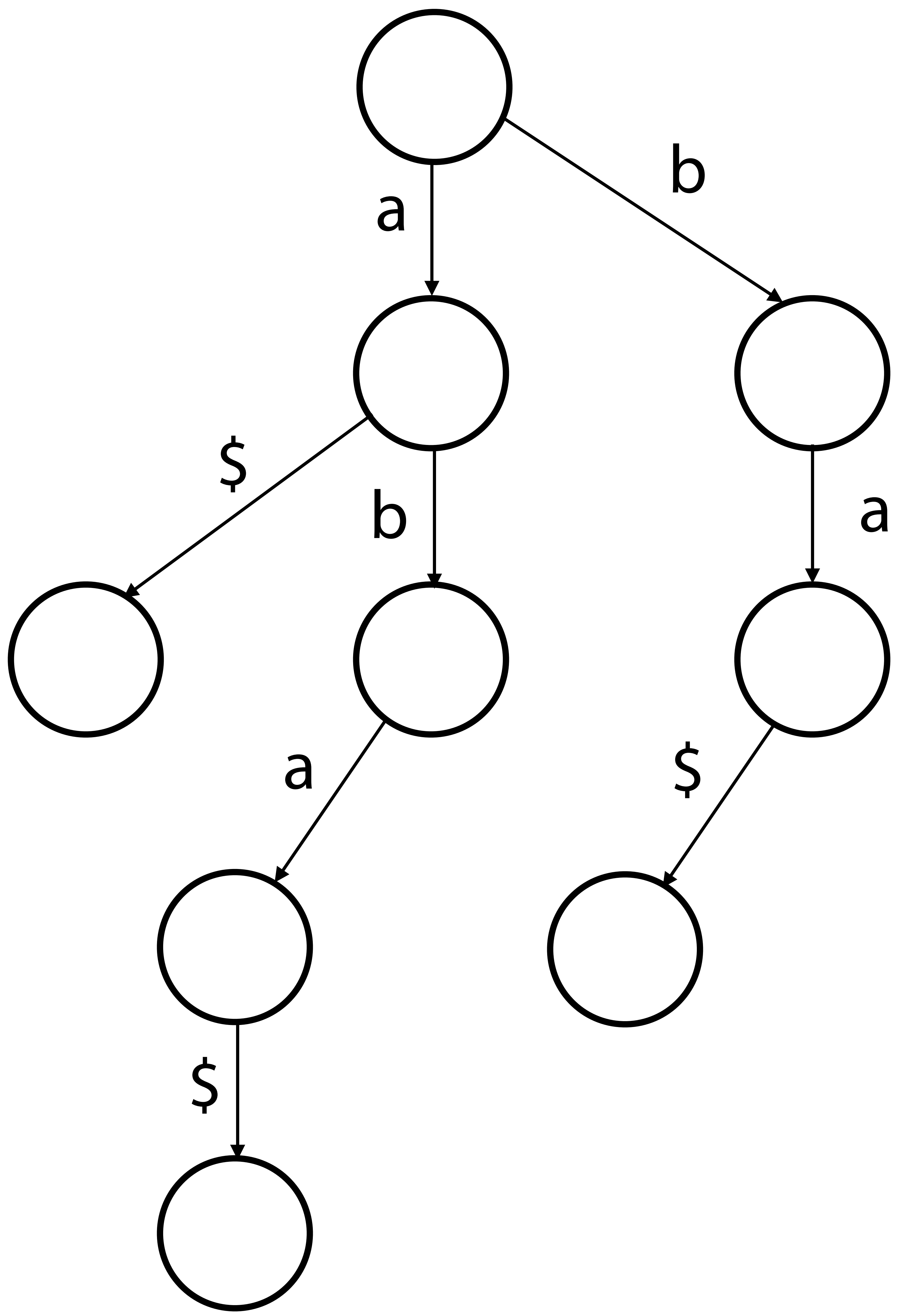
What's the suffix trie?



Suffix trie

T: aba$

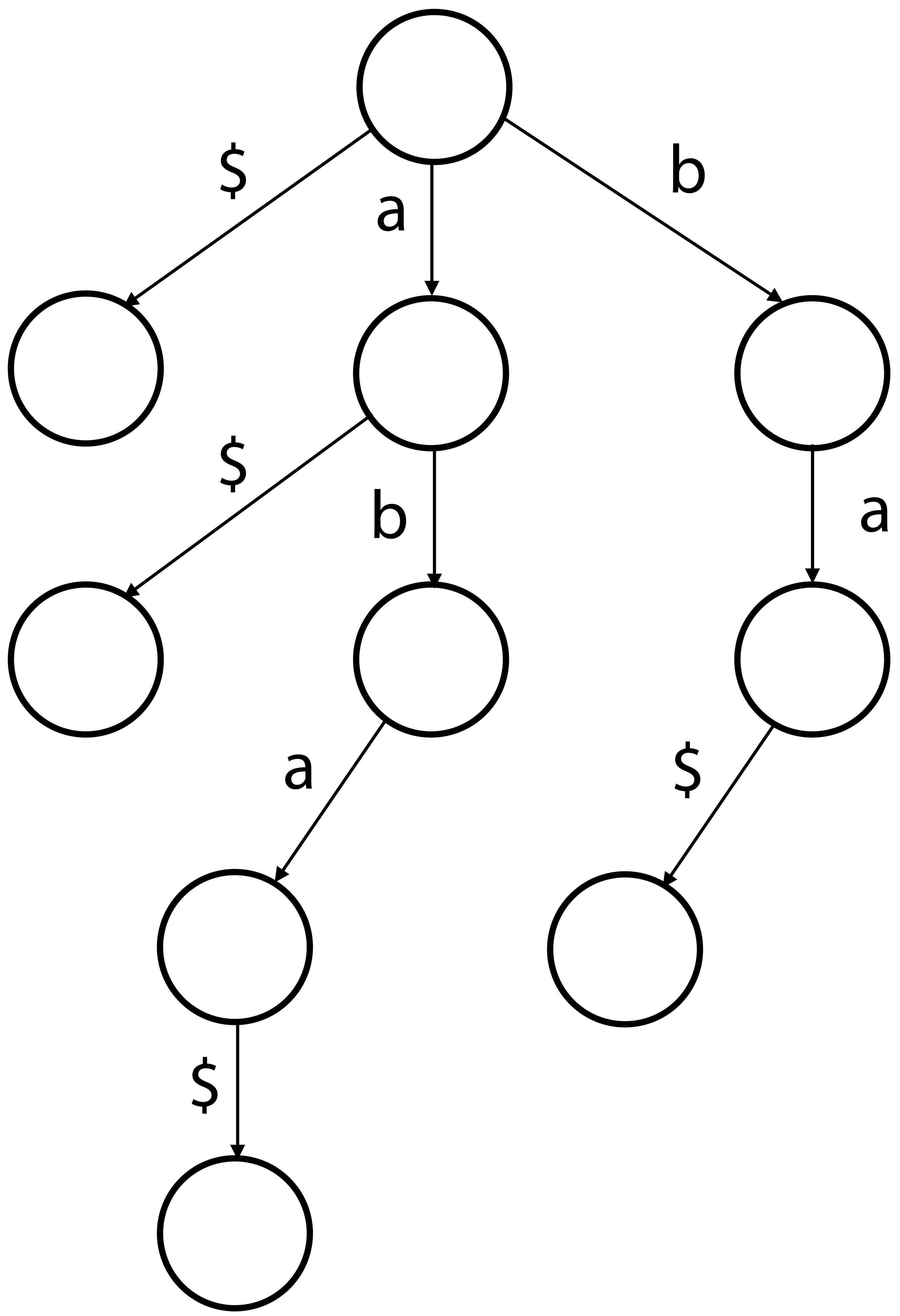
What's the suffix trie?



Suffix trie

T: aba$

What’s the suffix trie?



Suffix trie

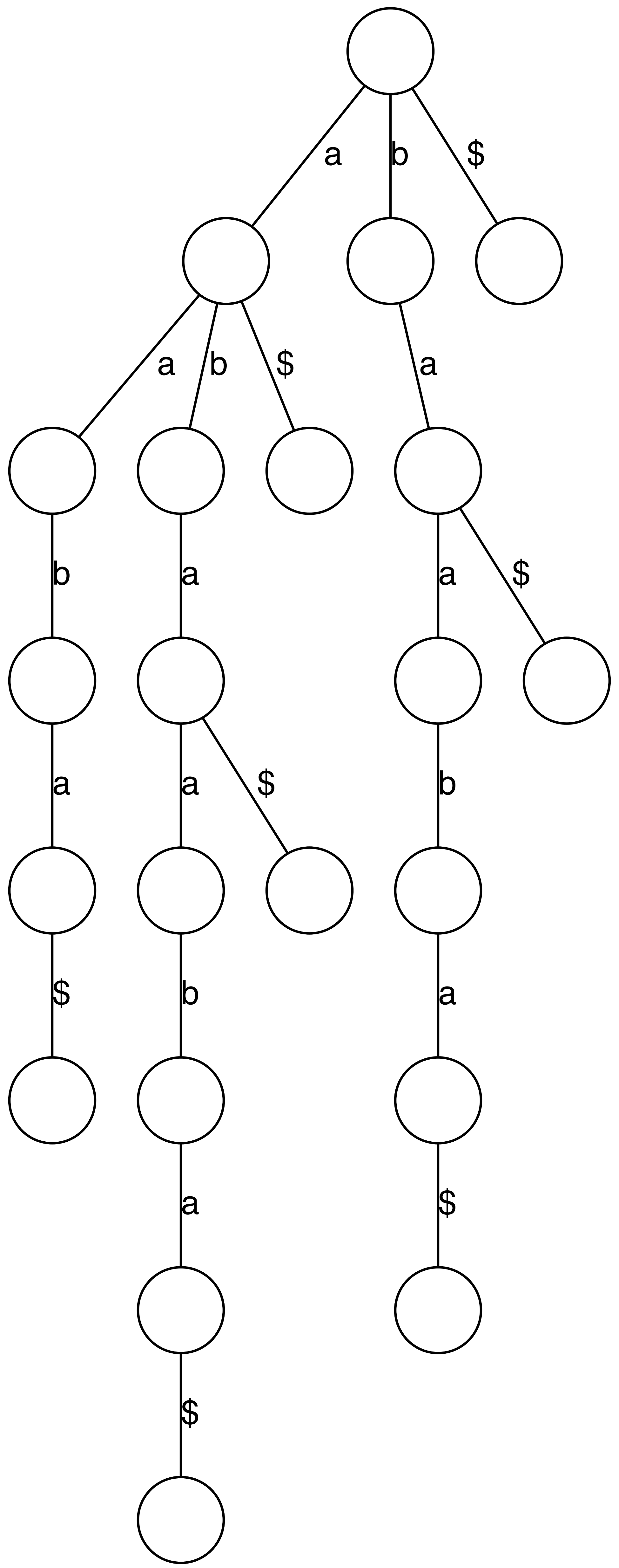
T: abaaba

Suffix trie

T: abaaba$

Suffix trie

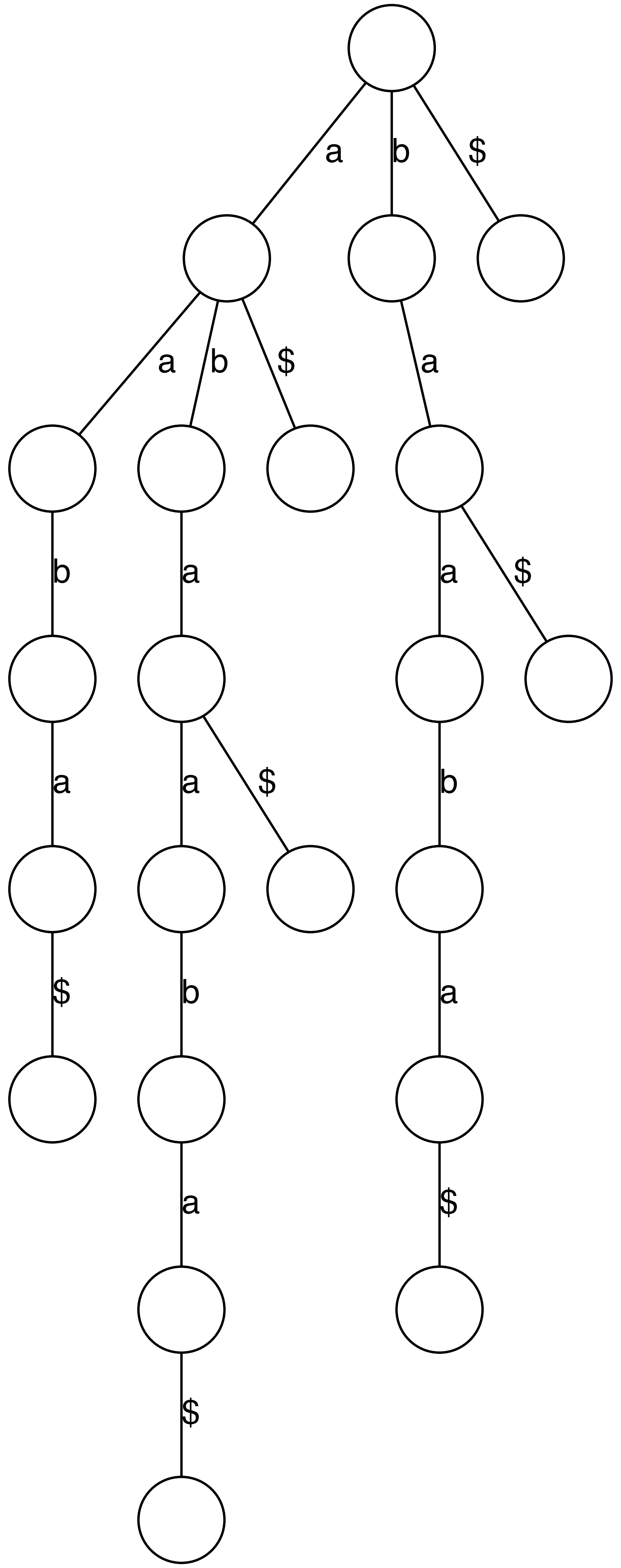
T: abaaba$



Suffix trie

T: abaaba$

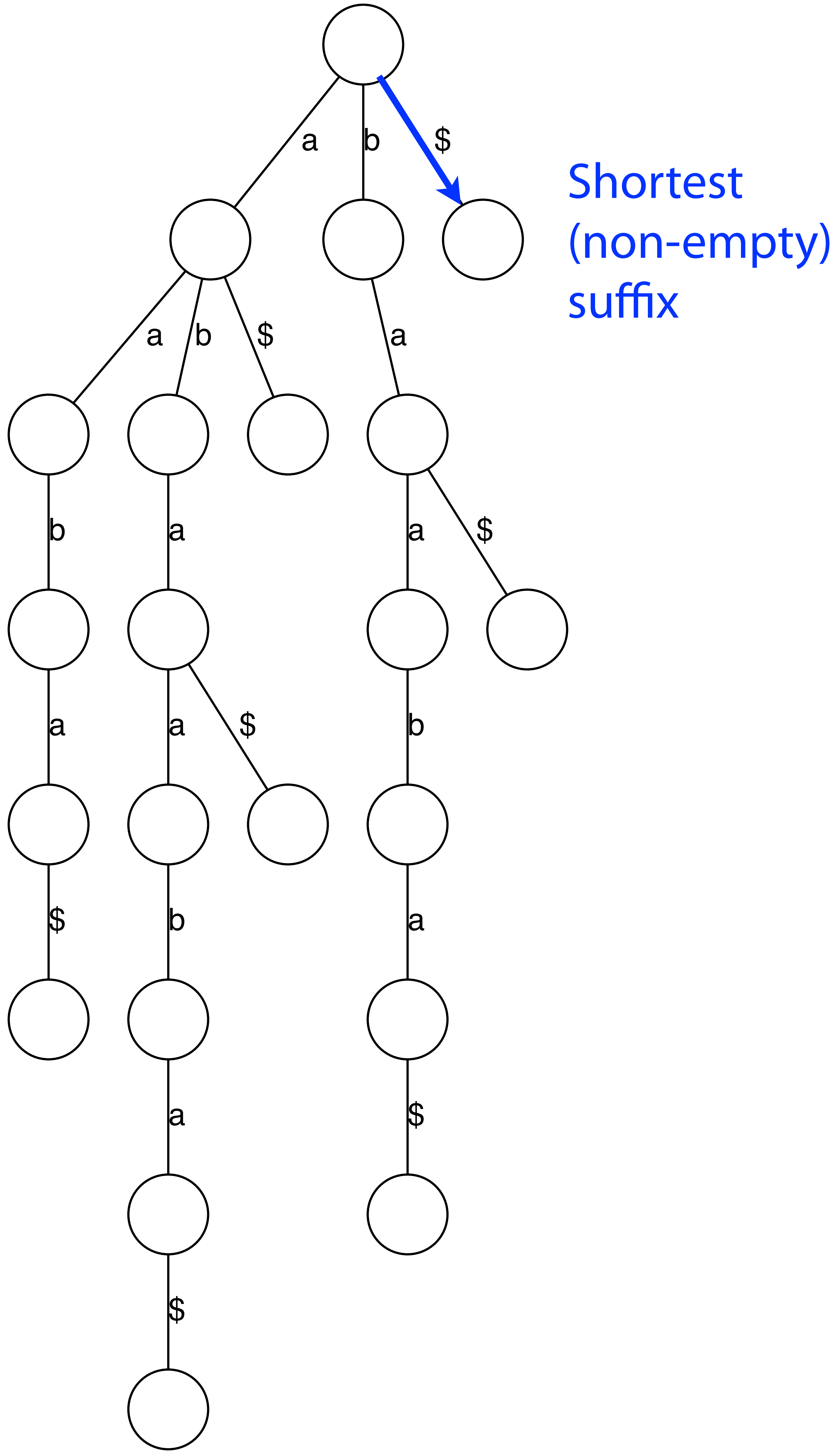
Each path from root to leaf represents a suffix; each suffix is represented by some path from root to leaf



Suffix trie

T: abaaba$

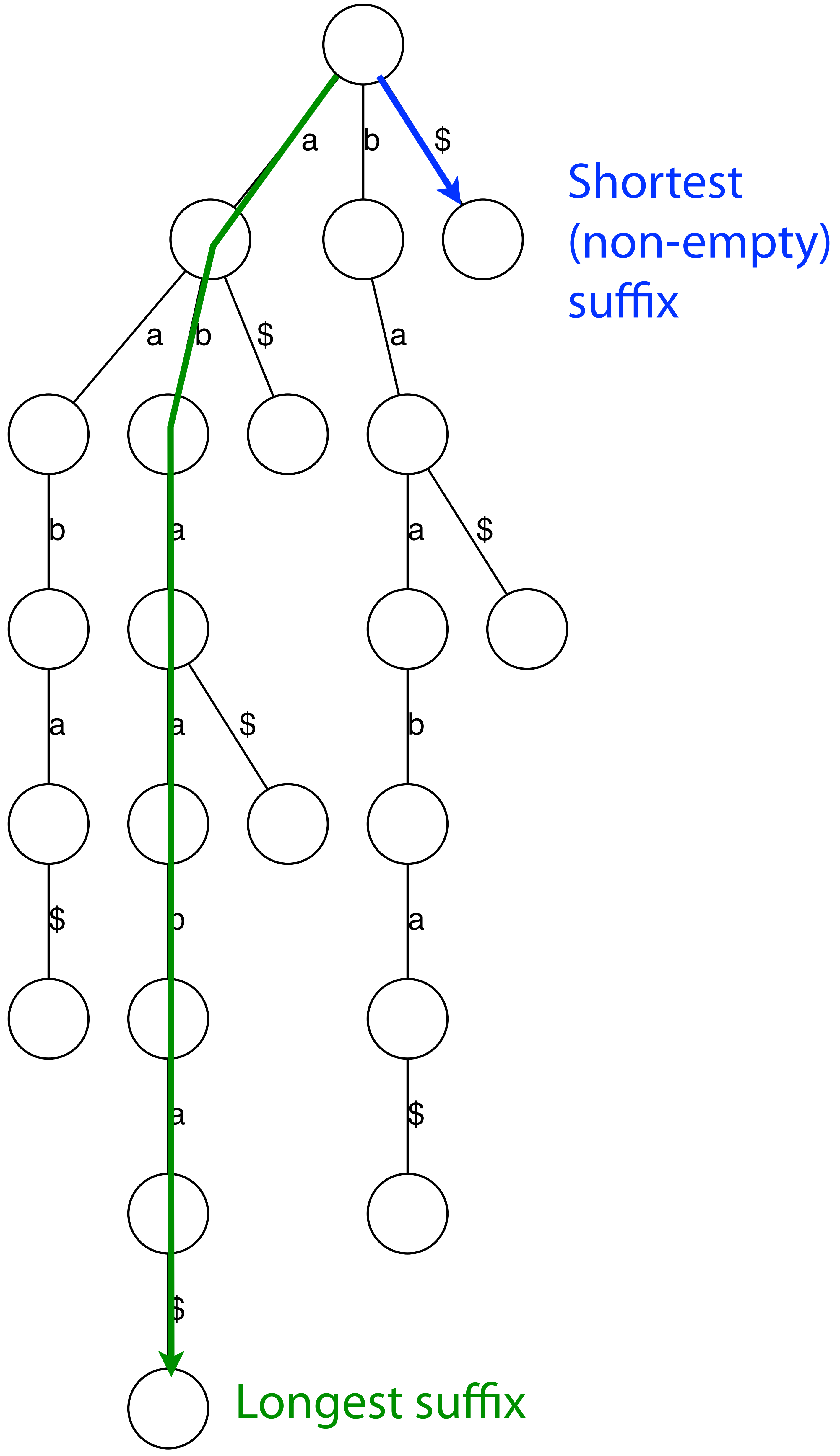
Each path from root to leaf represents a suffix; each suffix is represented by some path from root to leaf



Suffix trie

T: abaaba$

Each path from root to leaf represents a suffix; each suffix is represented by some path from root to leaf

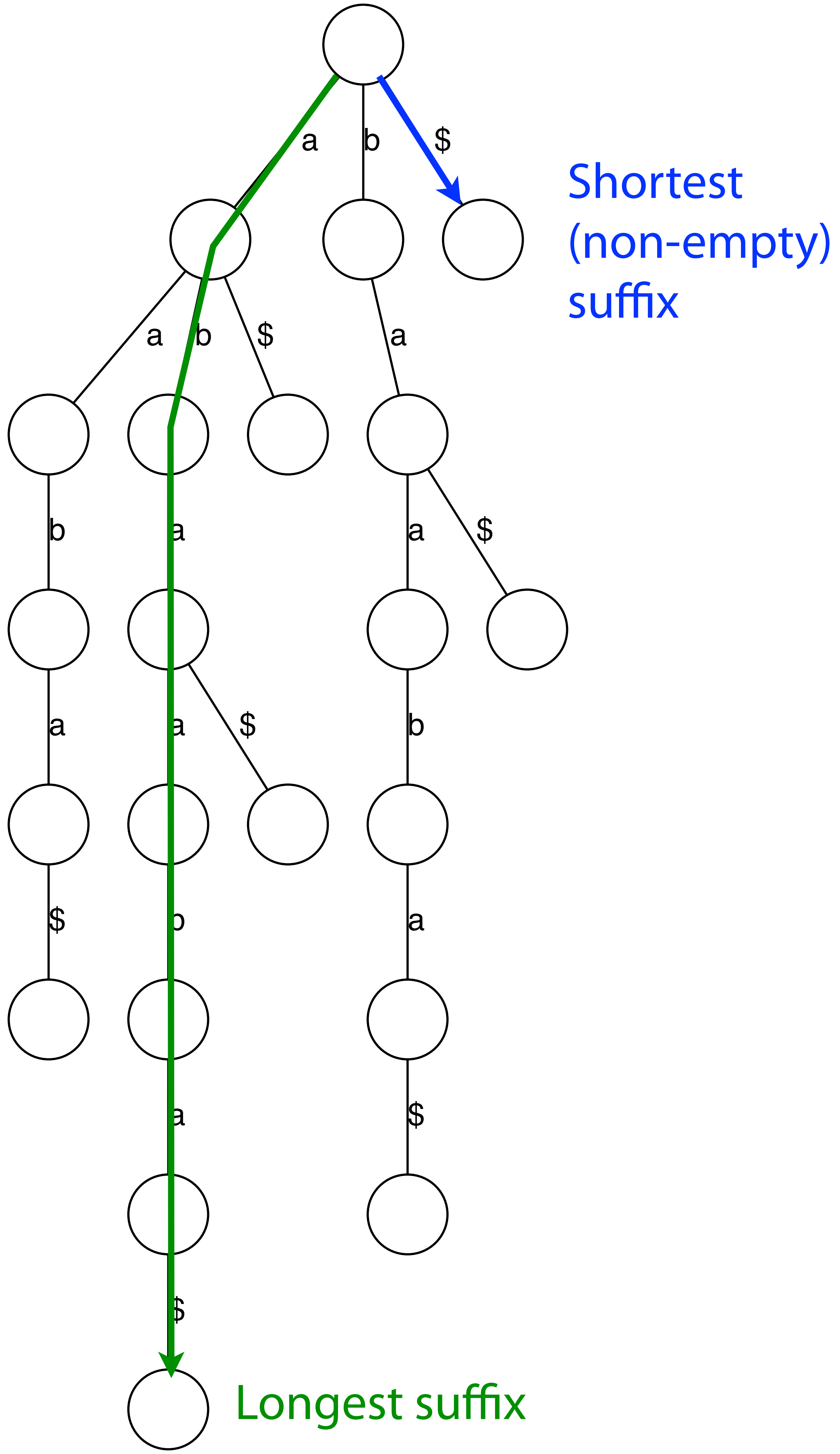


Suffix trie

T: abaaba$

Each path from root to leaf represents a suffix; each suffix is represented by some path from root to leaf

Would this still be the case if we hadn't added $?

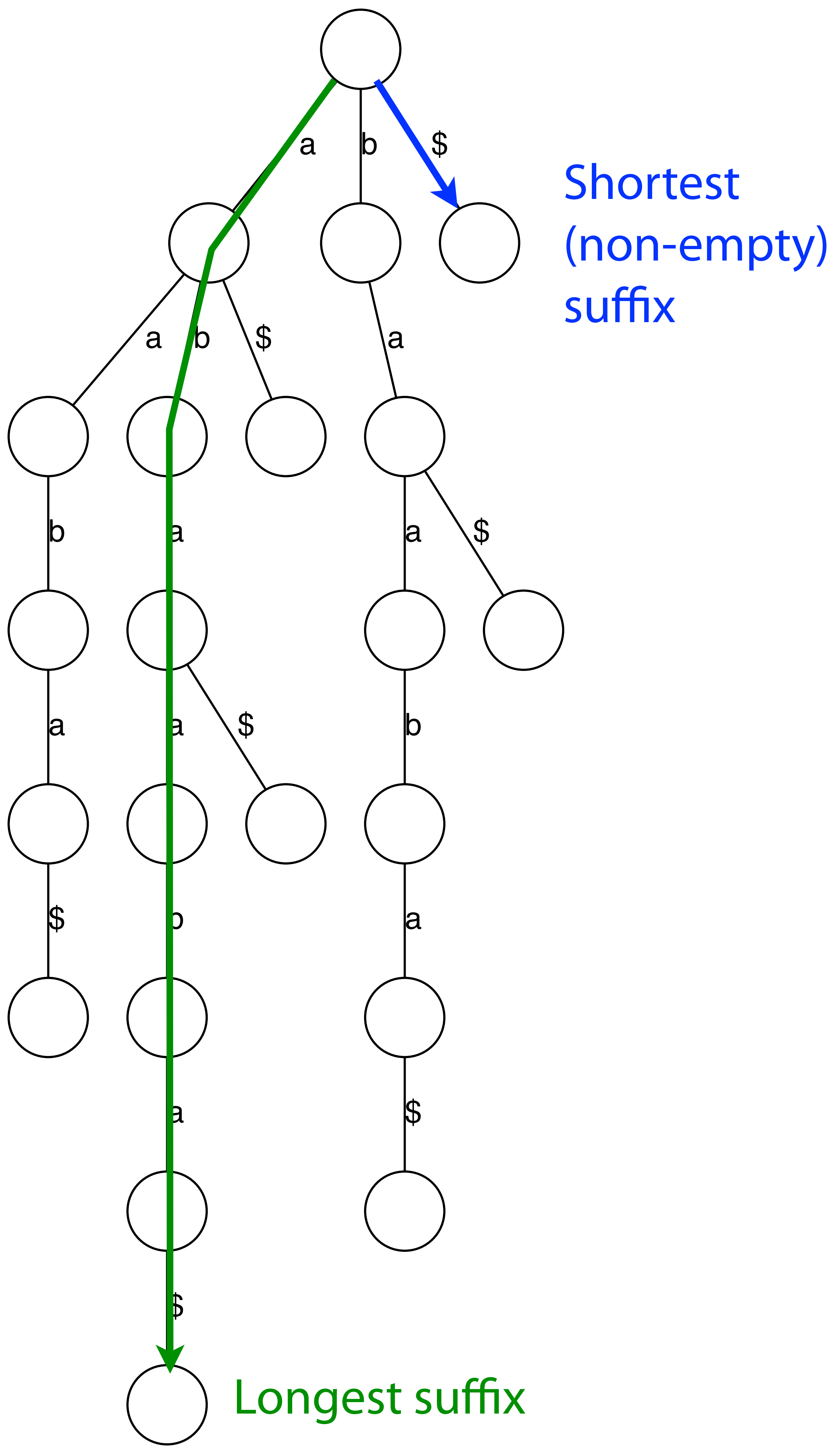


Suffix trie

T: abaaba *T*$: abaaba$

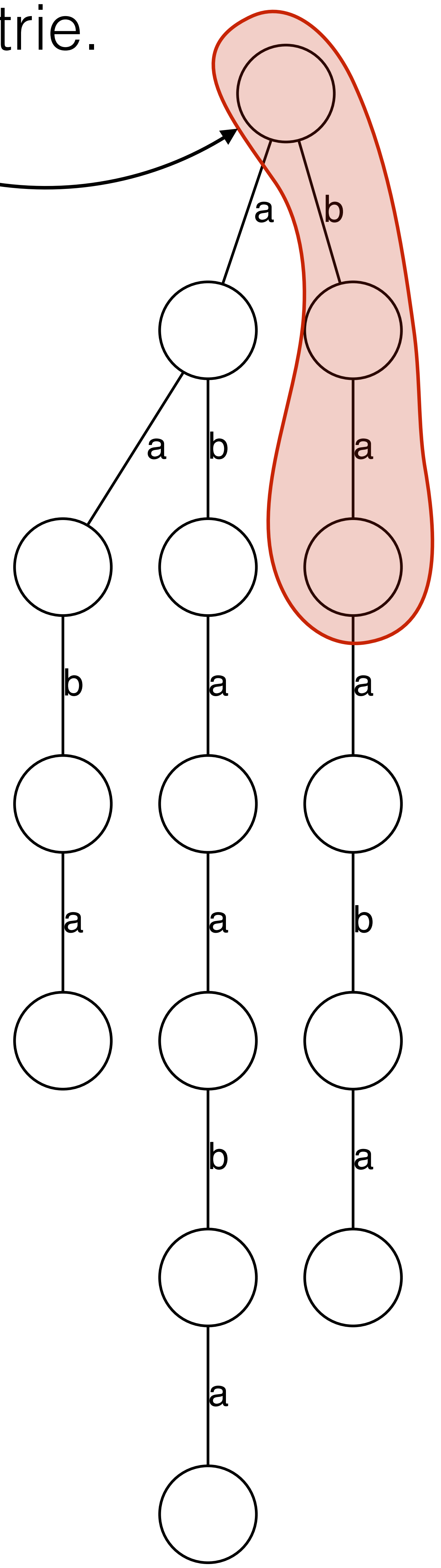
Each path from root to leaf represents a suffix; each suffix is represented by some path from root to leaf

Would this still be the case if we hadn't added $?



Suffix trie

Without $, no way to
spell out this suffix & end
at a leaf in the trie.



T: abaaba

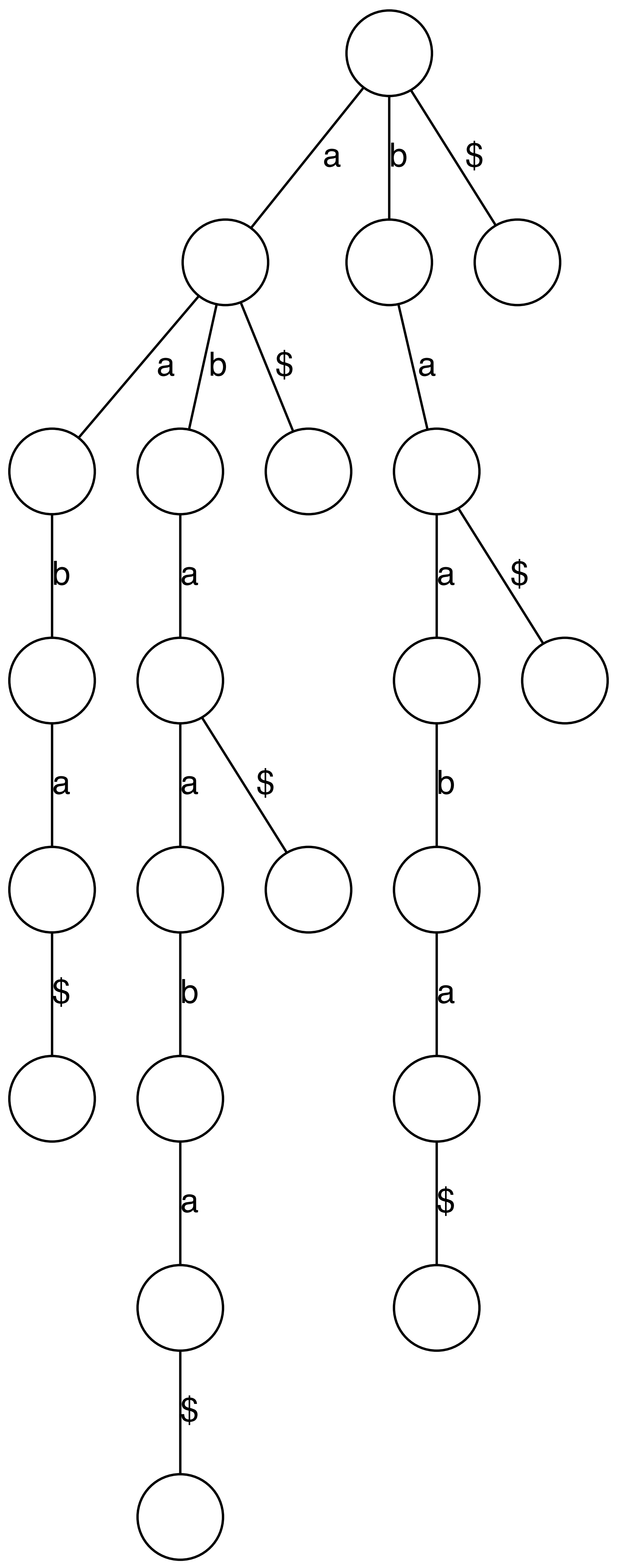
Each path from root to leaf represents a suffix; each suffix is represented by some path from root to leaf

Would this still be the case if we hadn't added **$**? **No**

Without the $, we have problems; e.g. here "ba" is a prefix of "baaba"

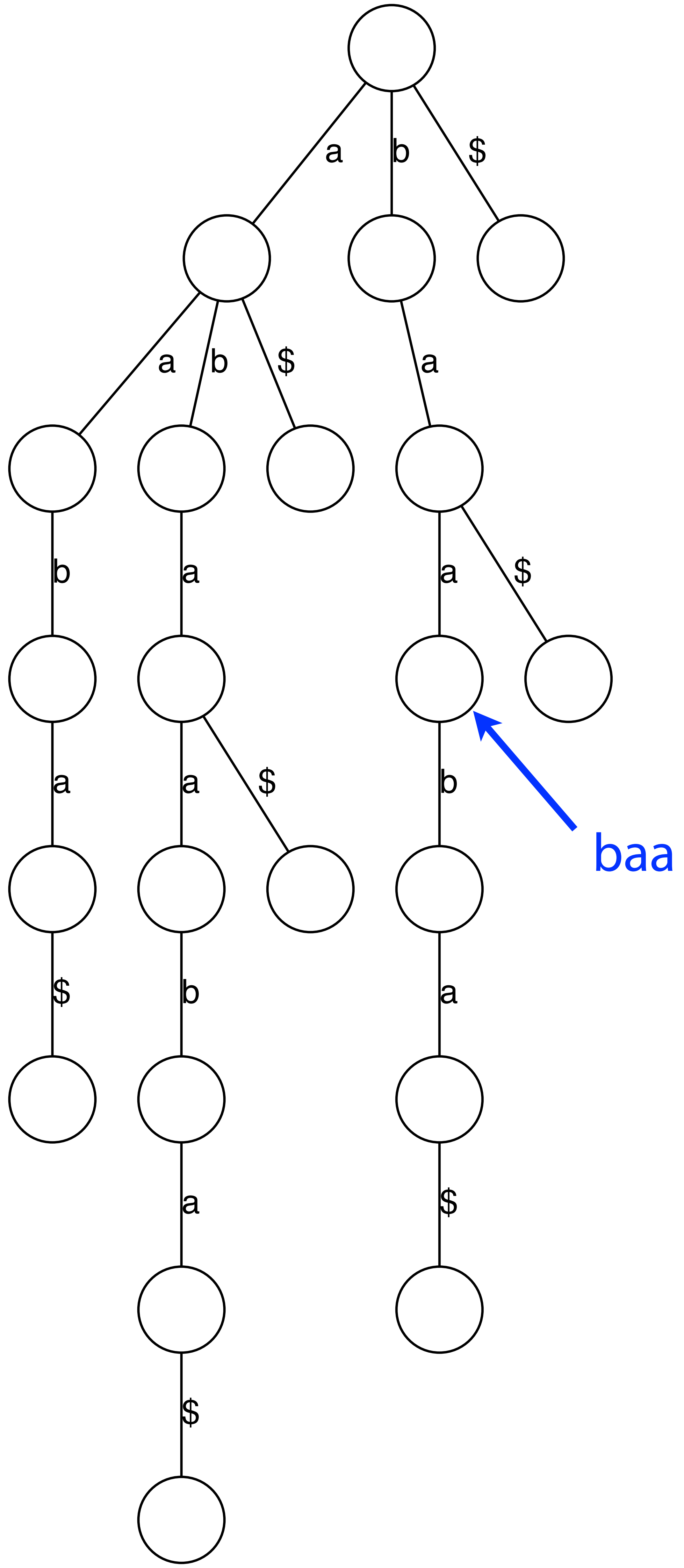
Suffix trie

Think of each node as having a label, spelling out characters on path from root to node



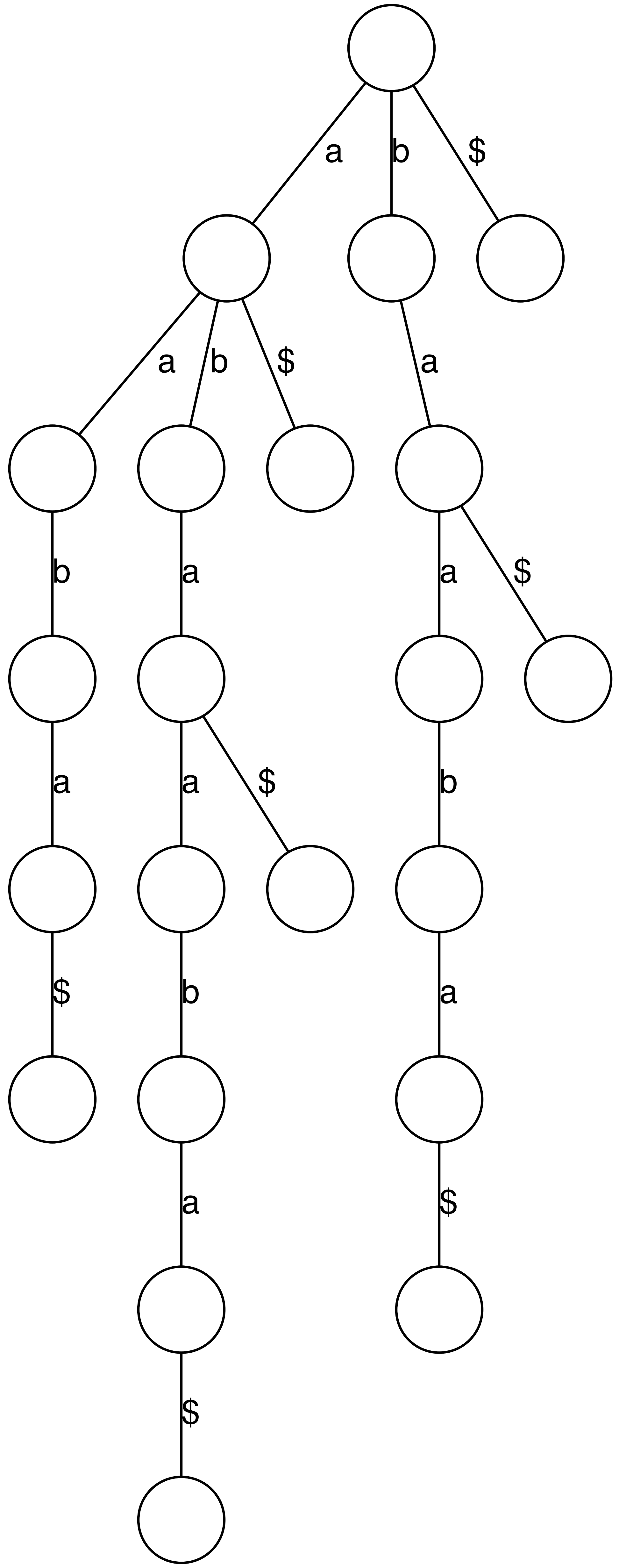
Suffix trie

Think of each node as having a label, spelling out characters on path from root to node



Suffix trie

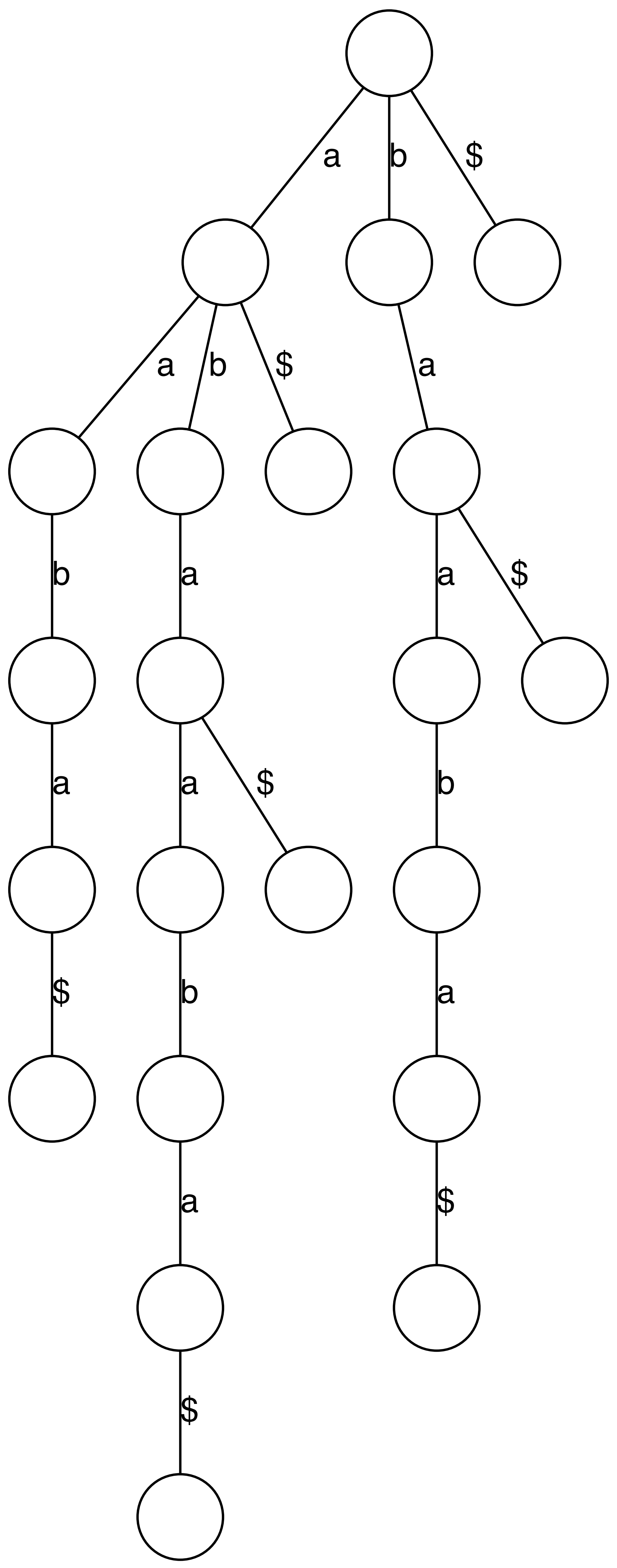
How do we check whether a string S is a substring of T?



Suffix trie

How do we check whether a string S is a substring of T?

Note: Each of T's substrings is spelled out along a path from the root.

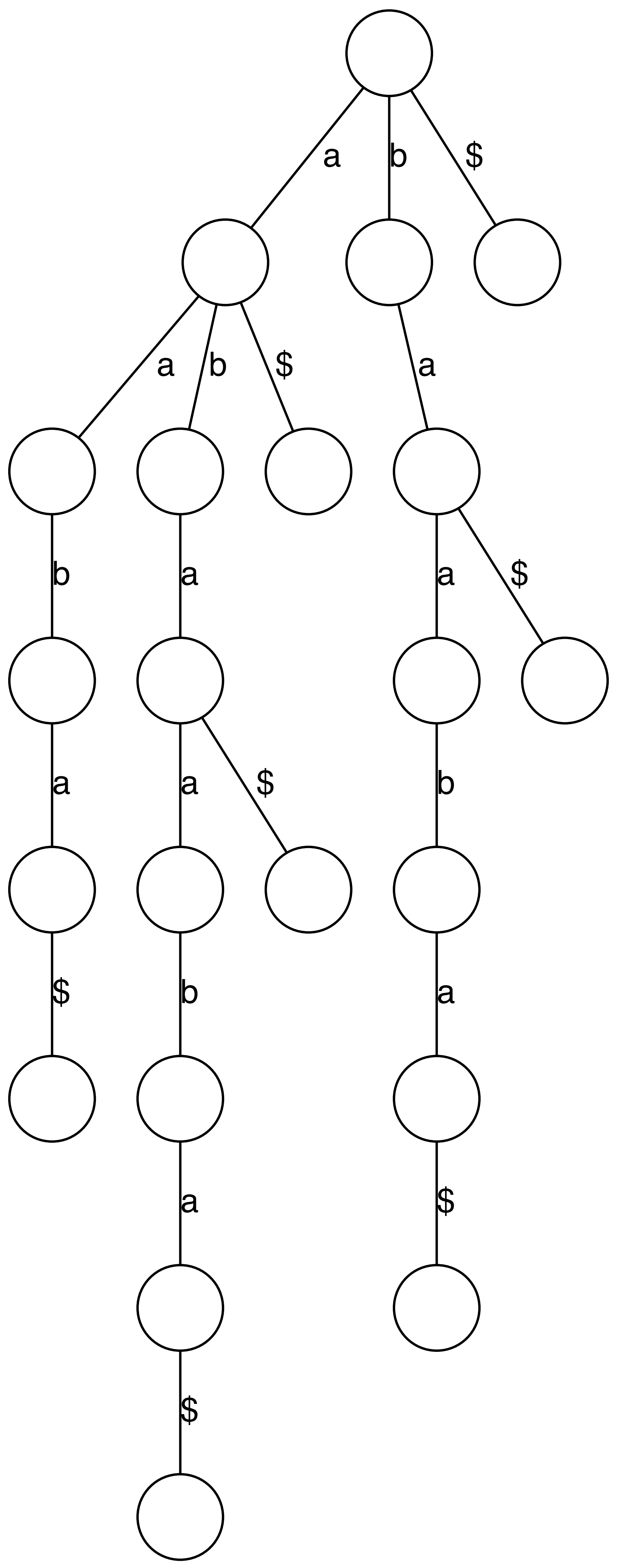


Suffix trie

How do we check whether a string S is a substring of T?

Note: Each of T's substrings is spelled out along a path from the root.

Every substring is a prefix of some suffix of T.



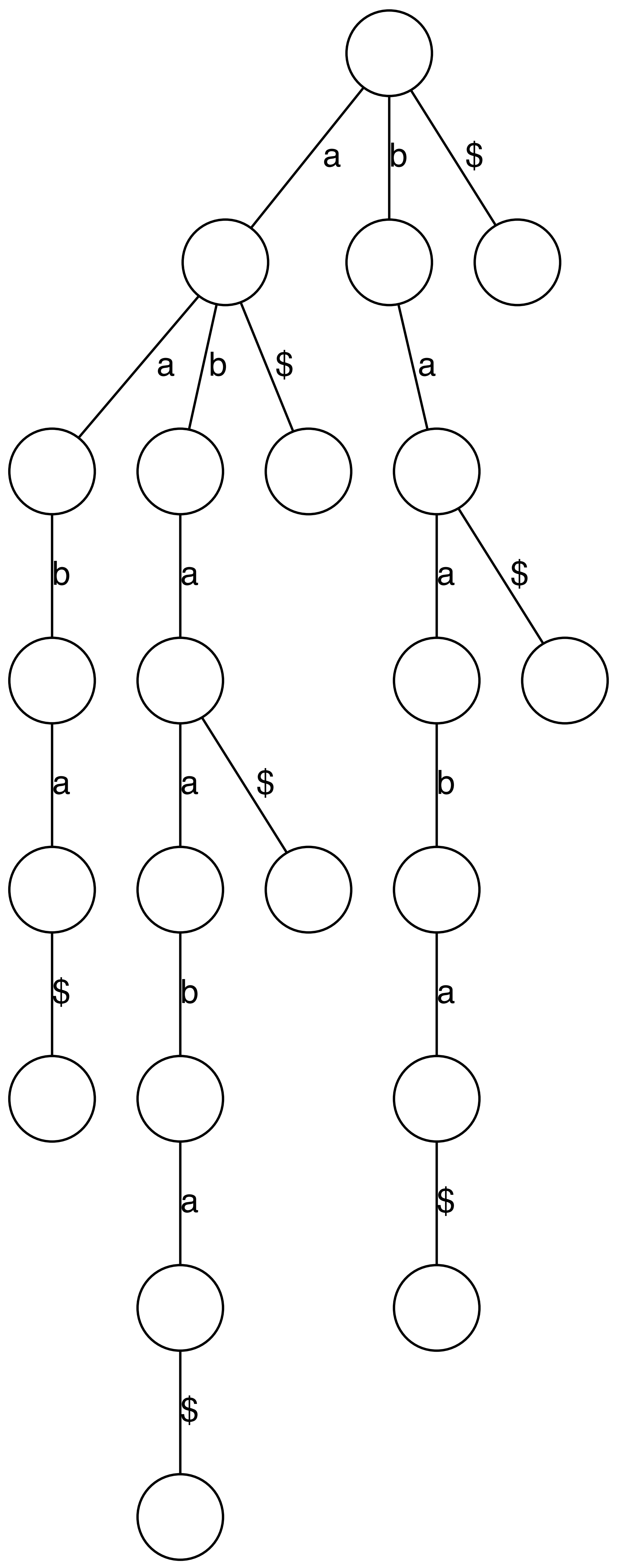
Suffix trie

How do we check whether a string S is a substring of T?

Note: Each of T's substrings is spelled out along a path from the root.

Every substring is a prefix of some suffix of T.

Start at the root and follow the edges labeled with the characters of S



Suffix trie

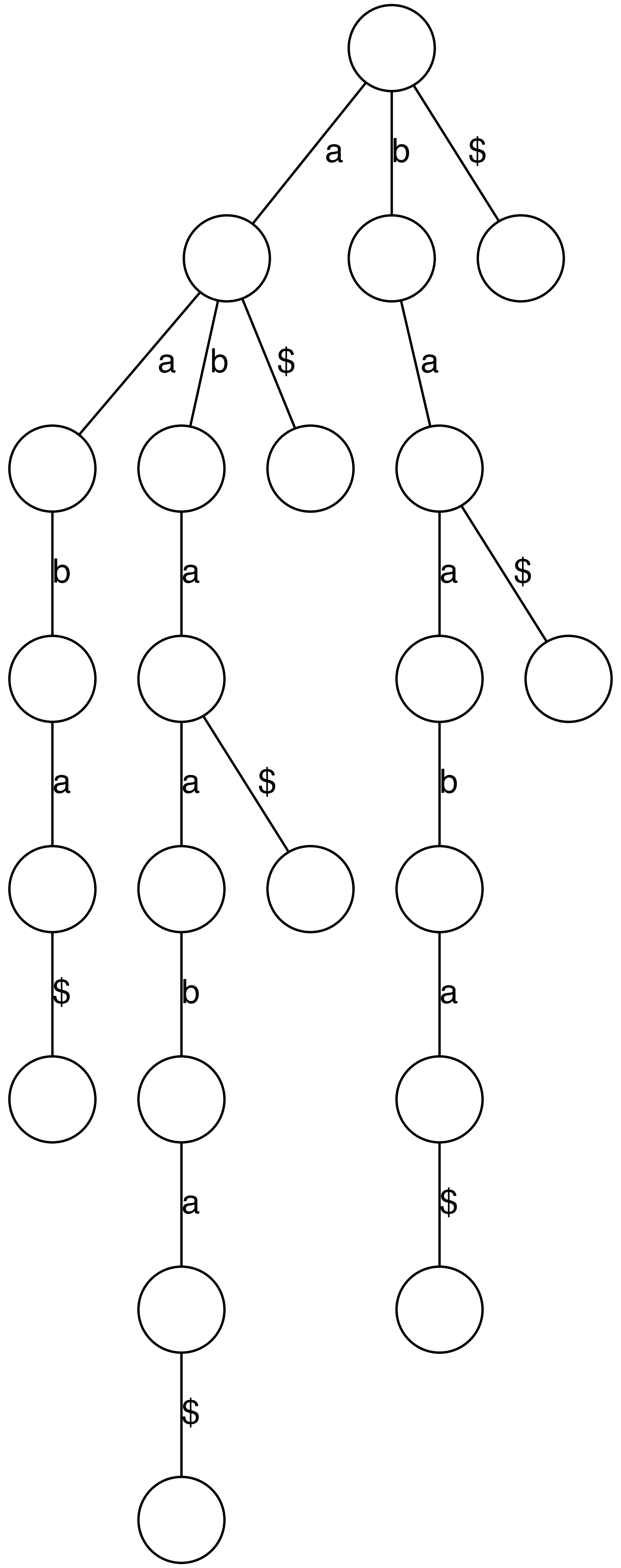
How do we check whether a string S is a substring of T?

Note: Each of T's substrings is spelled out along a path from the root.

Every substring is a prefix of some suffix of T.

Start at the root and follow the edges labeled with the characters of S

If we "fall off" the trie -- i.e. there is no outgoing edge for next character of S, then S is not a substring of T



Suffix trie

How do we check whether a string S is a substring of T?

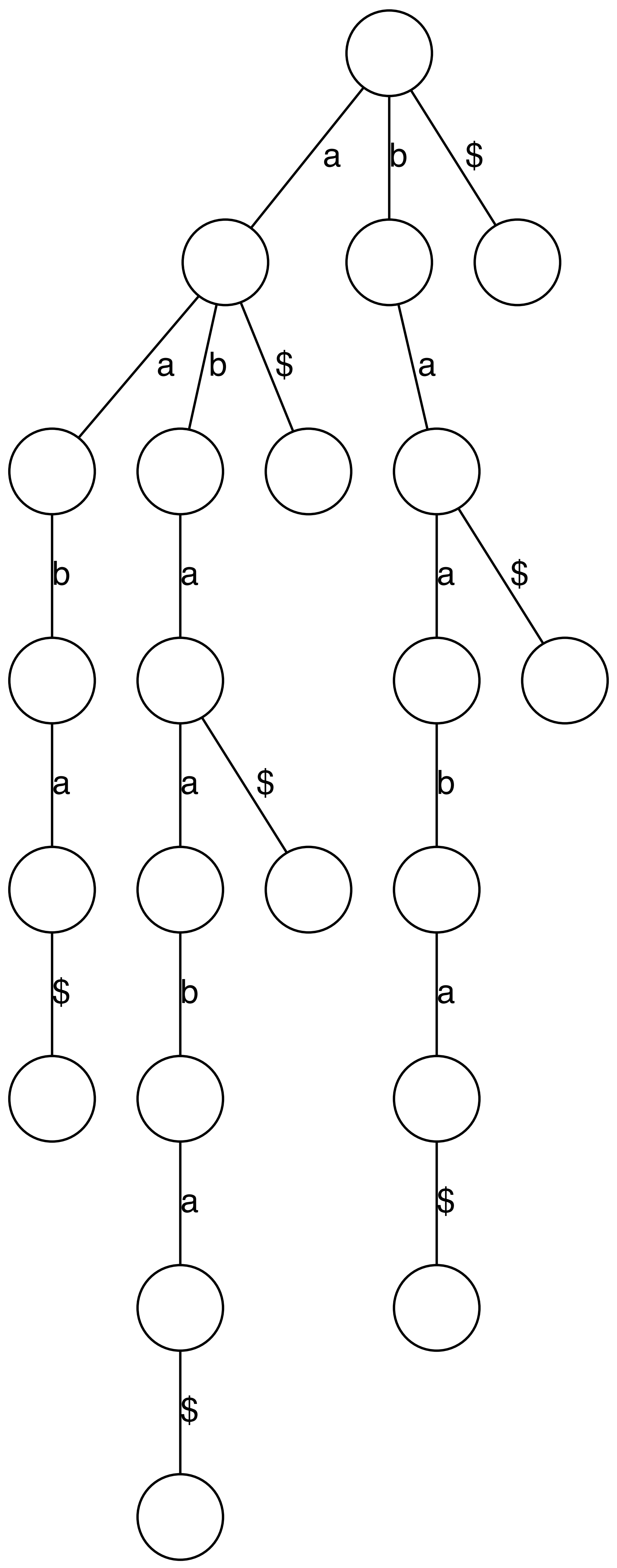
Note: Each of T's substrings is spelled out along a path from the root.

Every substring is a prefix of some suffix of T.

Start at the root and follow the edges labeled with the characters of S

If we "fall off" the trie -- i.e. there is no outgoing edge for next character of S, then S is not a substring of T

If we exhaust S without falling off, S is a substring of T



Suffix trie

How do we check whether a string S is a substring of T?

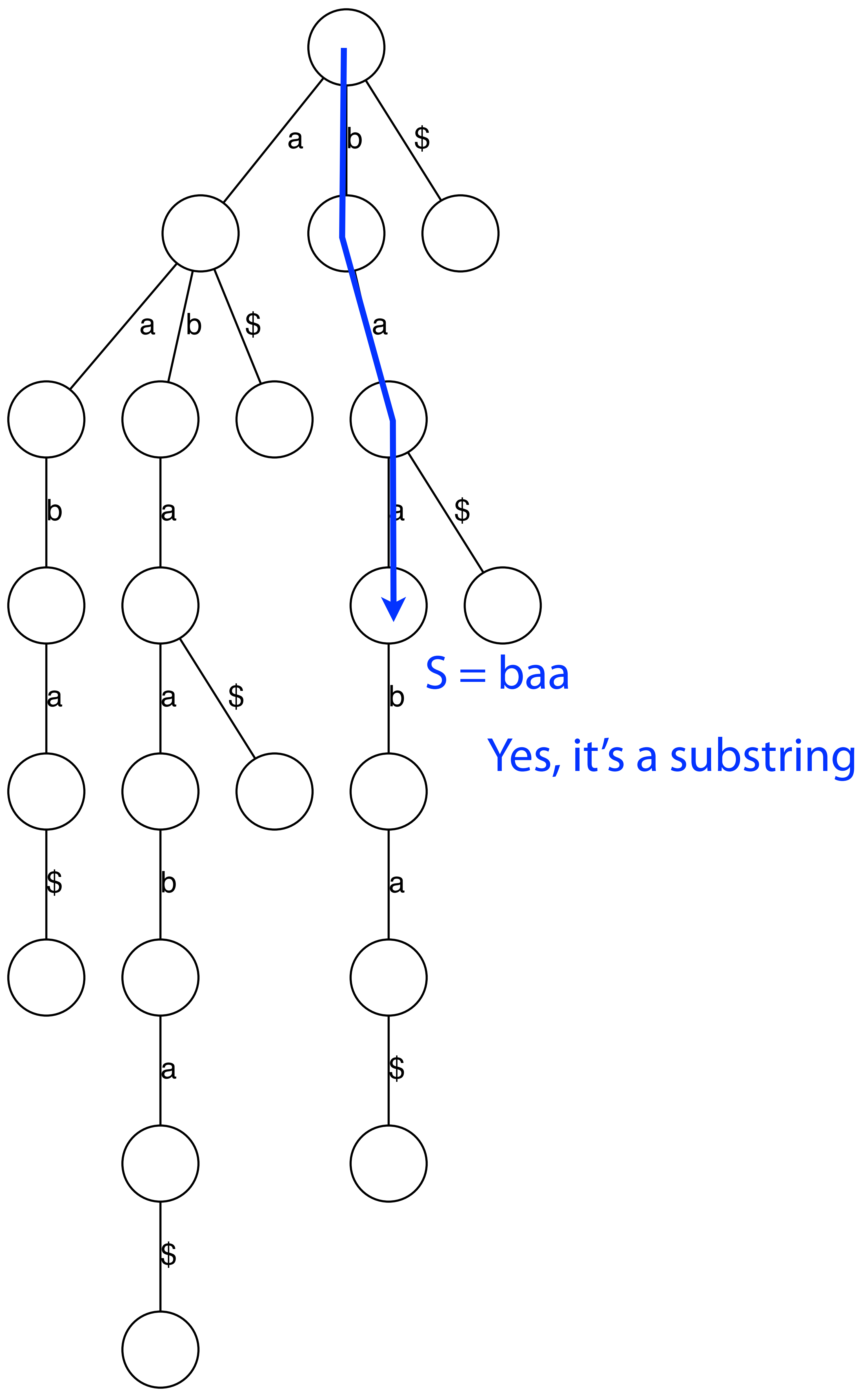
Note: Each of T's substrings is spelled out along a path from the root.

Every substring is a prefix of some suffix of T.

Start at the root and follow the edges labeled with the characters of S

If we "fall off" the trie -- i.e. there is no outgoing edge for next character of S, then S is not a substring of T

If we exhaust S without falling off, S is a substring of T



Suffix trie

How do we check whether a string S is a substring of T?

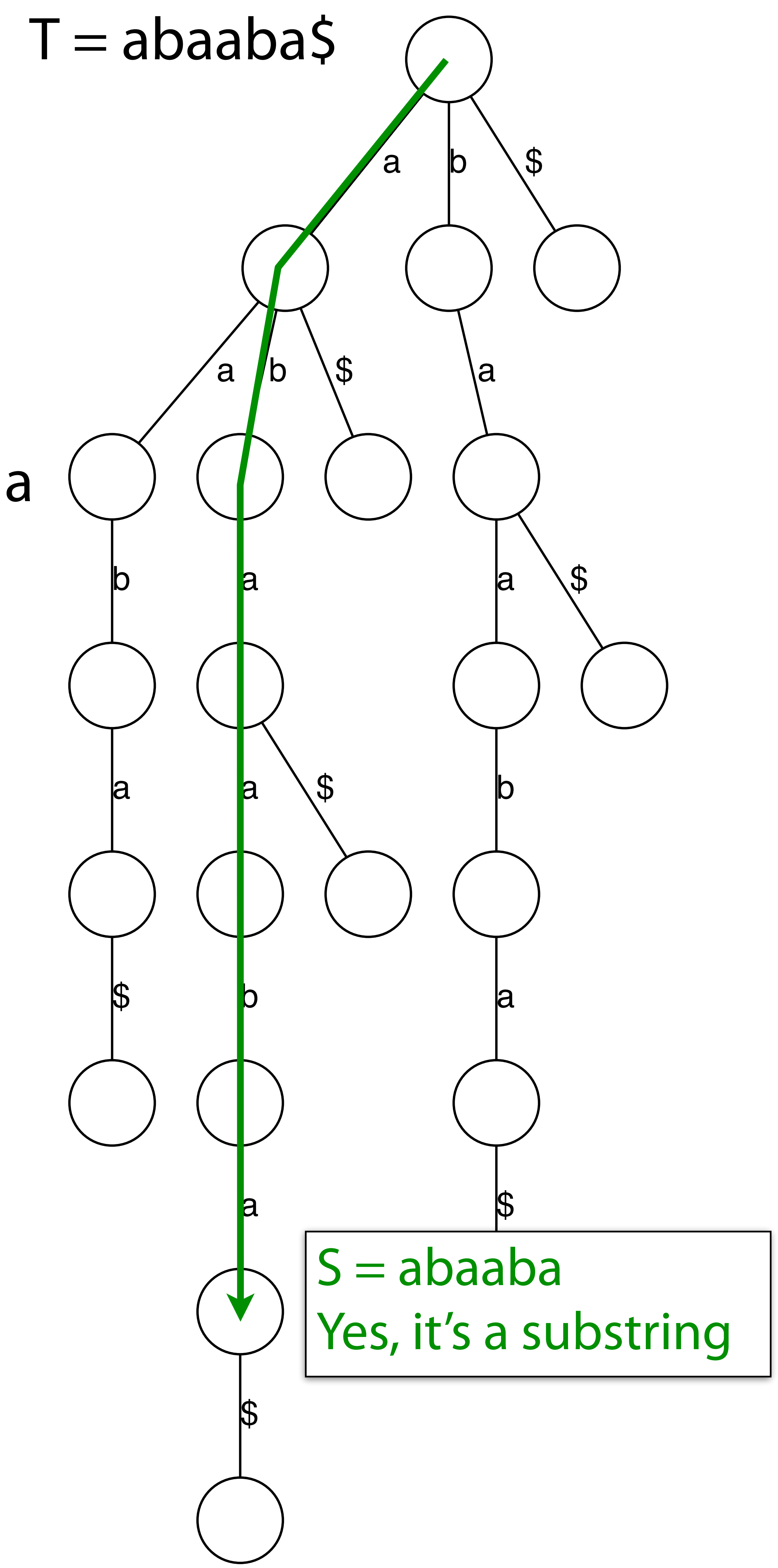
Note: Each of T's substrings is spelled out along a path from the root.

Every substring is a prefix of some suffix of T.

Start at the root and follow the edges labeled with the characters of S

If we "fall off" the trie -- i.e. there is no outgoing edge for next character of S, then S is not a substring of T

If we exhaust S without falling off, S is a substring of T



Suffix trie

How do we check whether a string S is a substring of T?

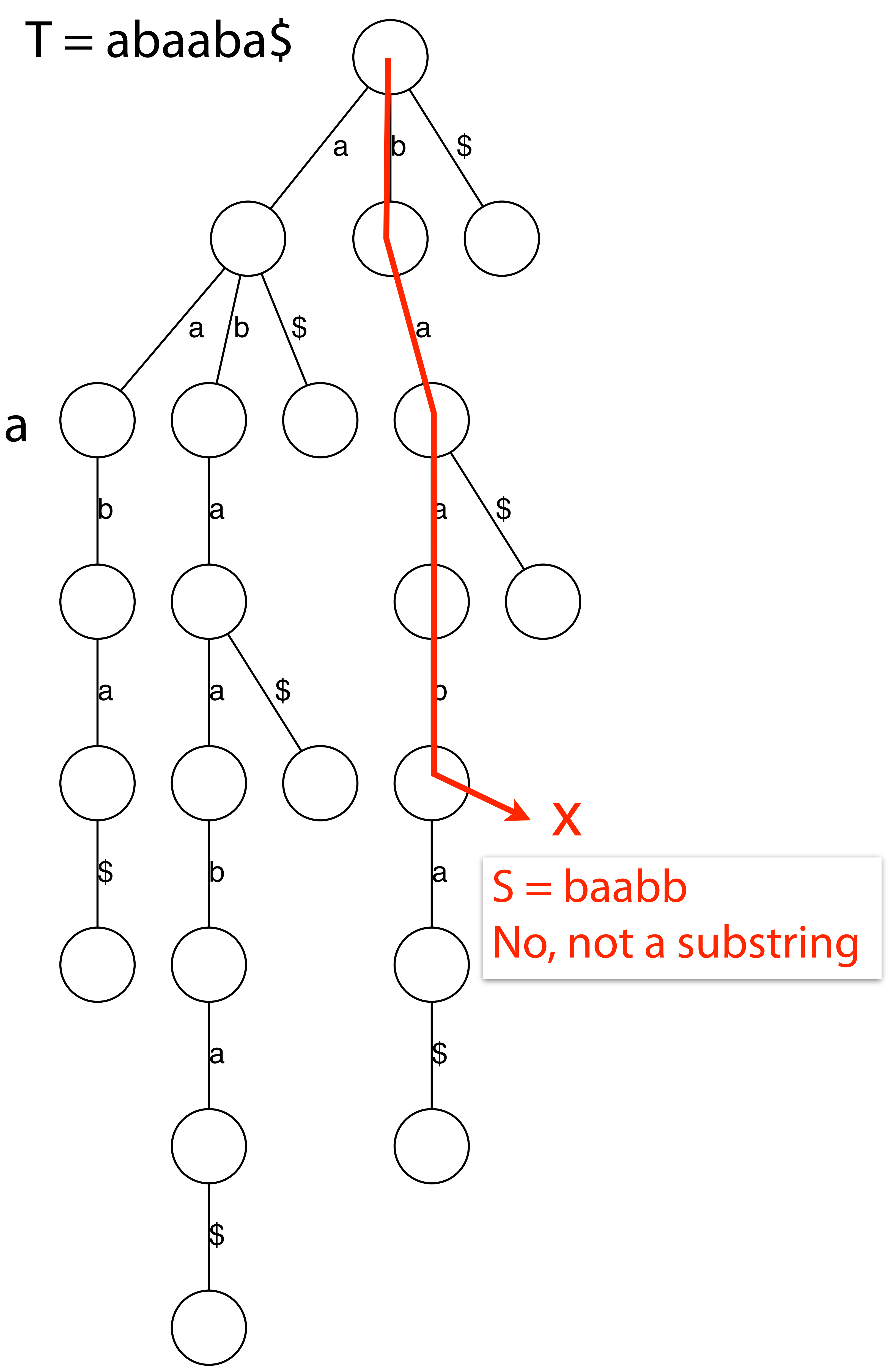
Note: Each of T's substrings is spelled out along a path from the root.

Every substring is a prefix of some suffix of T.

Start at the root and follow the edges labeled with the characters of S

If we "fall off" the trie -- i.e. there is no outgoing edge for next character of S, then S is not a substring of T

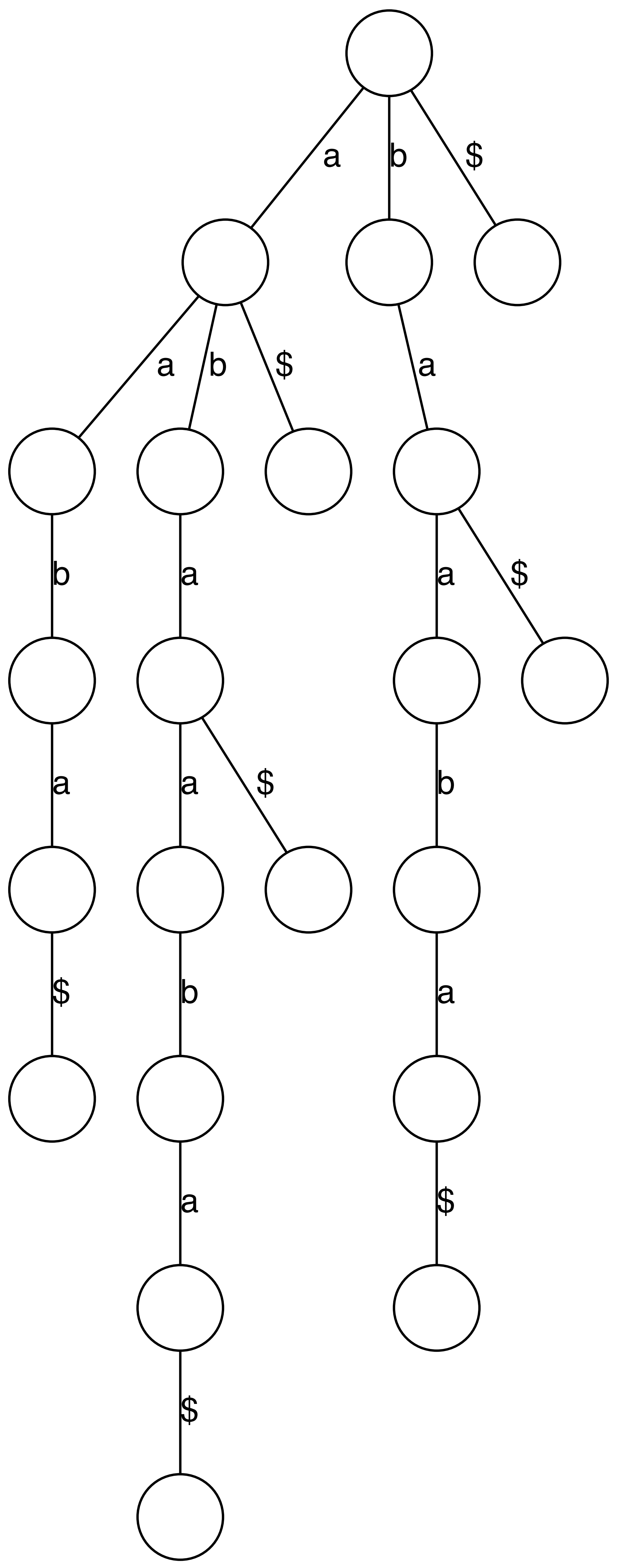
If we exhaust S without falling off, S is a substring of T



Suffix trie

How do we check whether a string S is a suffix of T?

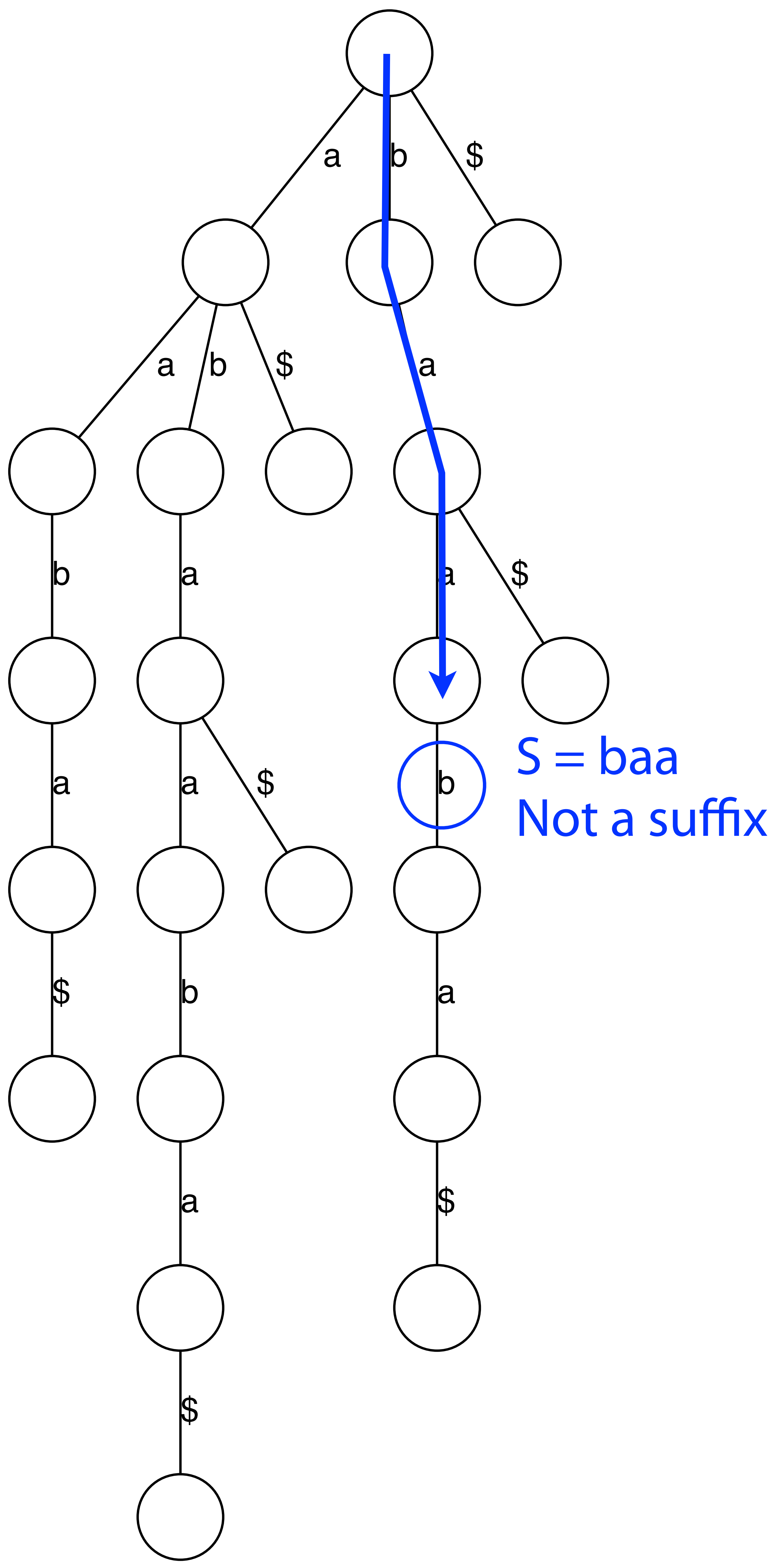
Same procedure as for substring, but additionally check terminal node for $ child



Suffix trie

How do we check whether a string S is a suffix of T?

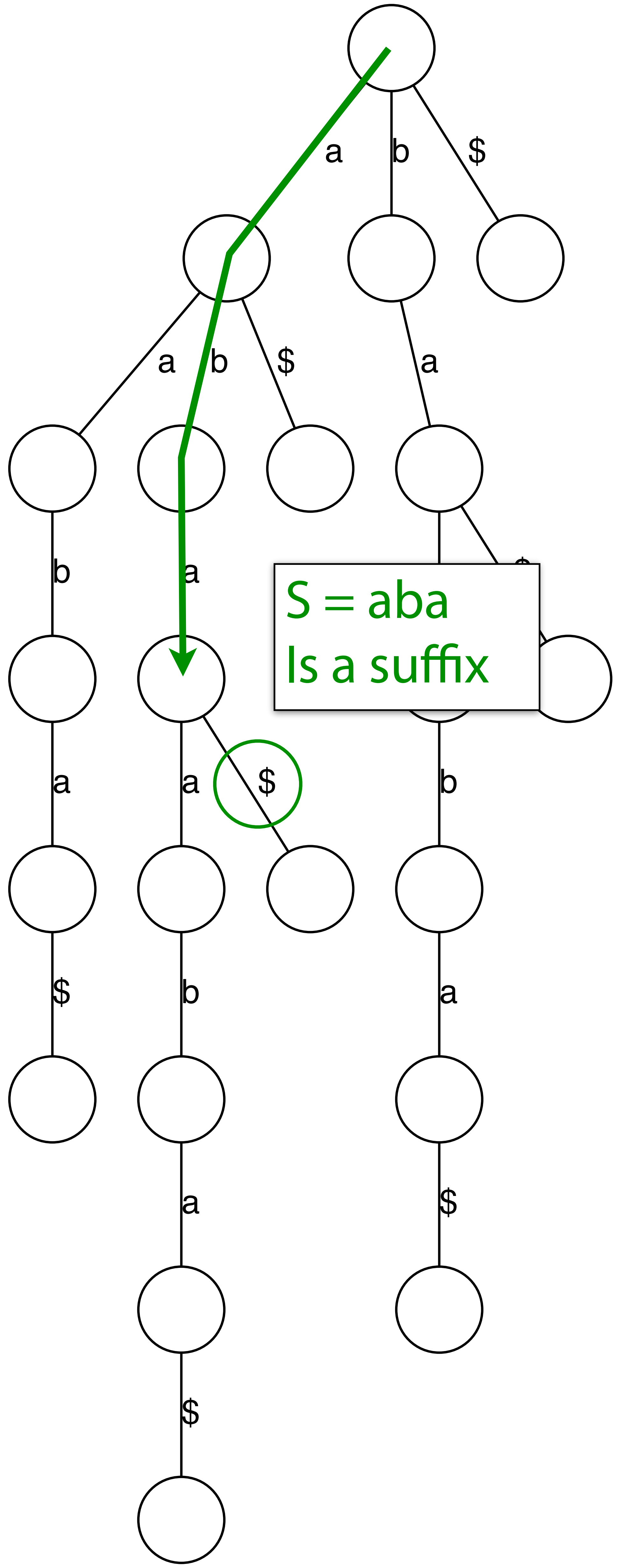
Same procedure as for substring, but additionally check terminal node for $ child



Suffix trie

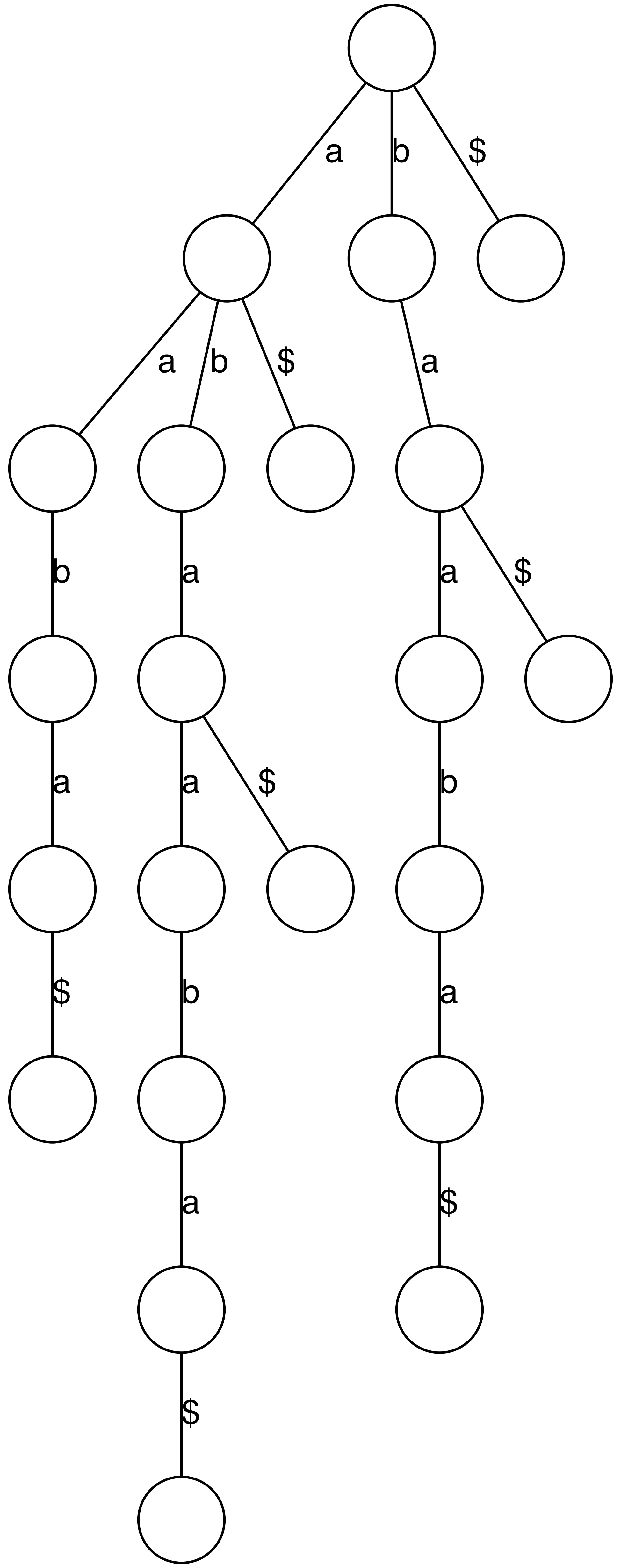
How do we check whether a string S is a suffix of T?

Same procedure as for substring, but additionally check terminal node for $ child



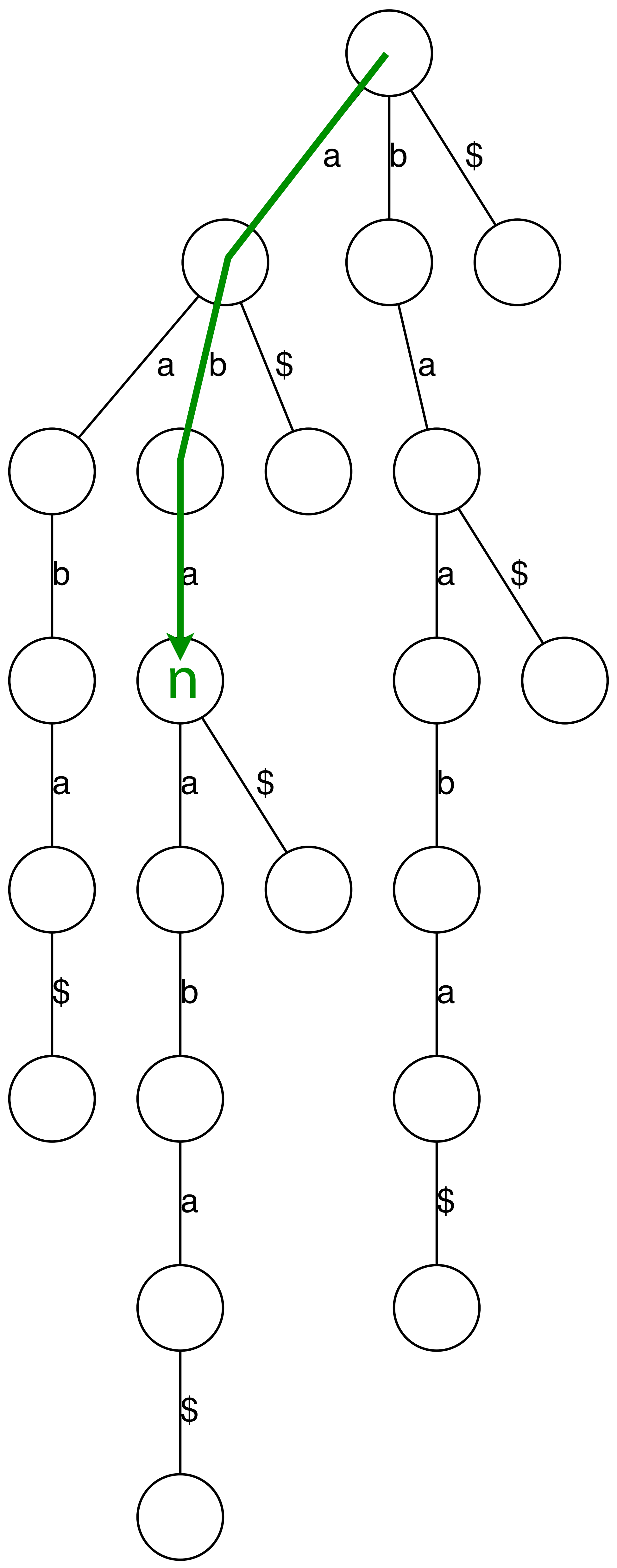
Suffix trie

How do we count the number of times a string S occurs as a substring of T?



Suffix trie

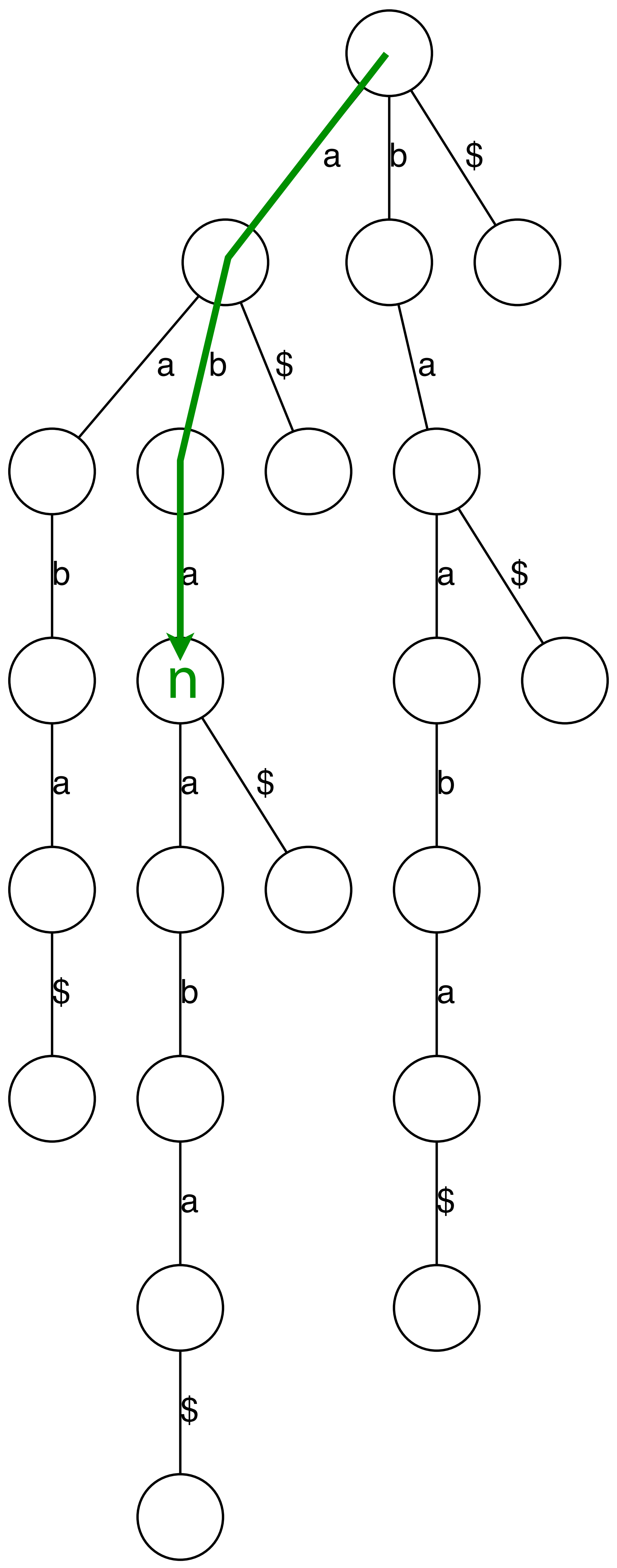
How do we count the number of times a string S occurs as a substring of T?



Suffix trie

How do we count the number of times a string S occurs as a substring of T?

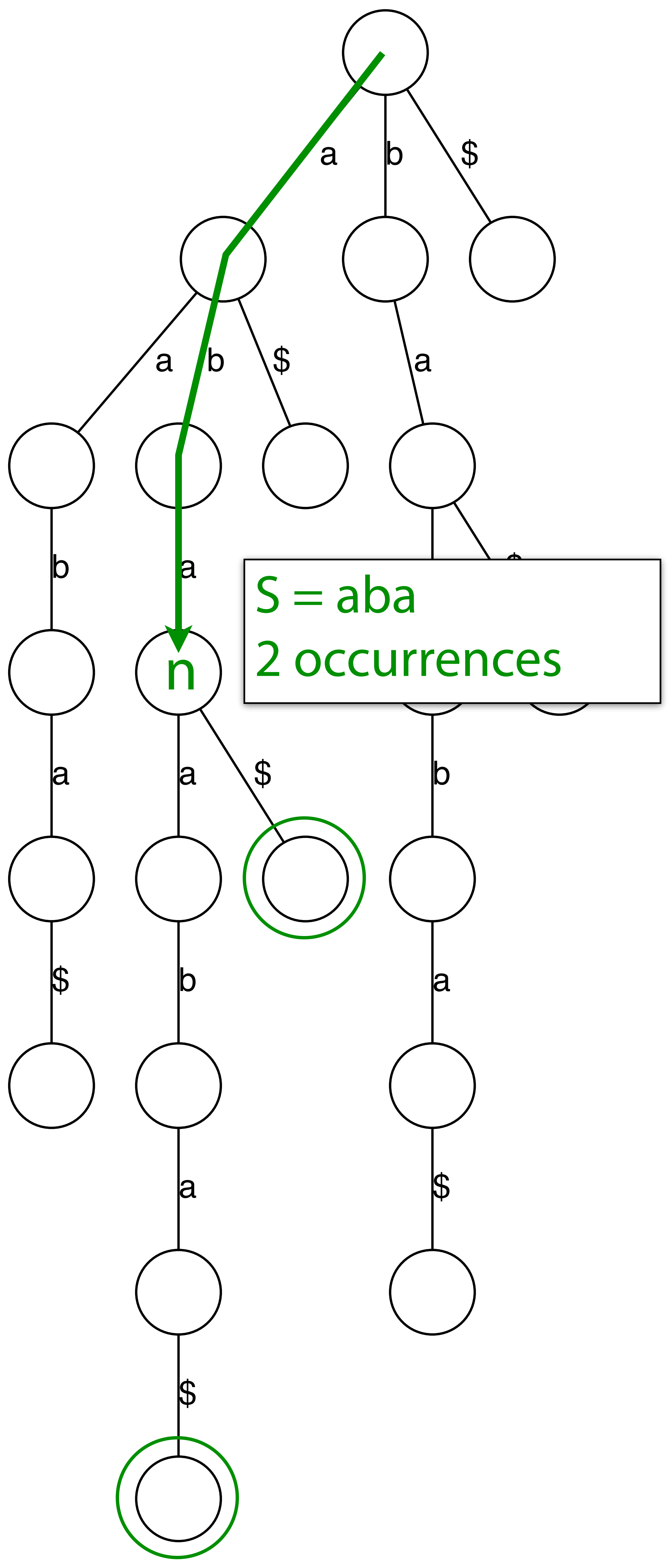
Follow path labeled with S. If we fall off, answer is 0. If we end up at node n, answer equals # of leaves in subtree rooted at n.



Suffix trie

How do we count the number of times a string S occurs as a substring of T?

Follow path labeled with S. If we fall off, answer is 0. If we end up at node n, answer equals # of leaves in subtree rooted at n.

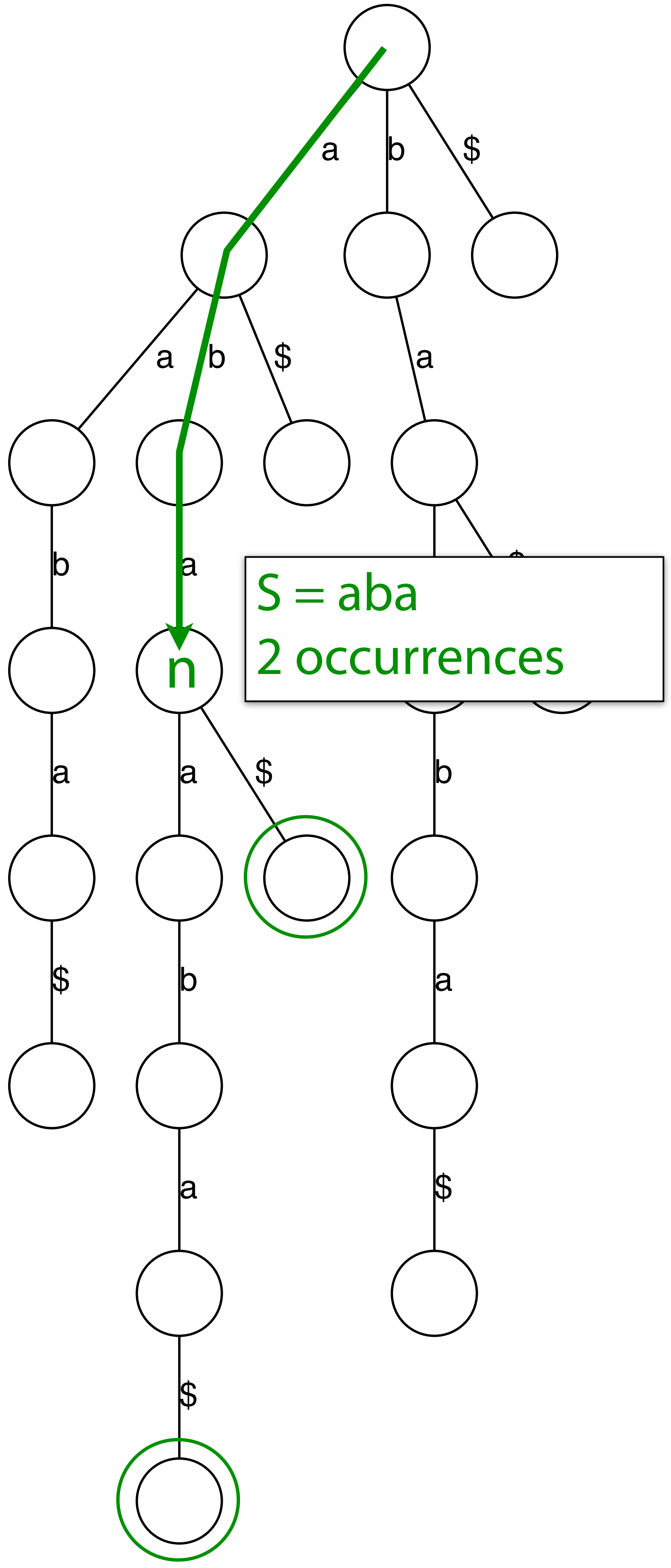


Suffix trie

How do we count the number of times a string S occurs as a substring of T?

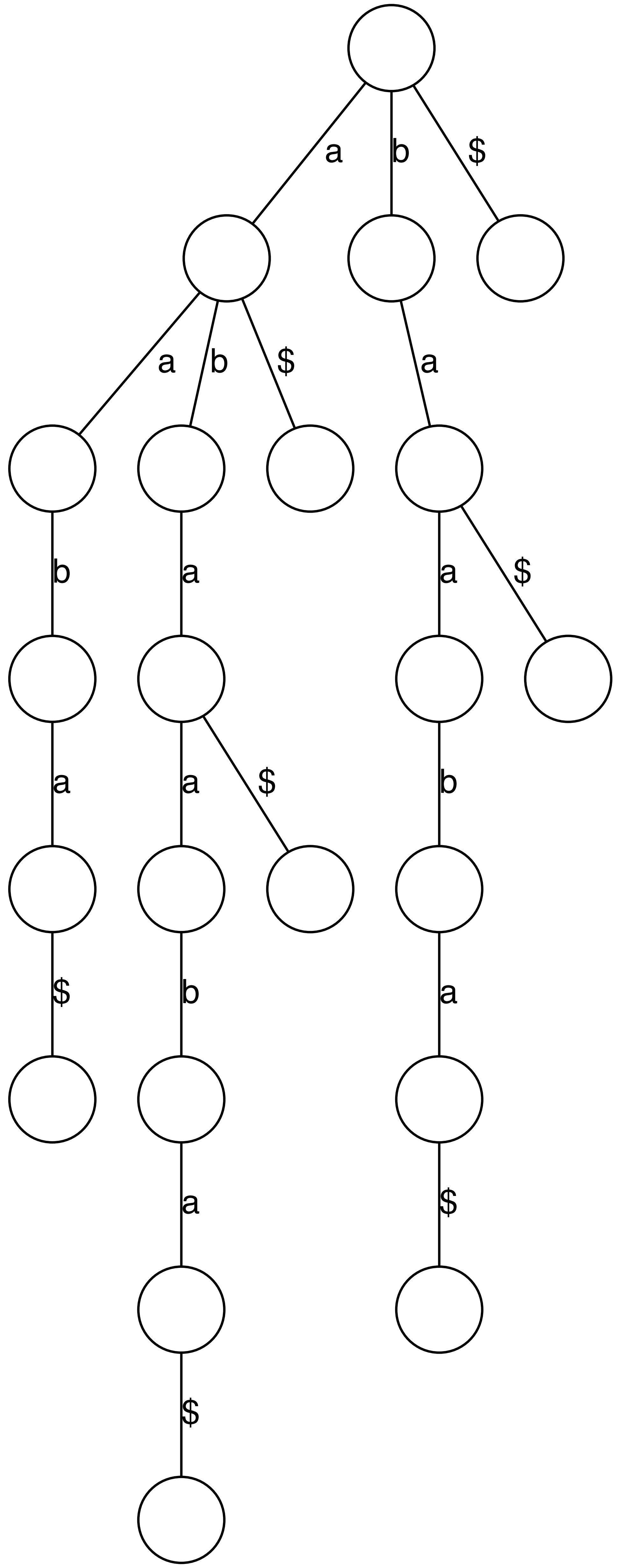
Follow path labeled with S. If we fall off, answer is 0. If we end up at node n, answer equals # of leaves in subtree rooted at n.

Leaves can be counted with depth-first traversal.



Suffix trie

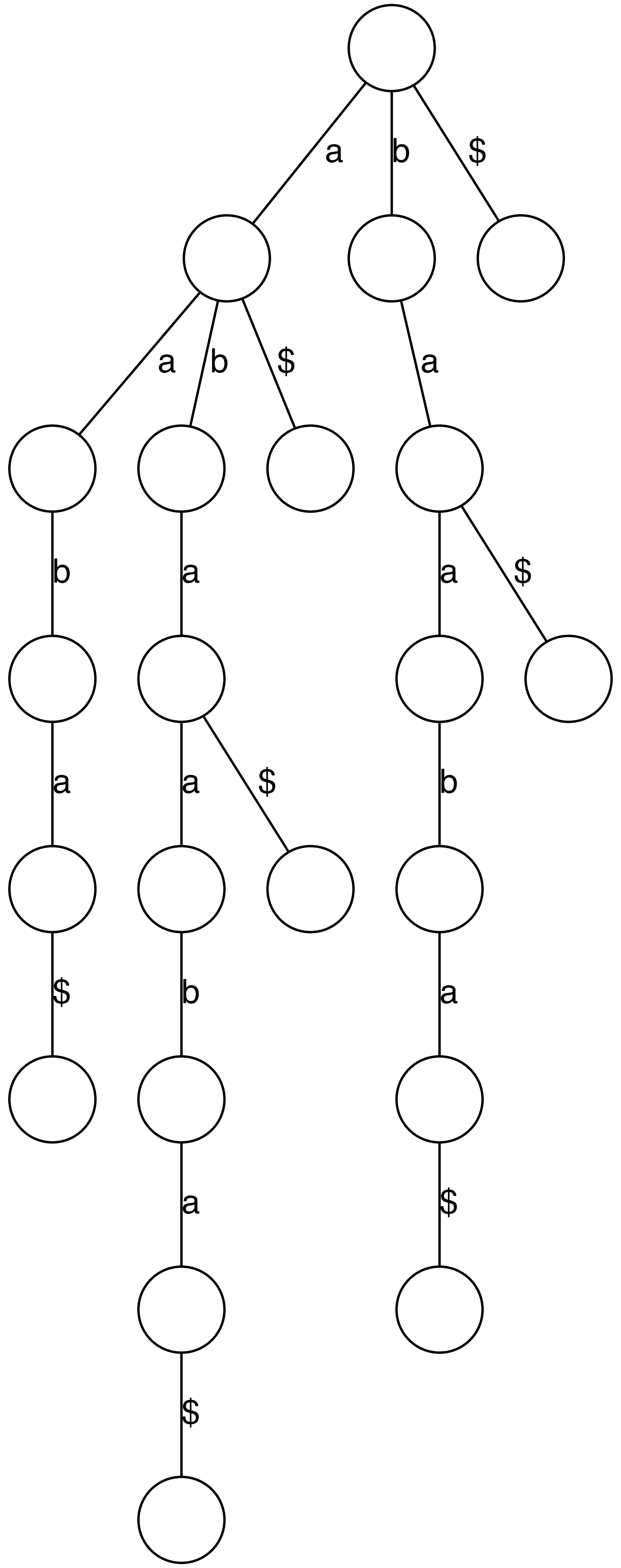
How do we find the longest repeated substring of T?



Suffix trie

How do we find the longest repeated substring of T?

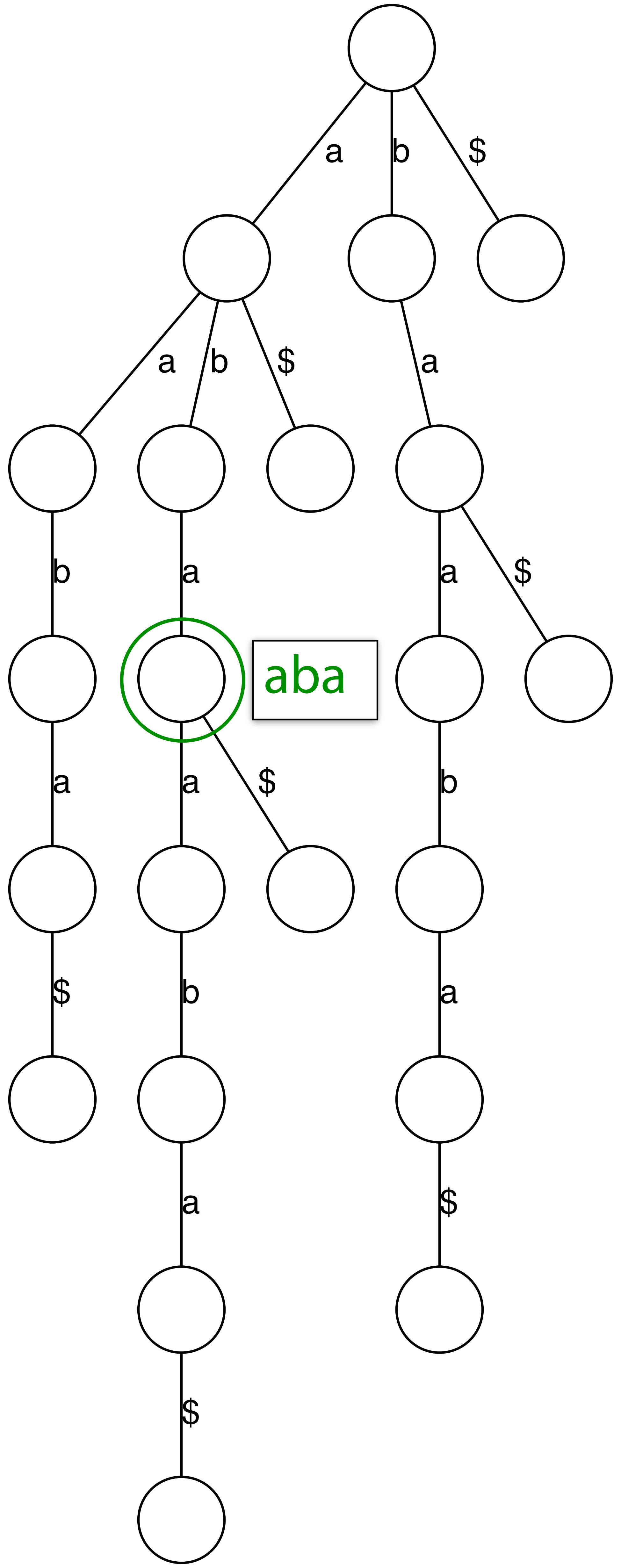
Find the deepest node with more than one child



Suffix trie

How do we find the longest repeated substring of T?

Find the deepest node with more than one child



Suffix trie

How many nodes does the suffix trie have?

Is there a class of string where the number of suffix trie nodes grows linearly with m?

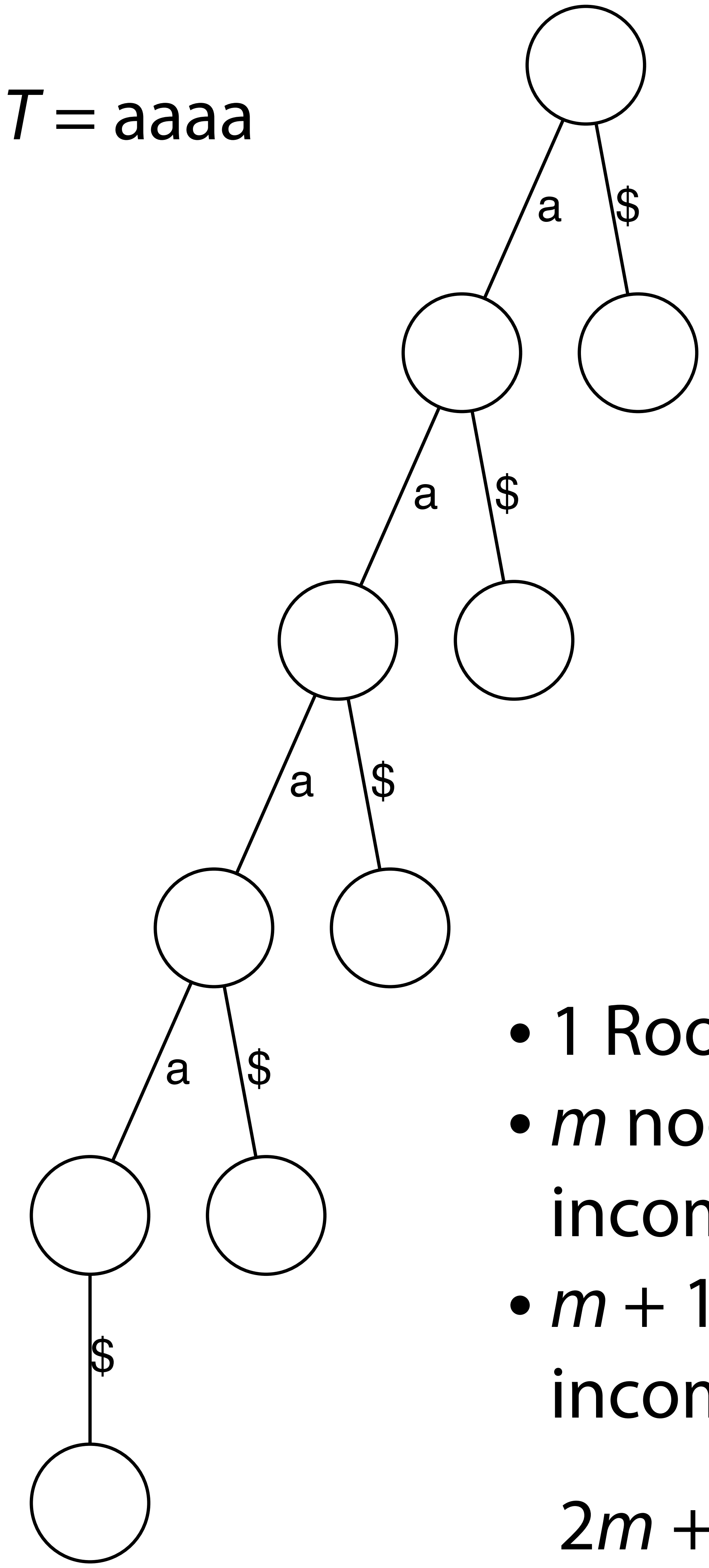
Suffix trie

How many nodes does the suffix trie have?

Is there a class of string where the number of suffix trie nodes grows linearly with *m*?

Yes: e.g. a string of *m* a's in a row (a^m)

T = aaaa



- 1 Root
- *m* nodes with incoming **a** edge
- *m* + 1 nodes with incoming **$** edge

2*m* + 2 nodes

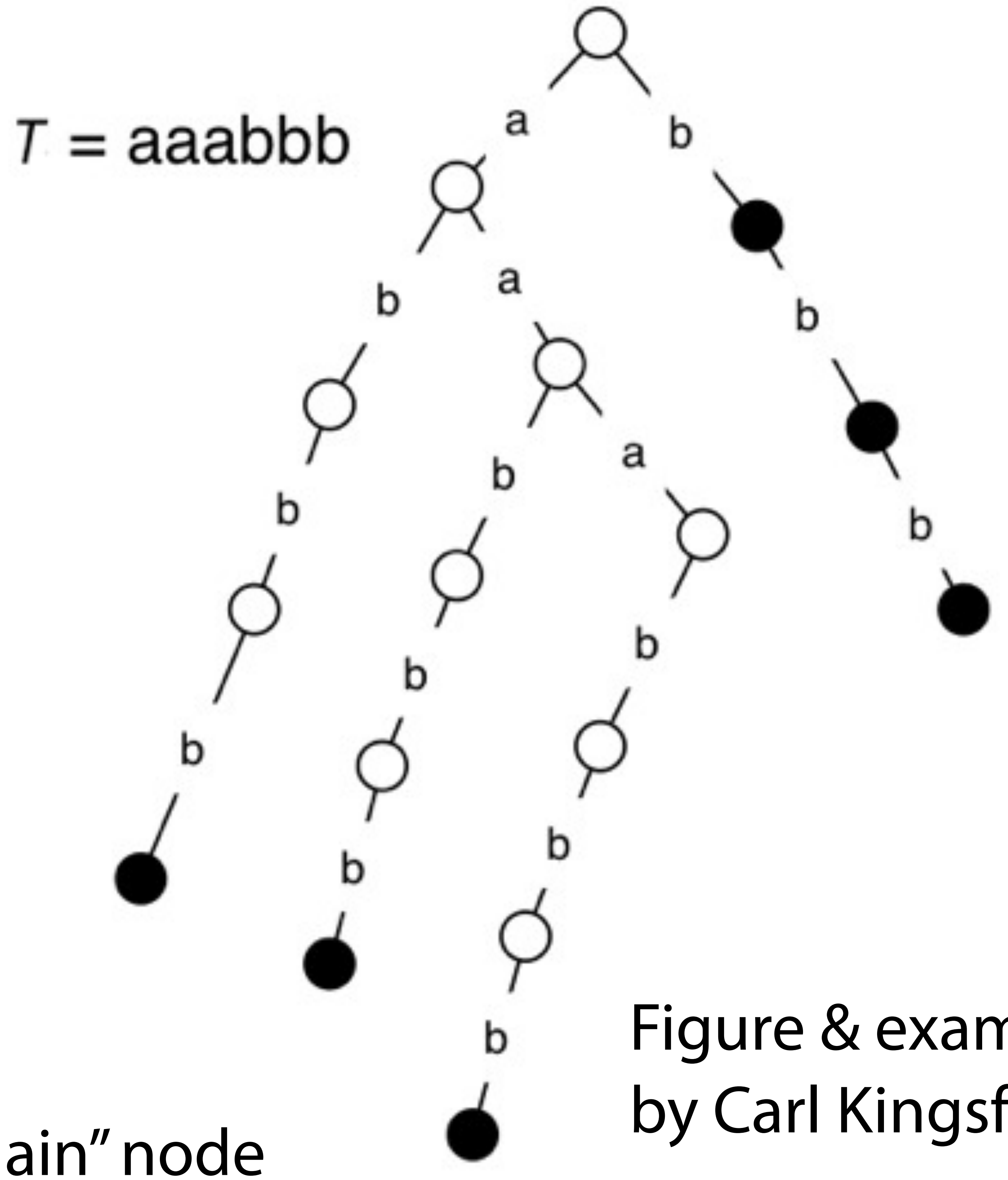
Suffix trie

Is there a class of string where the number of suffix trie nodes grows with m^2?

Suffix trie

Is there a class of string where the number of suffix trie nodes grows with m^2?

Yes: $a^n b^n$



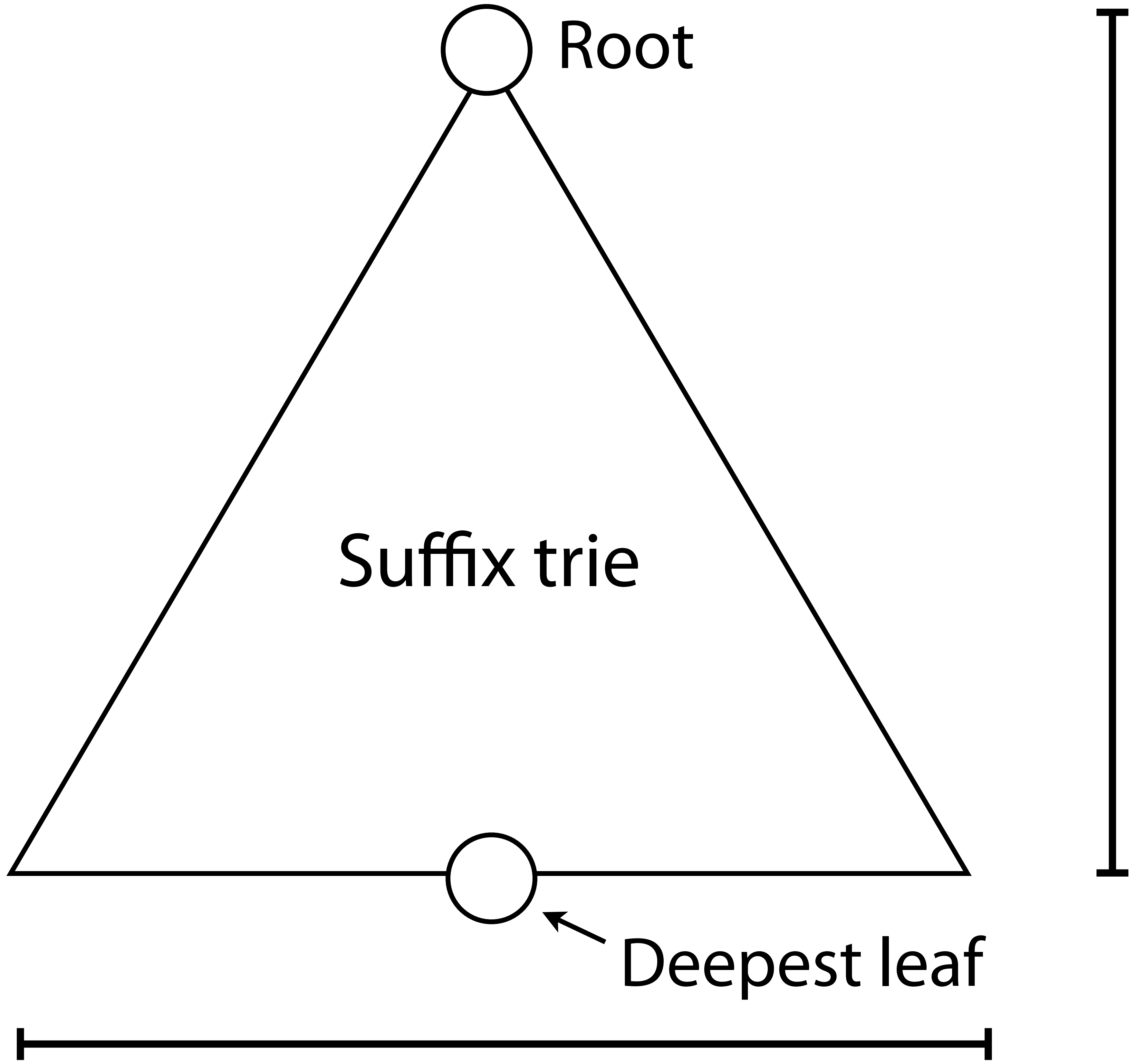
- 1 root
- n nodes along "b chain," right
- n nodes along "a chain," middle
- n chains of n "b" nodes hanging off each "a chain" node
- $2n + 1$ **\$** leaves (not shown)

$n^2 + 4n + 2$ nodes, where $m = 2n$

Figure & example by Carl Kingsford

Suffix trie: upper bound on size

Could worst-case # nodes be worse than O(m^2)?



Max # nodes from *top to bottom*
= length of longest suffix + 1
= m + 1

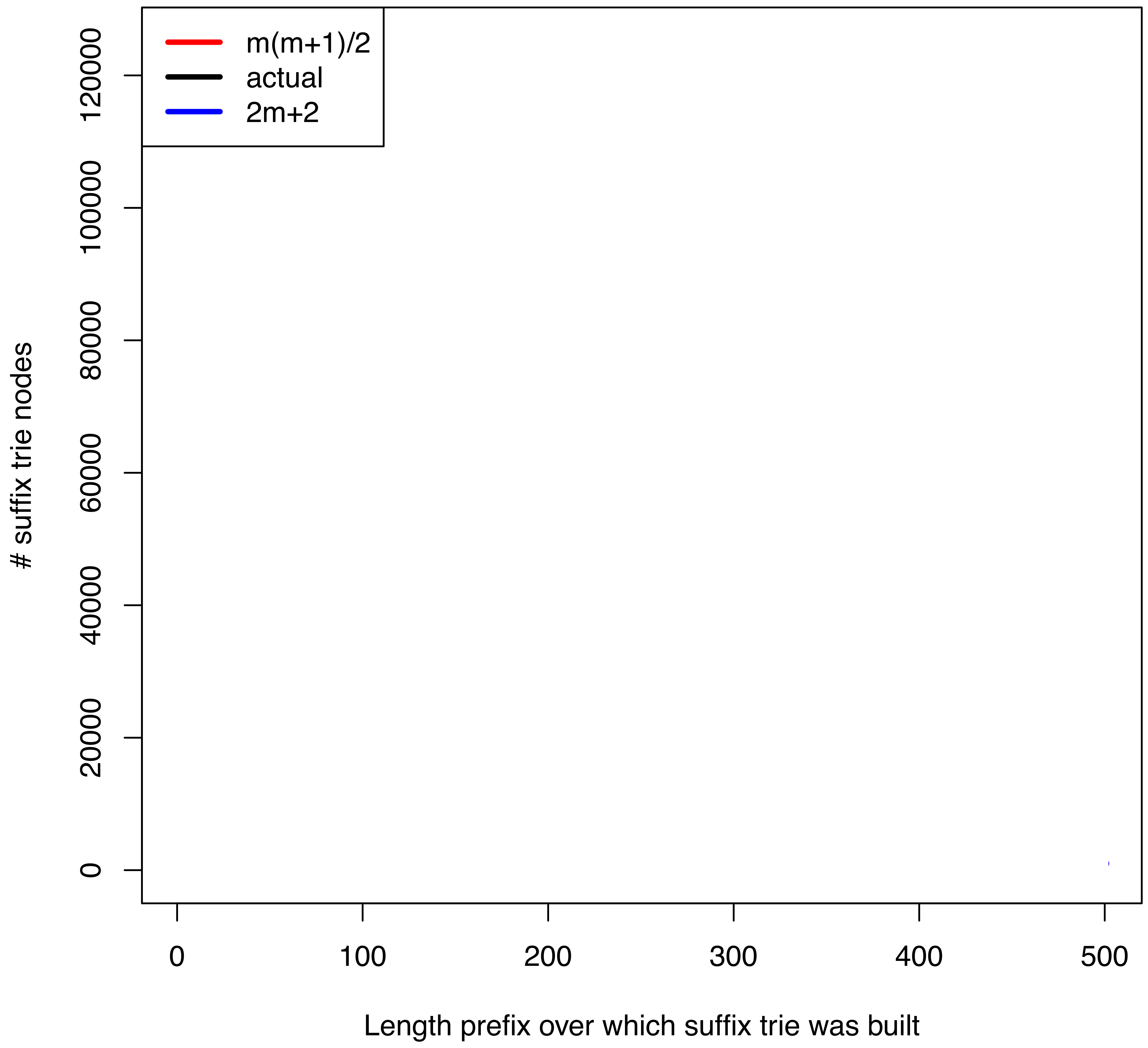
Max # nodes from *left to right*
= max # distinct substrings of any length
≤ m

$O(m^2)$ is worst case

Suffix trie: actual growth

Built suffix tries for the first 500 prefixes of the lambda phage virus genome

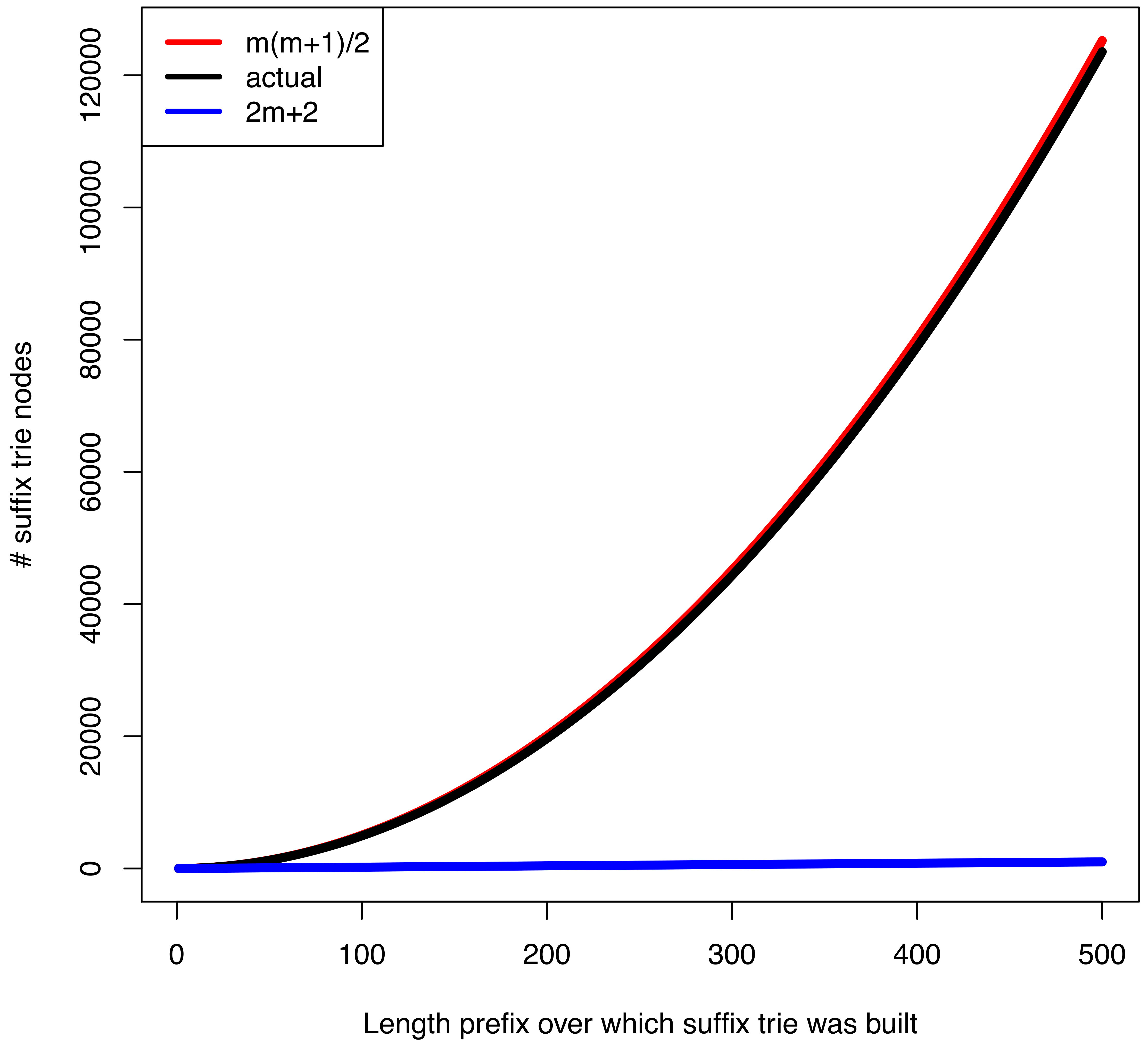
Black curve shows how # nodes increases with prefix length



Suffix trie: actual growth

Built suffix tries for the first 500 prefixes of the lambda phage virus genome

Black curve shows how # nodes increases with prefix length

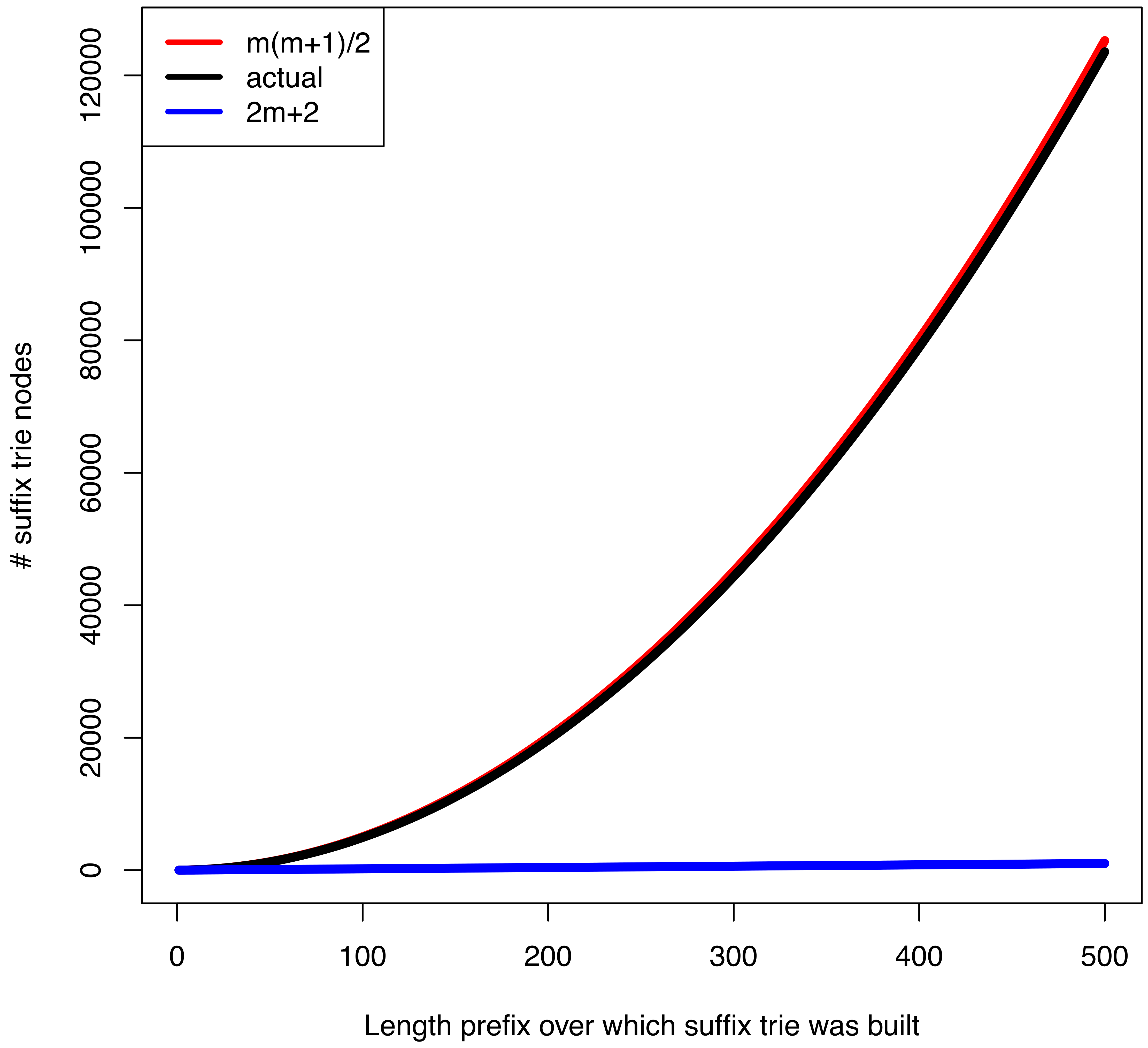


Suffix trie: actual growth

Built suffix tries for the first 500 prefixes of the lambda phage virus genome

Black curve shows how # nodes increases with prefix length

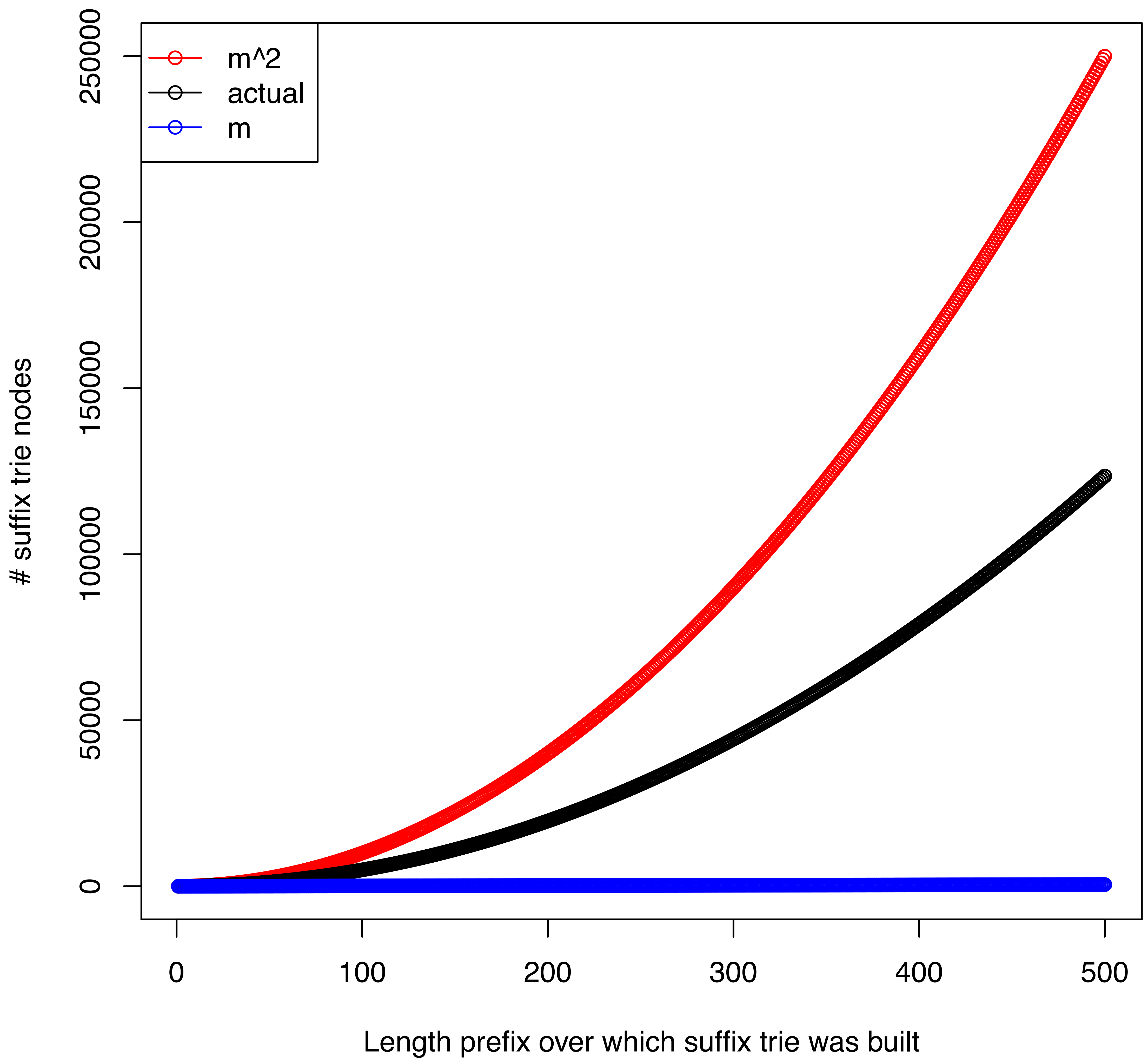
Actual growth much closer to worst case than to best!



Suffix trie: actual growth

Built suffix tries for the first 500 prefixes of the lambda phage virus genome

Black curve shows how # nodes increases with prefix length



Suffix tries ⇒ Suffix trees

Suffix Tree Definitions

A **Σ+**-tree is a rooted tree, T, where each edge is labeled with *non-empty strings*, where no node has two outgoing edges labeled with strings having the same *first* character. T is **compact** if all internal nodes have ≥ 2 children.

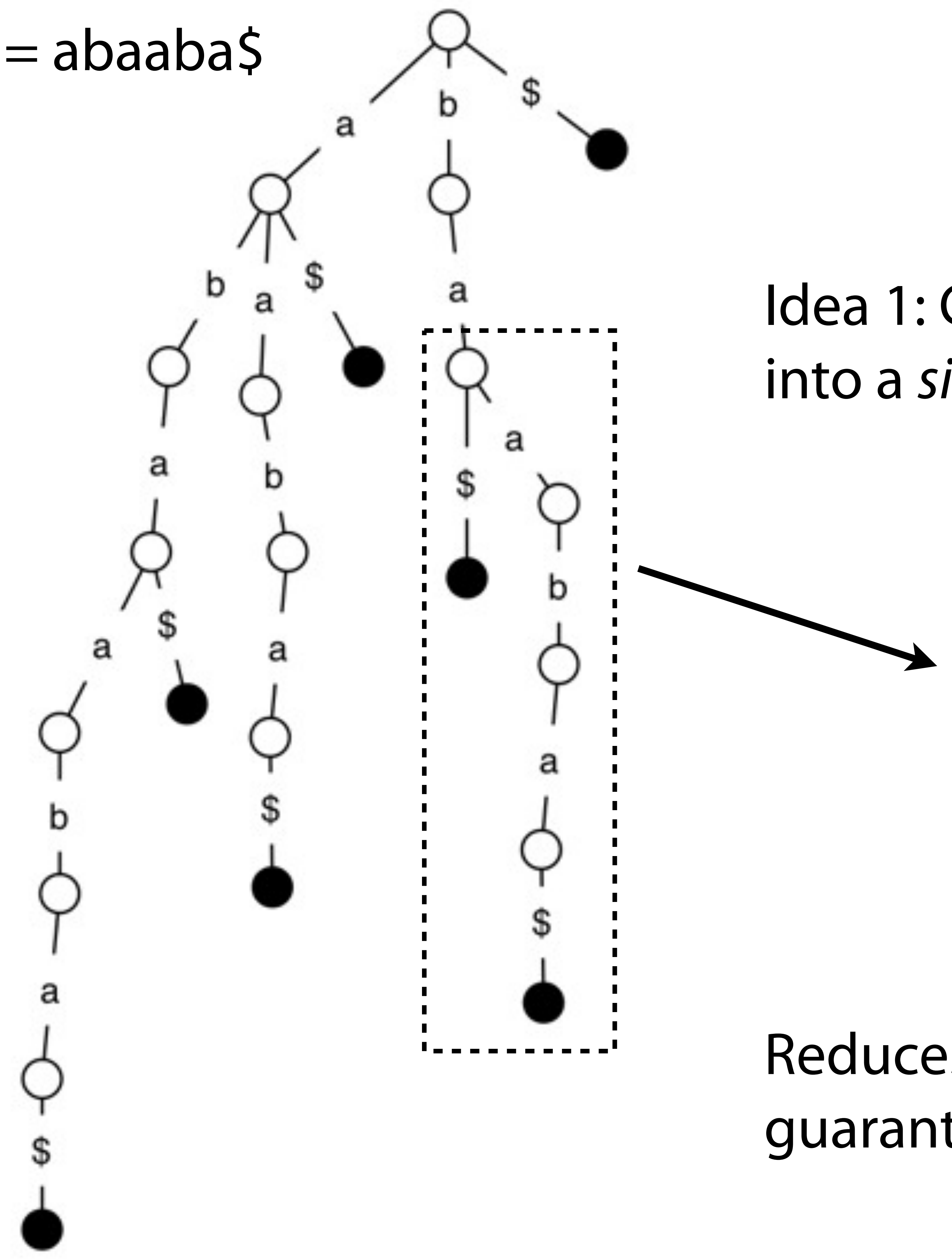
- for a node v in T, **depth**(v) or **node-depth**(v) is the distance from v to the root.
- **node-depth**(r) = 0
- **string**(v) = concatenation of all characters on the path r ⇝ v
- **string-depth**(v) = |**string**(v)| (note: **string-depth**(v) ≥ **node-depth**(v))
- for a string x, if ∃ node v with **string**(v) = x, we say **node**(x) = v
- T **displays** string x if ∃ node v and string y such that xy = **string**(v)
- **words**(T) = { x | T **displays** x}
- A **suffix tree** of string s is a compact Σ+-tree such that **words**(T) = {s' | s' is a substring of s}

Defs. from: http://profs.sci.univr.it/~liptak/ALBioinfo/files/sequence_analysis.pdf

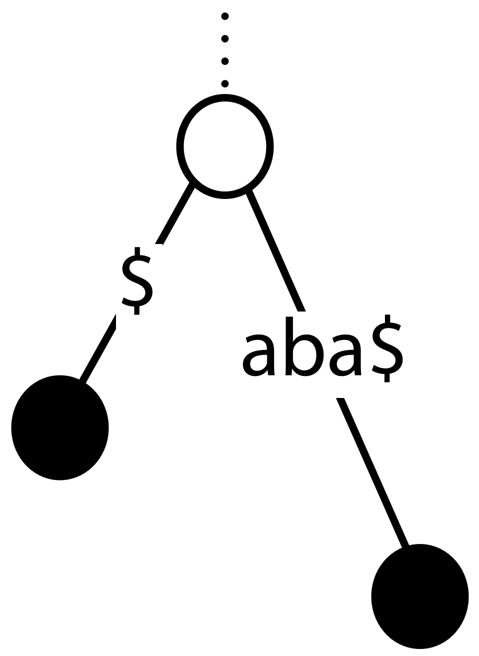
*

Suffix trie: making it smaller

T = abaaba\$



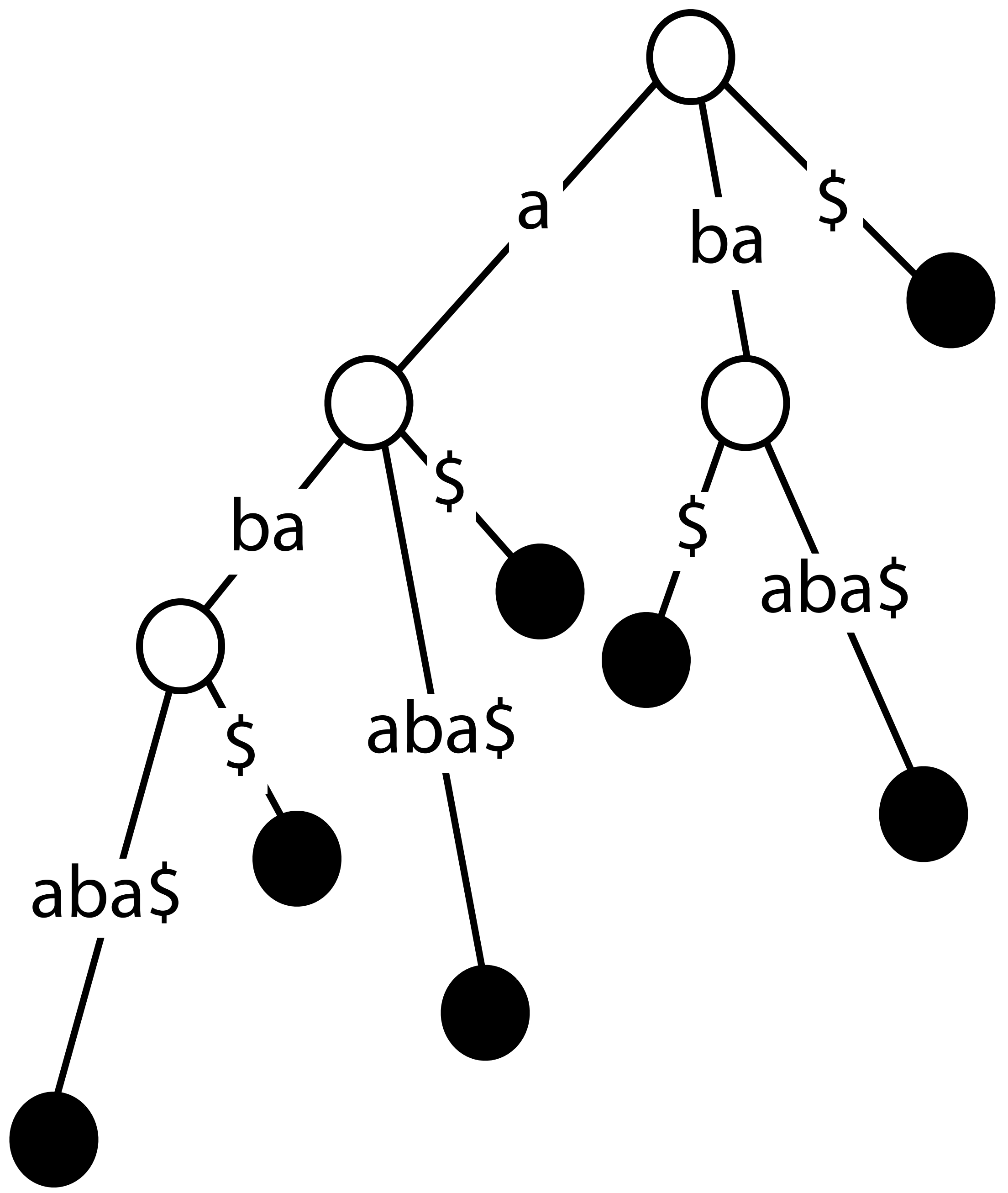
Idea 1: Coalesce non-branching paths into a *single edge* with a *string* label



Reduces # nodes, edges,
guarantees internal nodes have >1 child

Suffix tree

T = abaaba$



L leaves, I internal nodes, E edges

$E = L + I - 1$

$E \geq 2I$ (each internal node branches)

$L + I - 1 \geq 2I \Rightarrow I \leq L - 1$

but

$L \leq m$ (at most m suffixes)

$I \leq m - 1$

$E = L + I - 1 \leq 2m - 2$

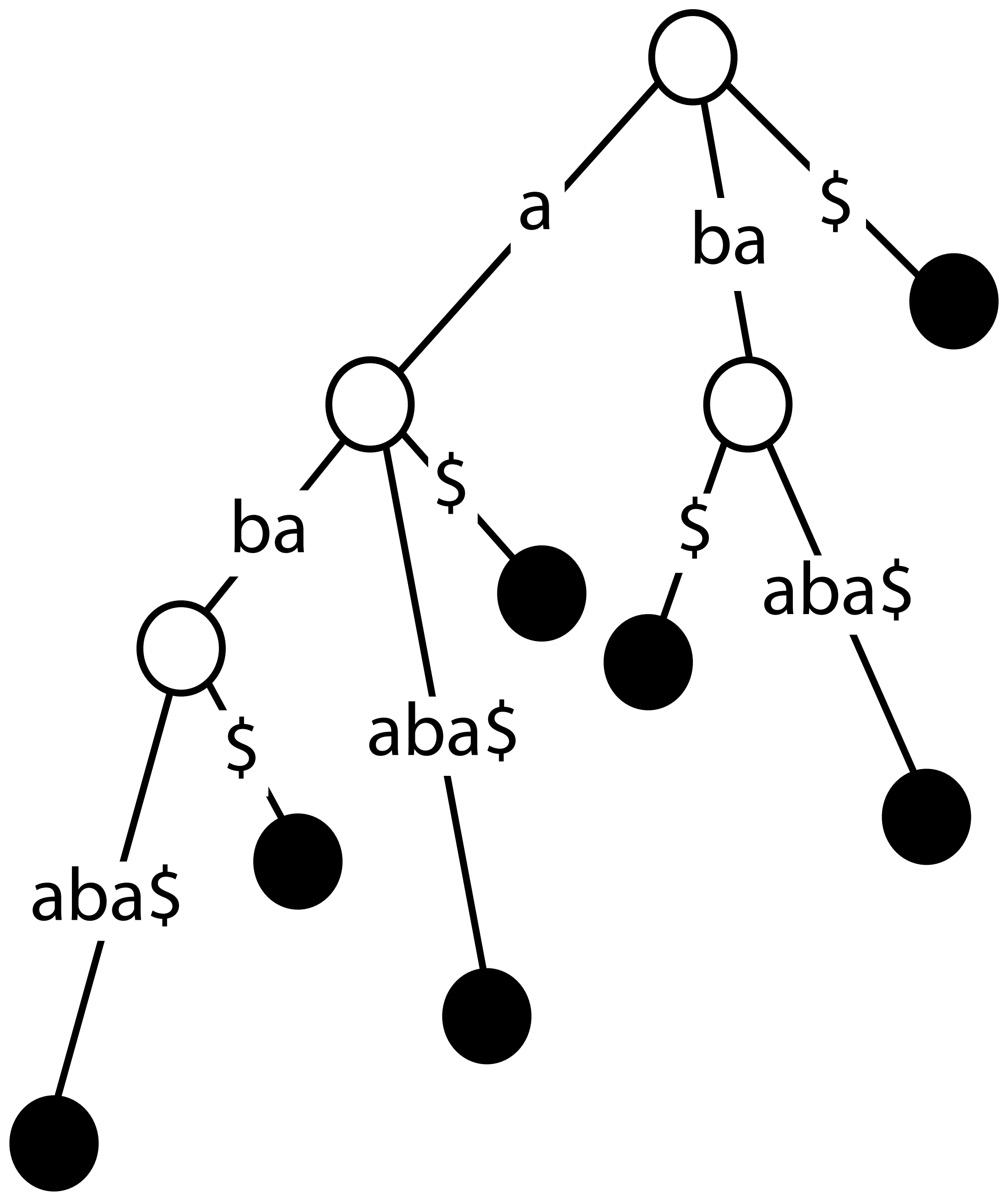
$E + L + I \leq 4m - 3 \in O(m)$

Is the total size O(m) now?

Suffix tree

L leaves, I internal nodes, E edges

T = abaaba$



$E = L + I - 1$

$E \geq 2I$ (each internal node branches)

$L + I - 1 \geq 2I \Rightarrow I \leq L - 1$

but

$L \leq m$ (at most m suffixes)

$I \leq m - 1$

$E = L + I - 1 \leq 2m - 2$

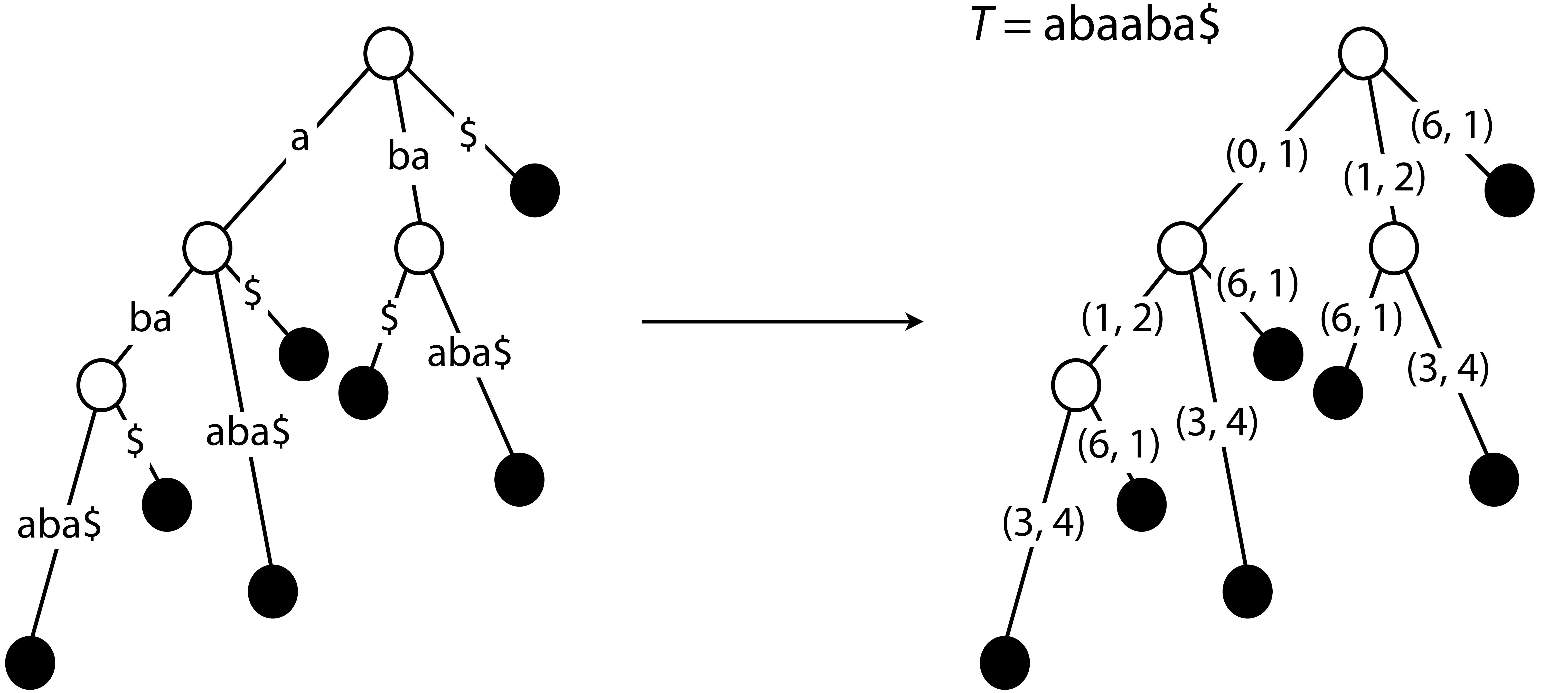
$E + L + I \leq 4m - 3 \in O(m)$

Is the total size O(m) now?

NO: The total length of edge labels is quadratic in m.

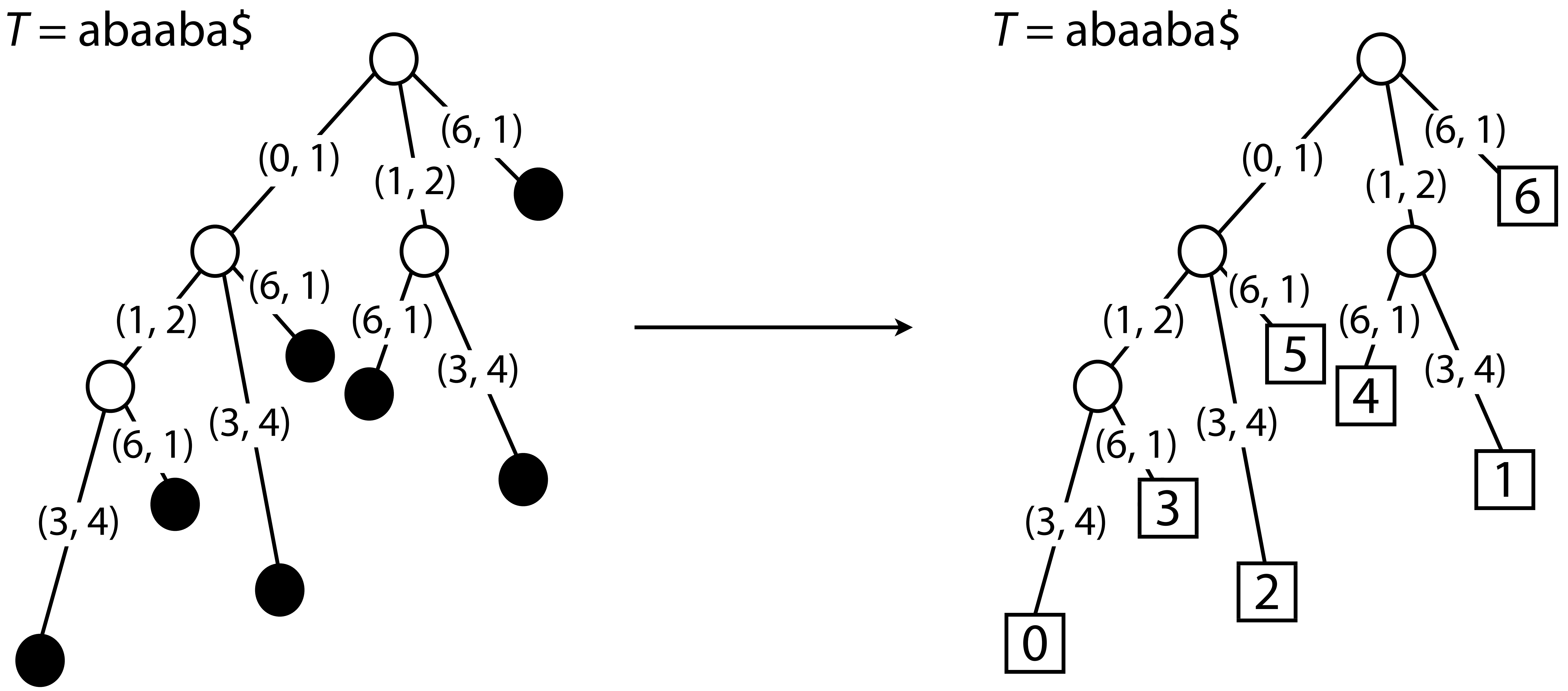
Suffix tree

T = abaaba$ Idea 2: Store T itself in addition to the tree. Convert tree's edge labels to (offset, length) pairs with respect to T.

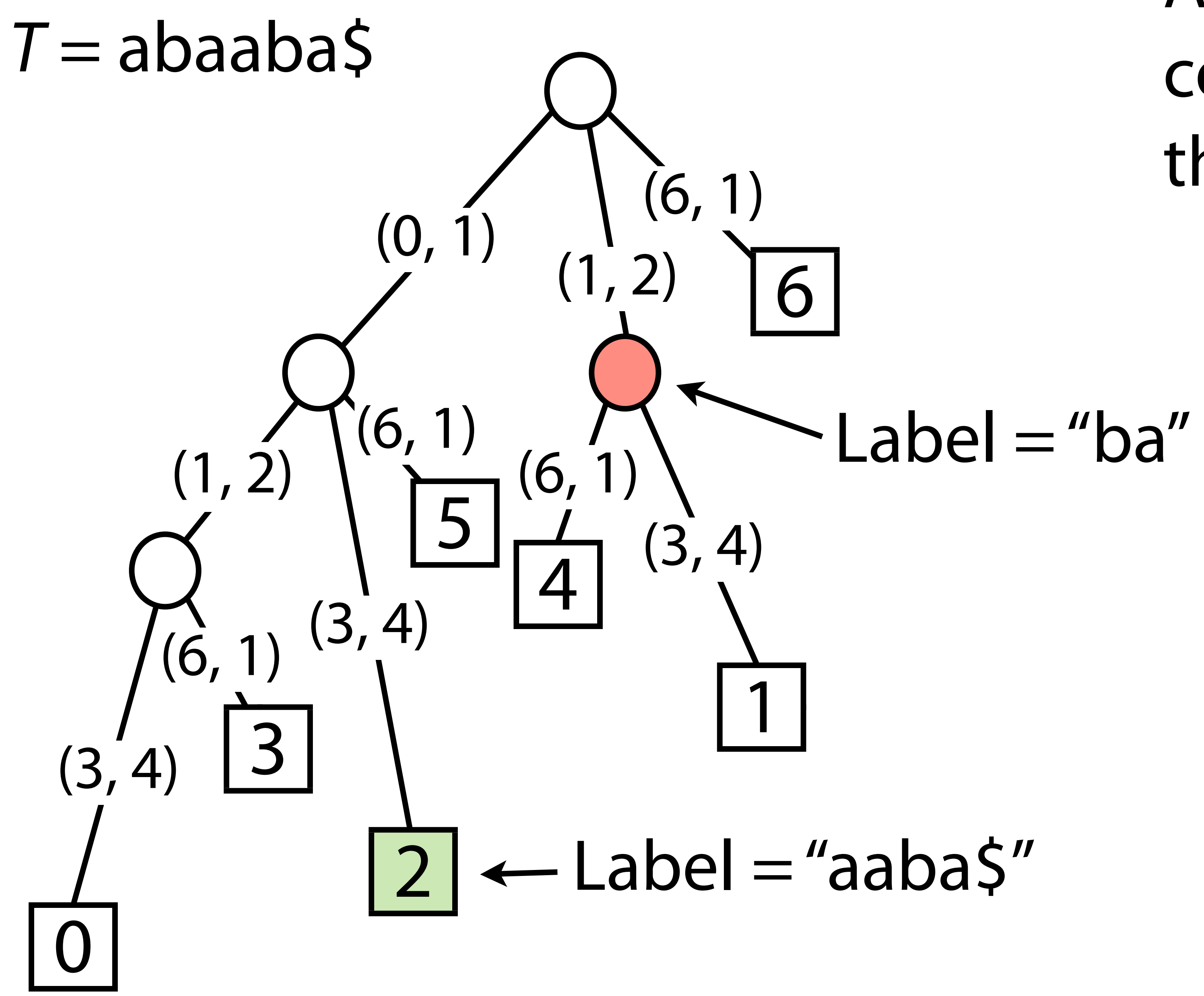


Space required for suffix tree is now $O(m)$

Suffix tree: leaves hold offsets where suffixes **begin**



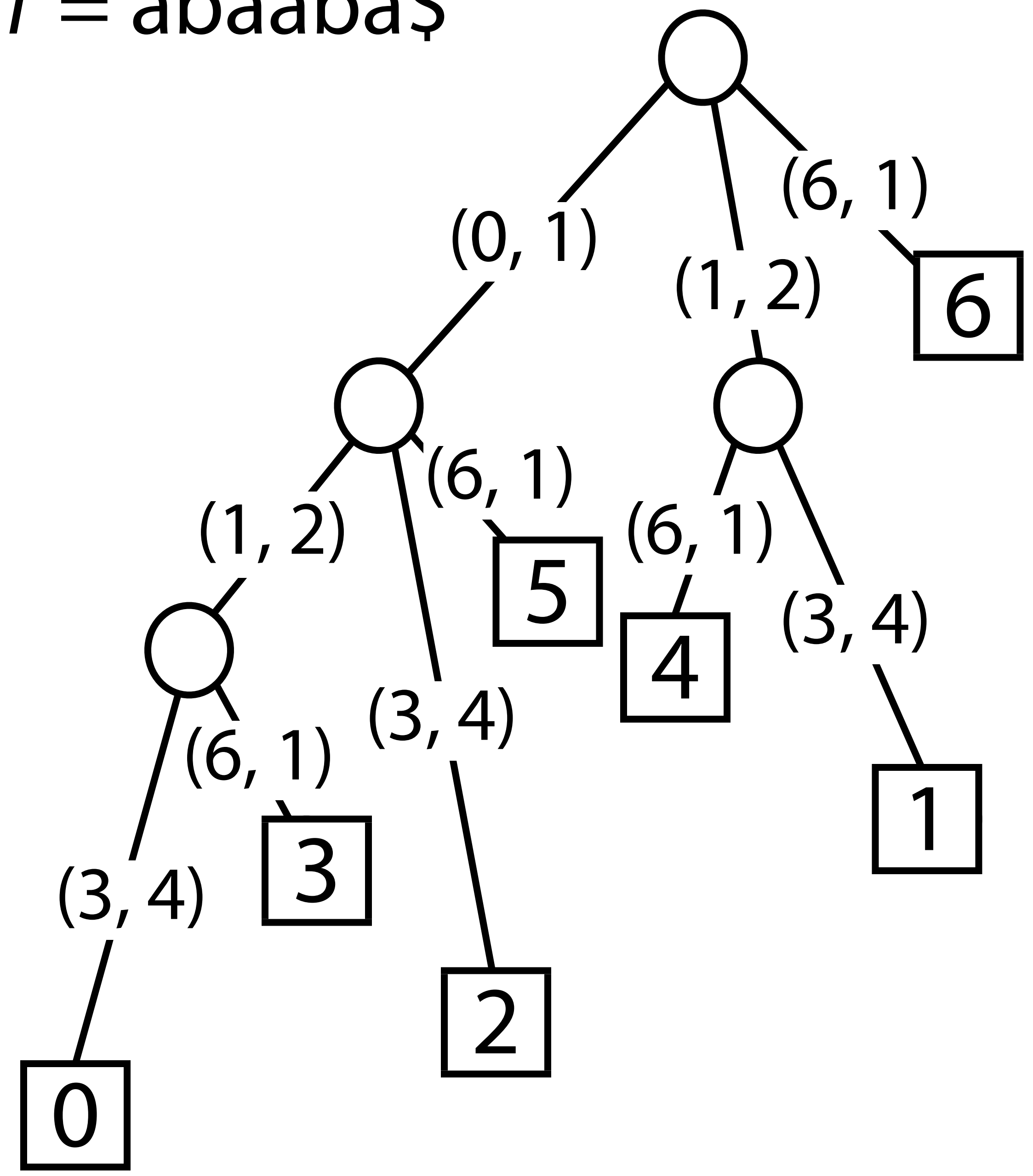
Suffix tree: labels



Again, each node's *label* equals the concatenated edge labels from the root to the node. These aren't stored explicitly.

Suffix tree: labels

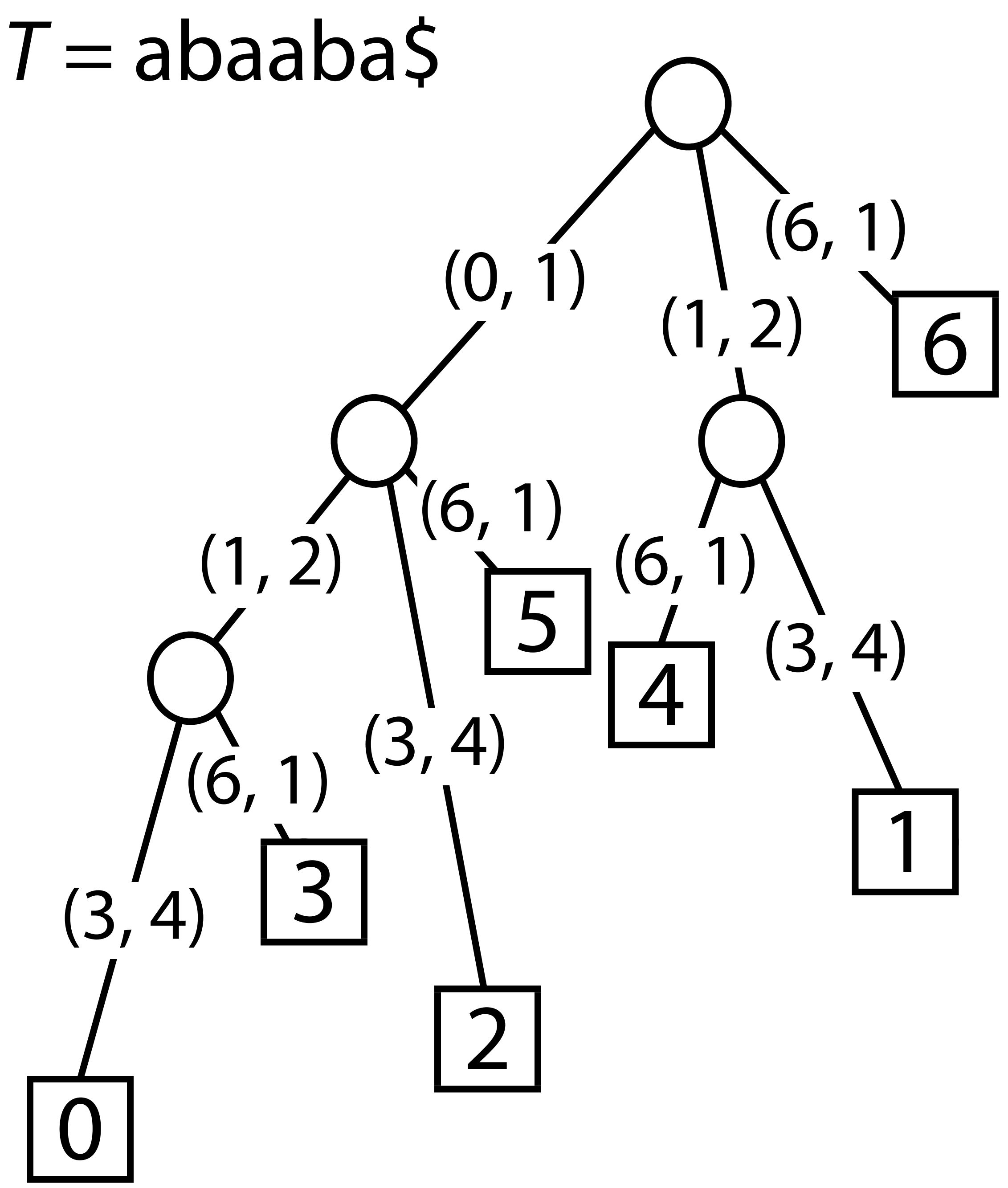
T = abaaba\$



Because edges can have string labels, we must distinguish two notions of "depth"

- **Node** depth: how many edges we must follow from the root to reach the node
- **Label** depth: total length of edge labels for edges on path from root to node

Suffix tree: space caveat



Minor point:

We say the space taken by the edge labels is O(m), because we keep 2 integers per edge and there are O(m) edges

To store one such integer, we need enough bits to distinguish m positions in T, i.e. ceil($\log_2 m$) bits. We usually ignore this factor, since 64 bits is plenty for all practical purposes.

Similar argument for the pointers / references used to distinguish tree nodes.

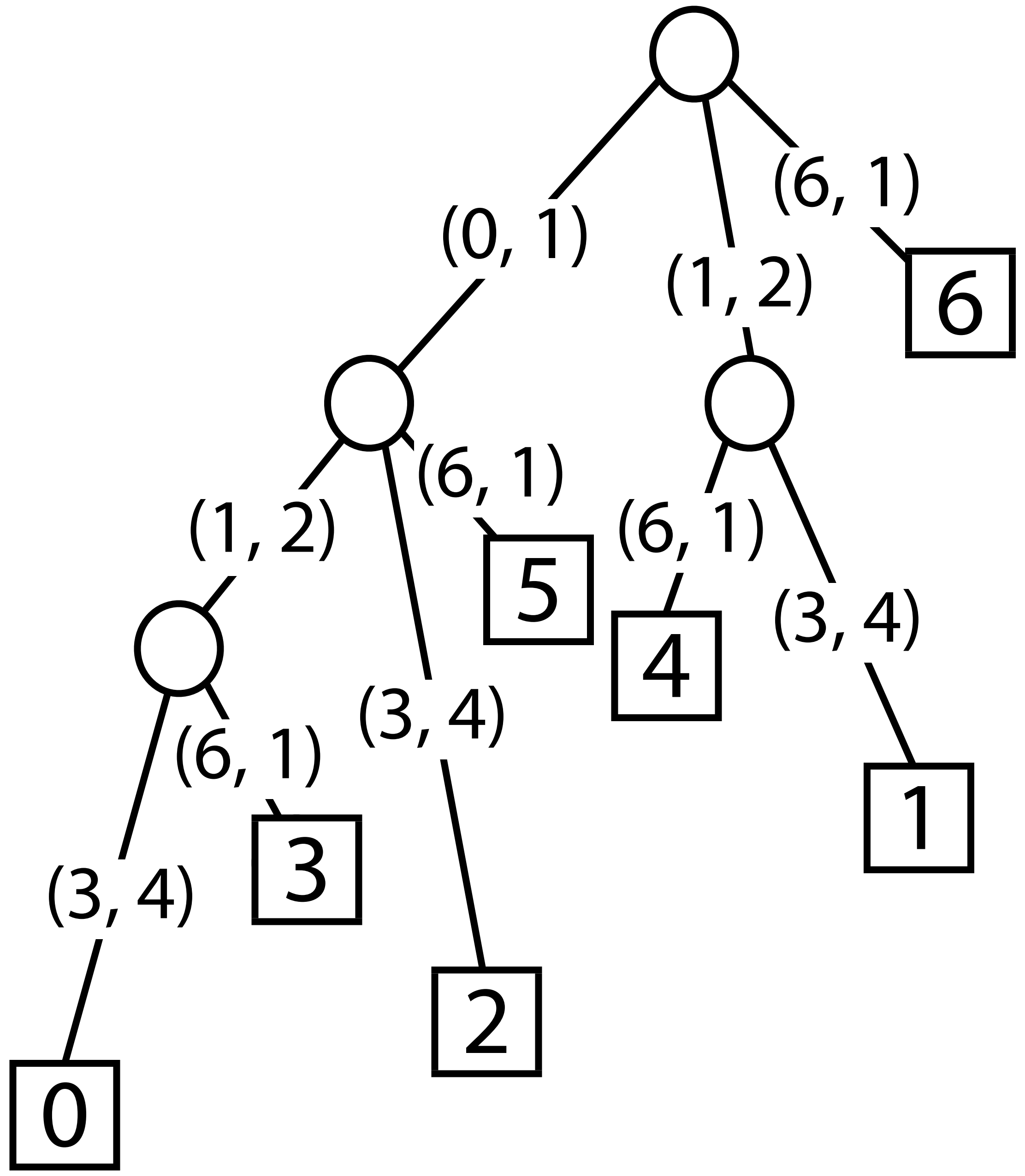
Suffix tree: building

Naive method 1: build a suffix trie, then coalesce non-branching paths and relabel edges

Naive method 2: build a single-edge tree representing only the longest suffix, then augment to include the 2nd-longest, then augment to include 3rd-longest, etc

Both are $O(m^2)$ time, but first uses $O(m^2)$ space while second uses $O(m)$

Naive method 2 is described in Gusfield 5.4



Suffix tree: building

Other methods for construction:

Ukkonen, Esko. "On-line construction of suffix trees."
Algorithmica 14.3 (1995): 249-260.

$O(m)$ time and space

Has *online* property: if T arrives one character at a time, algorithm efficiently updates suffix tree upon each arrival

We won't cover it here; see Gusfield Ch. 6 for details

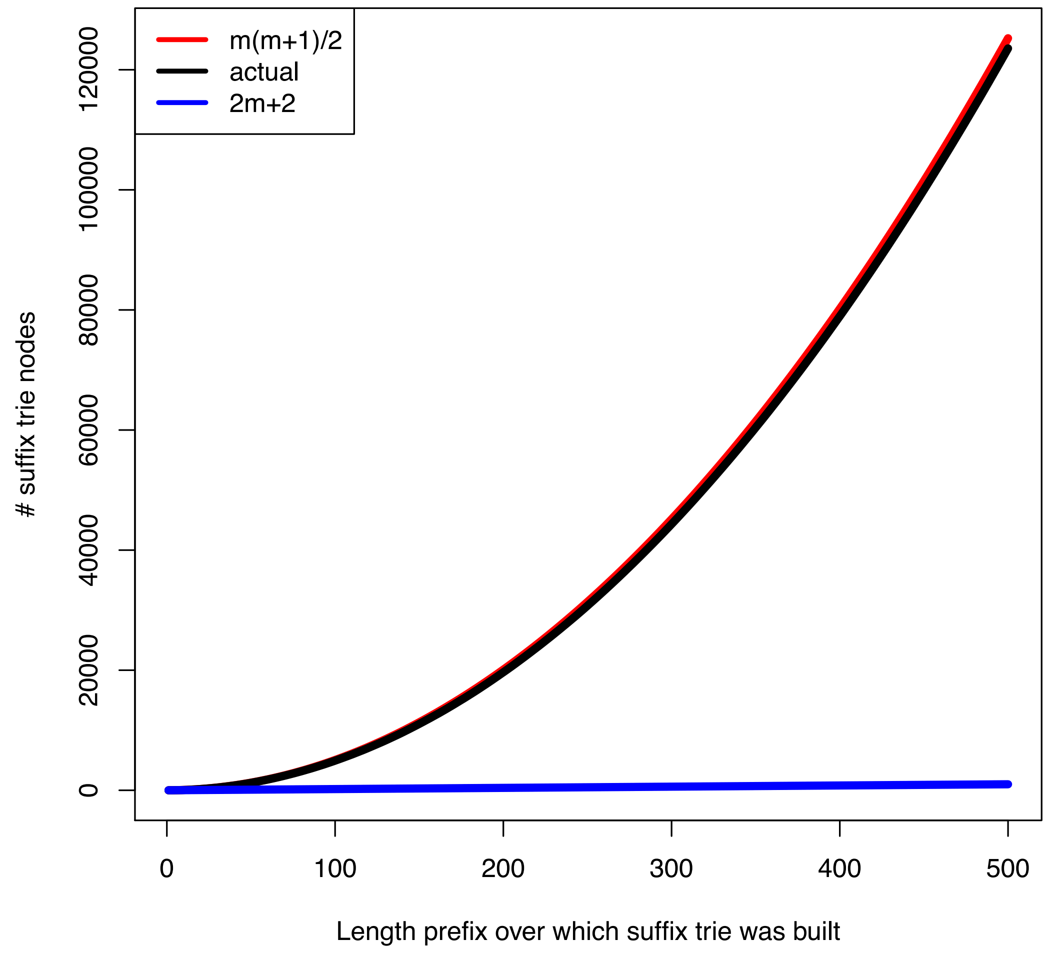
Or just Google "Ukkonen's algorithm"

Suffix tree: actual growth

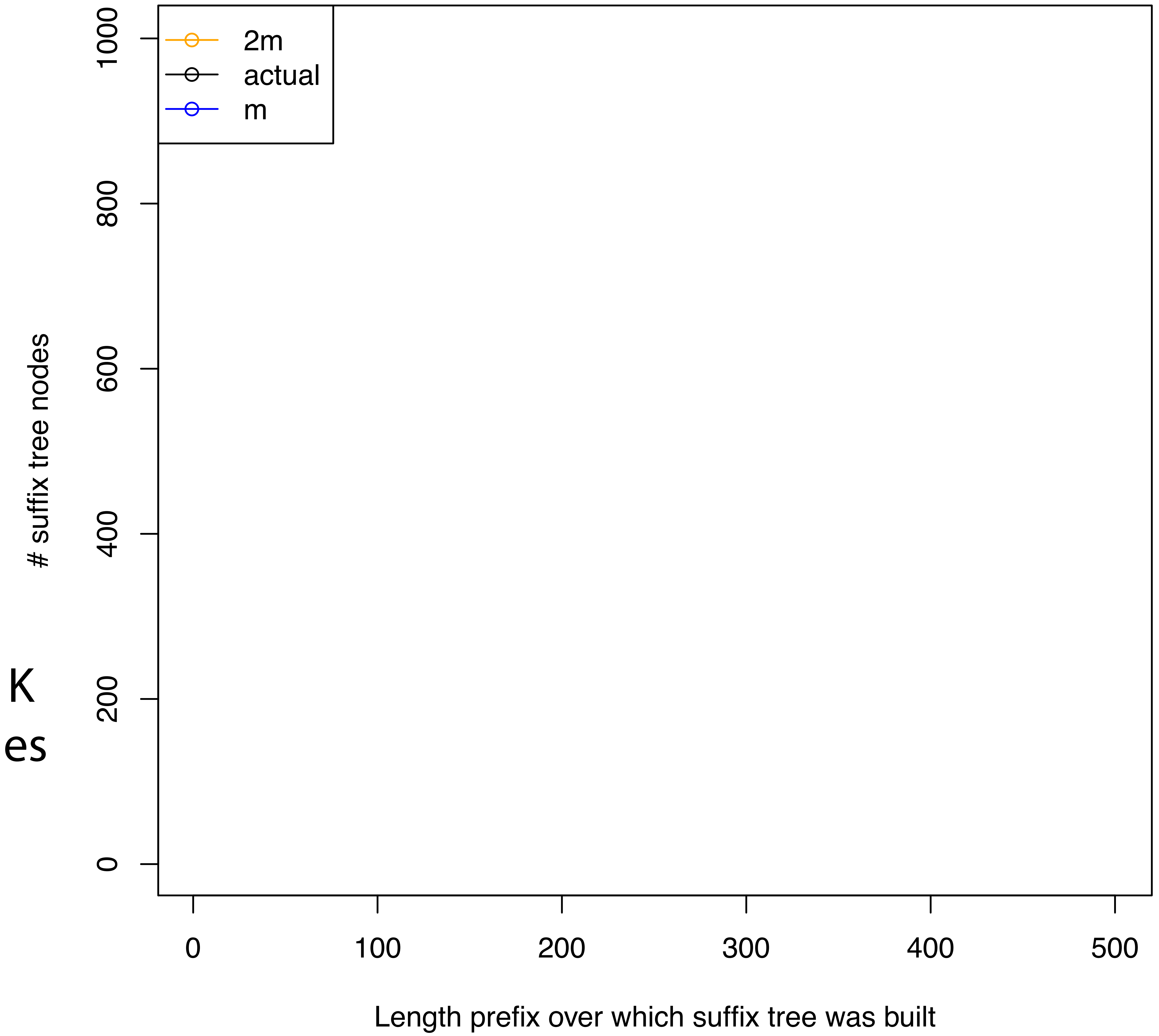
Built suffix trees for the first 500 prefixes of the lambda phage virus genome

Black curve shows # nodes increasing with prefix length

Remember suffix trie plot:



123 K nodes

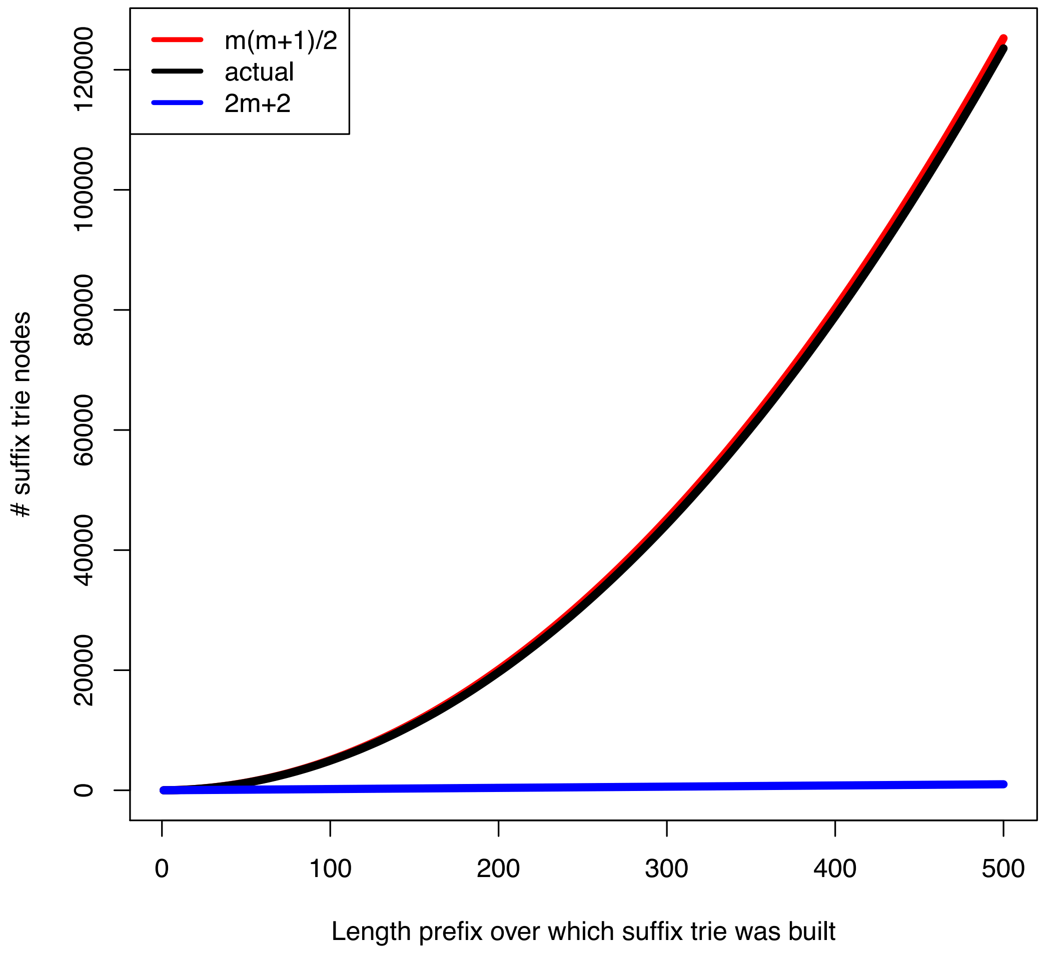


Suffix tree: actual growth

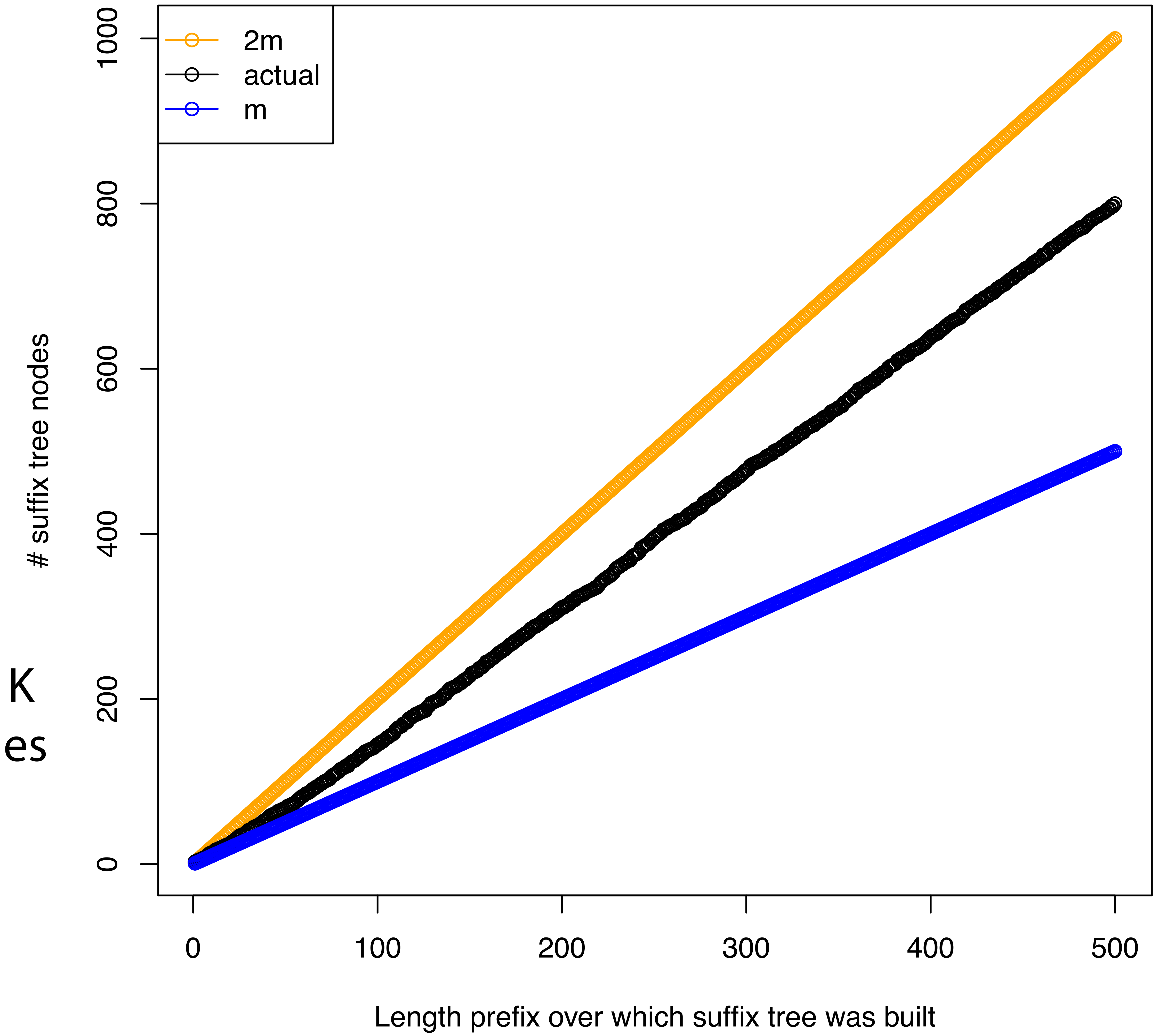
Built suffix trees for the first 500 prefixes of the lambda phage virus genome

Black curve shows # nodes increasing with prefix length

Remember suffix trie plot:

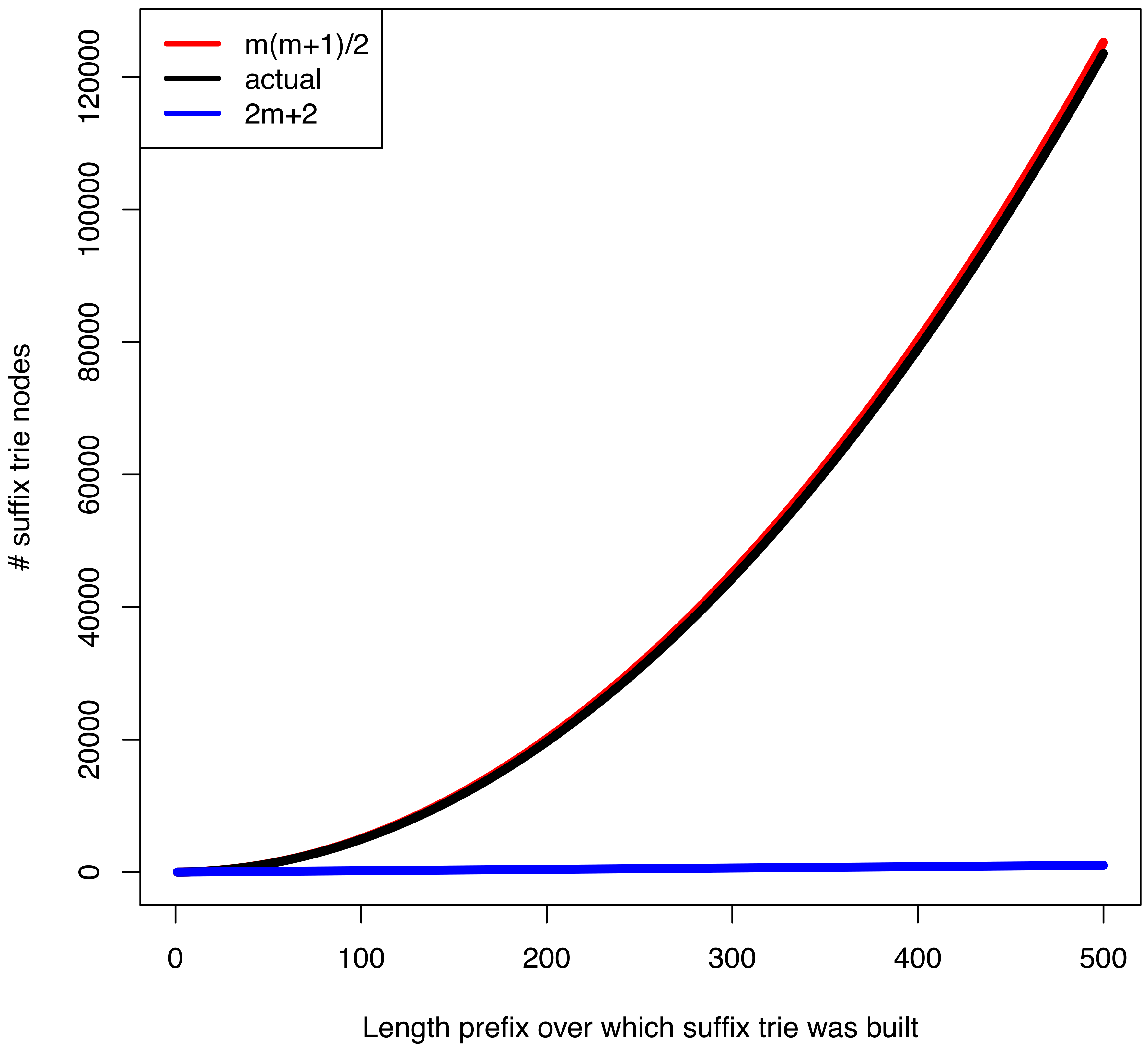


123 K nodes



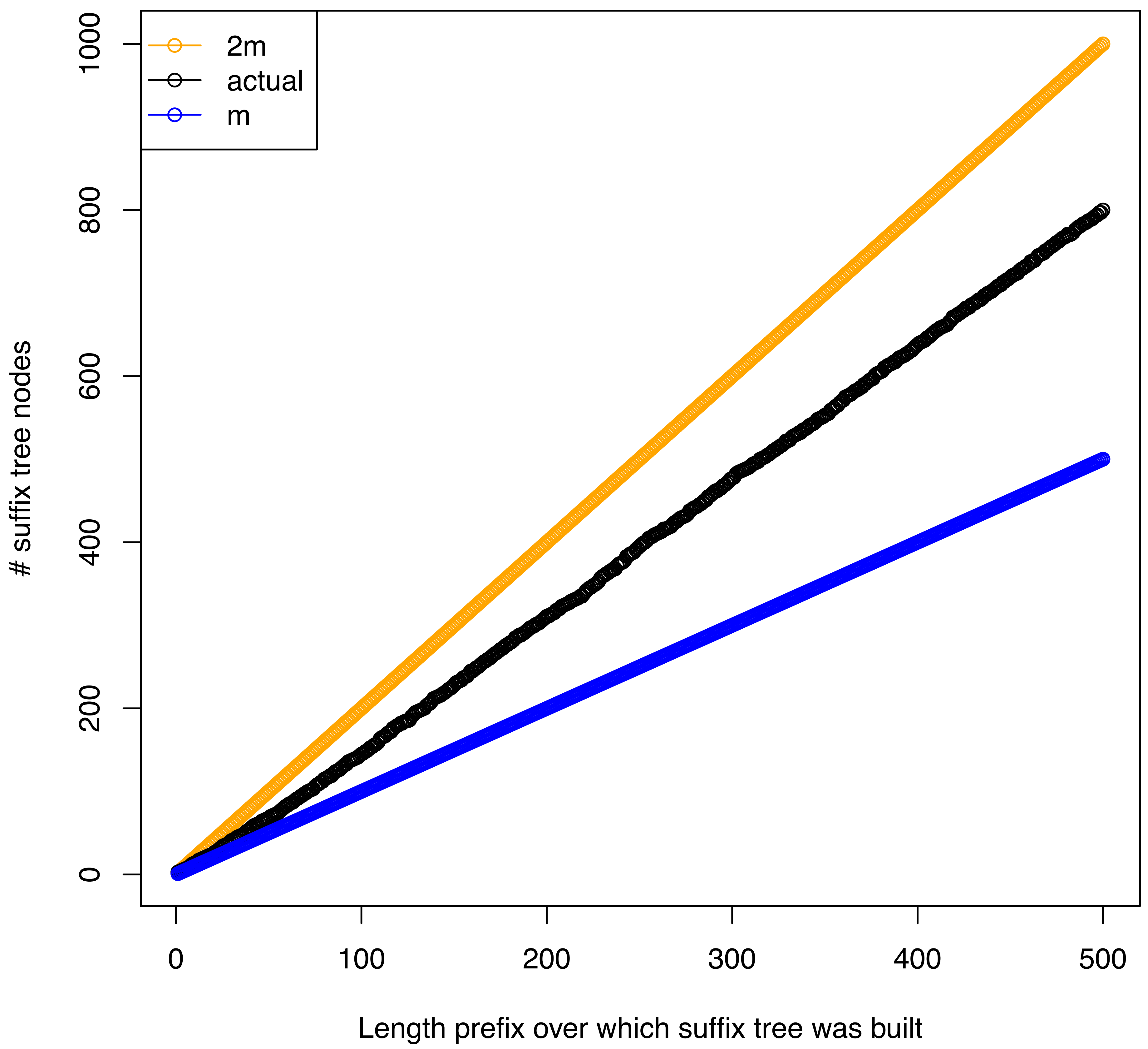
Suffix trie

>100K nodes



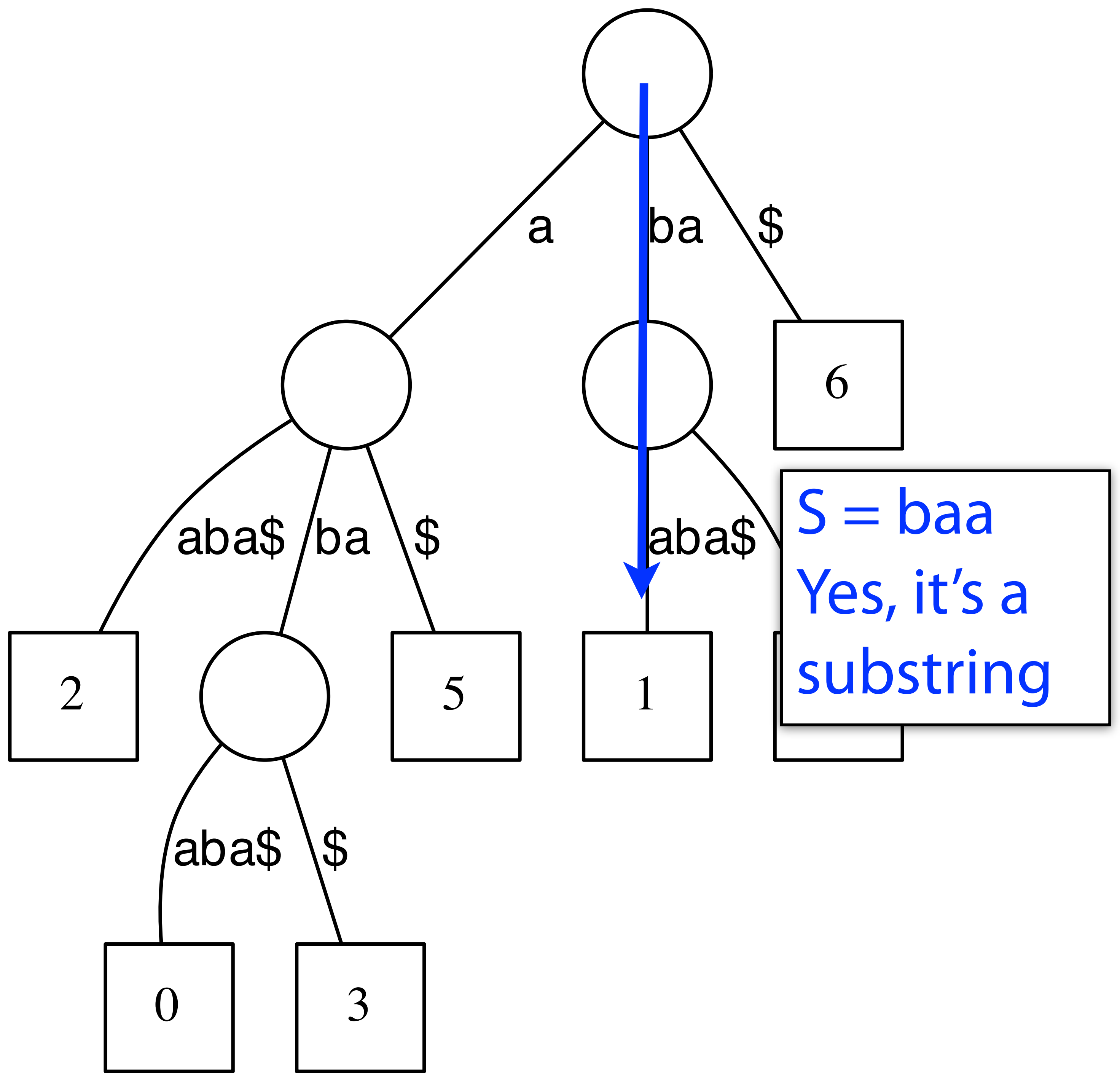
Suffix tree

<1K nodes



Suffix tree

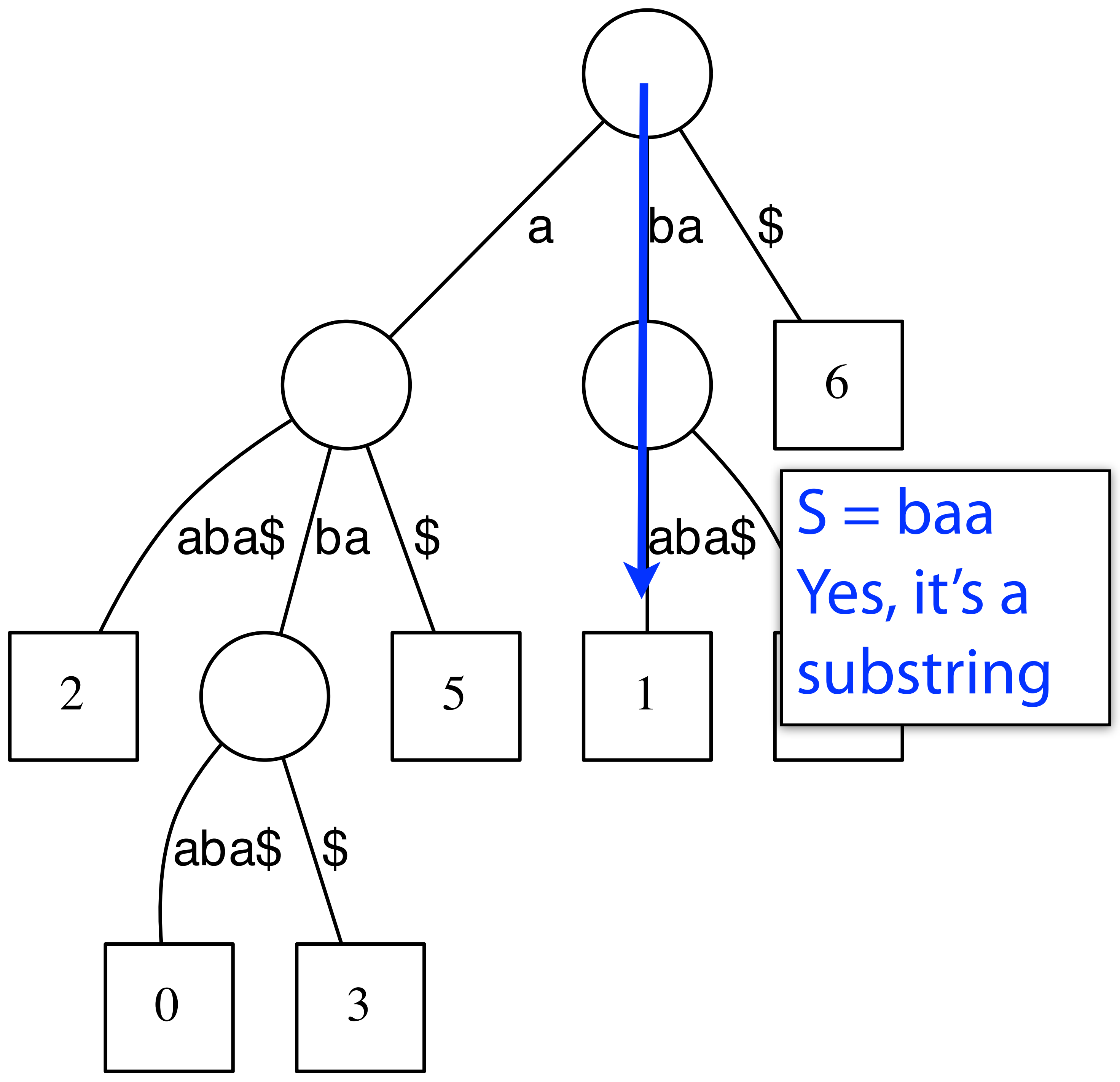
How do we check whether a string *S* is a substring of *T*?



Suffix tree

How do we check whether a string *S* is a substring of *T*?

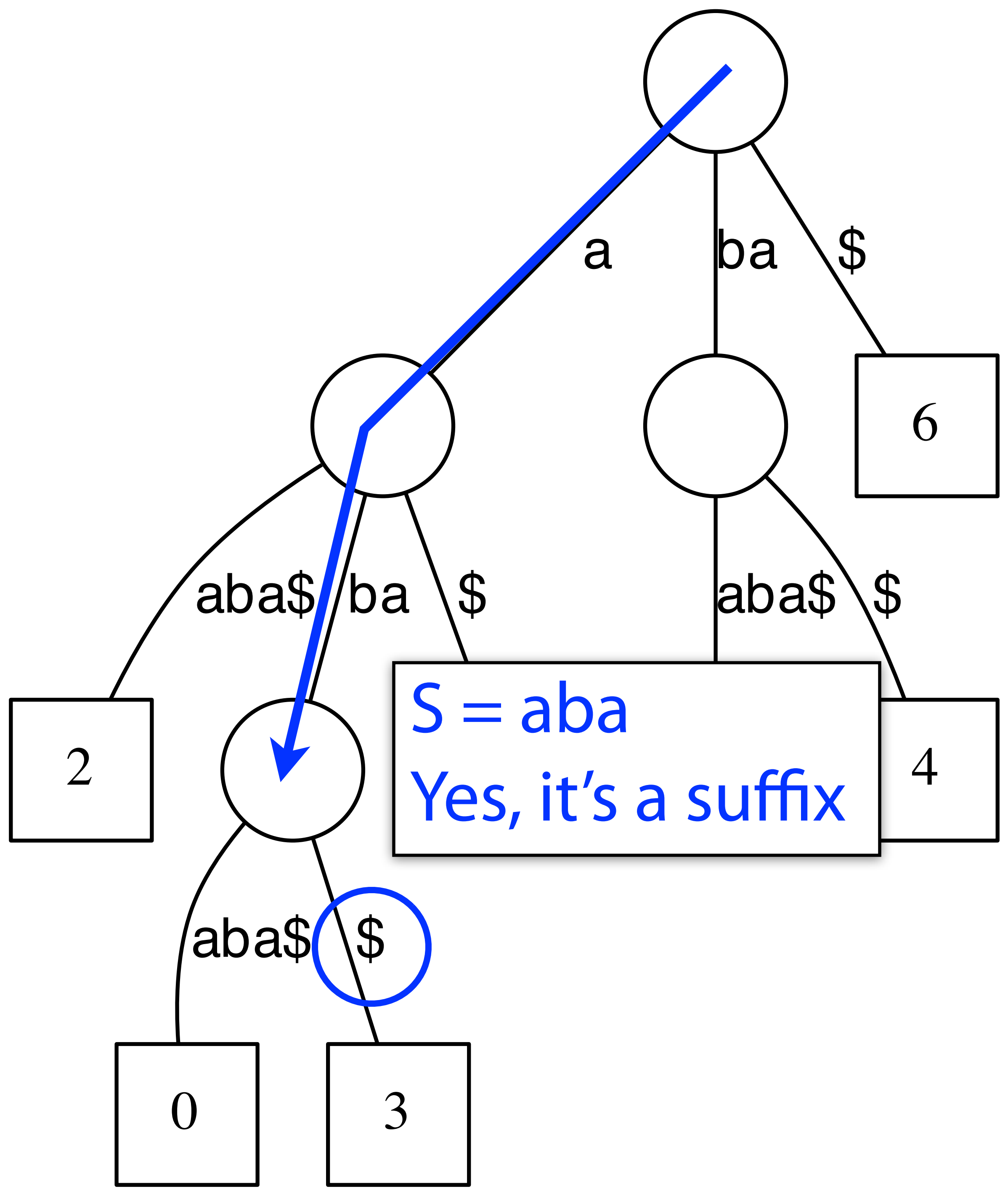
Essentially same procedure as for suffix trie, except we have to deal with coalesced edges



Suffix tree

How do we check whether a string *S* is a suffix of *T*?

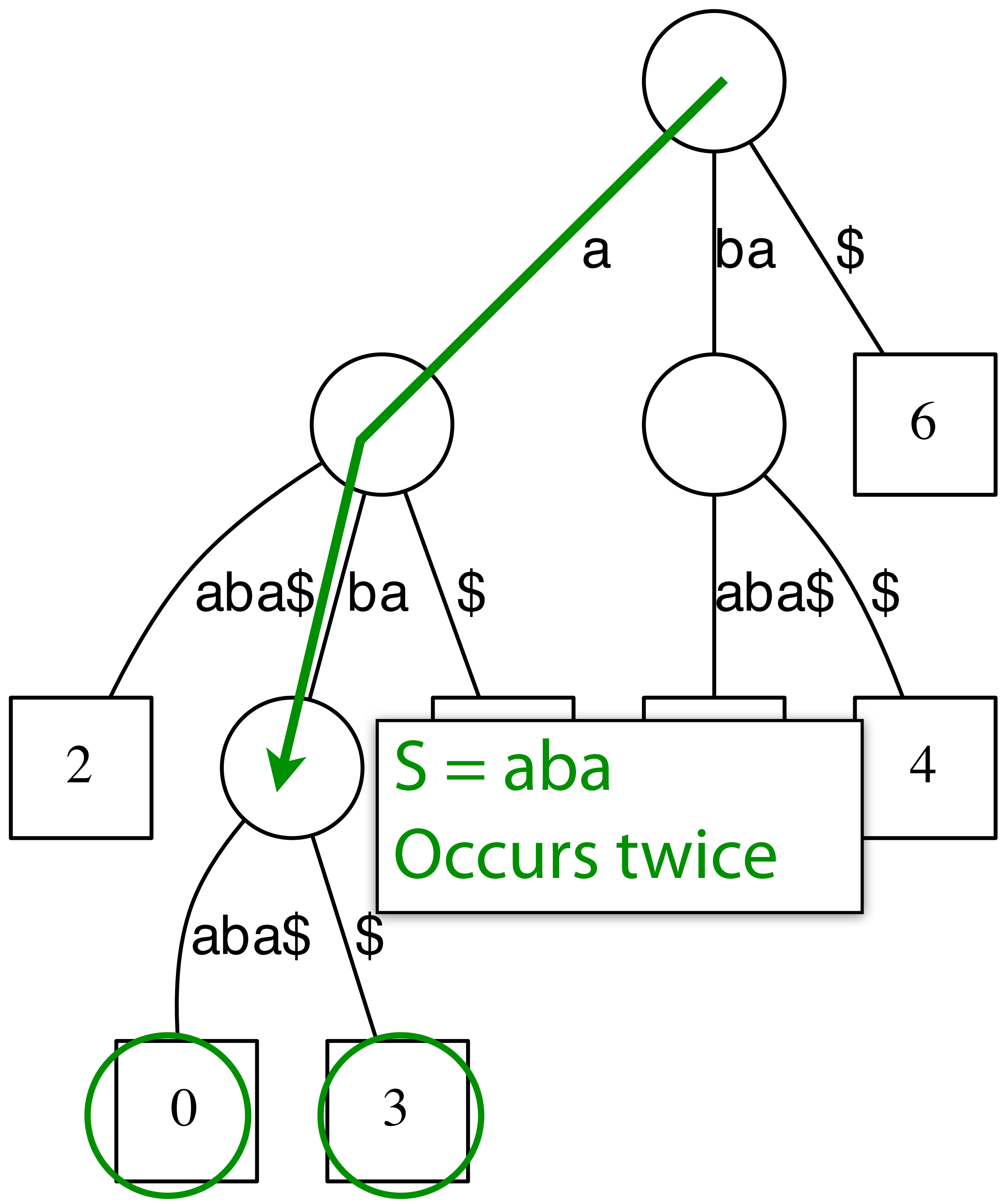
Essentially same procedure as for suffix trie, except we have to deal with coalesced edges



Suffix tree

How do we count the **number of times** a string *S* occurs as a substring of *T*?

Same procedure as for suffix trie



Suffix tree: applications

With suffix tree of T, we can find all matches of P to T. Let k = # matches.

E.g., P = ab, T = abaaba$

$O(n)$

Step 1: walk down ab path

If we "fall off" there are no matches

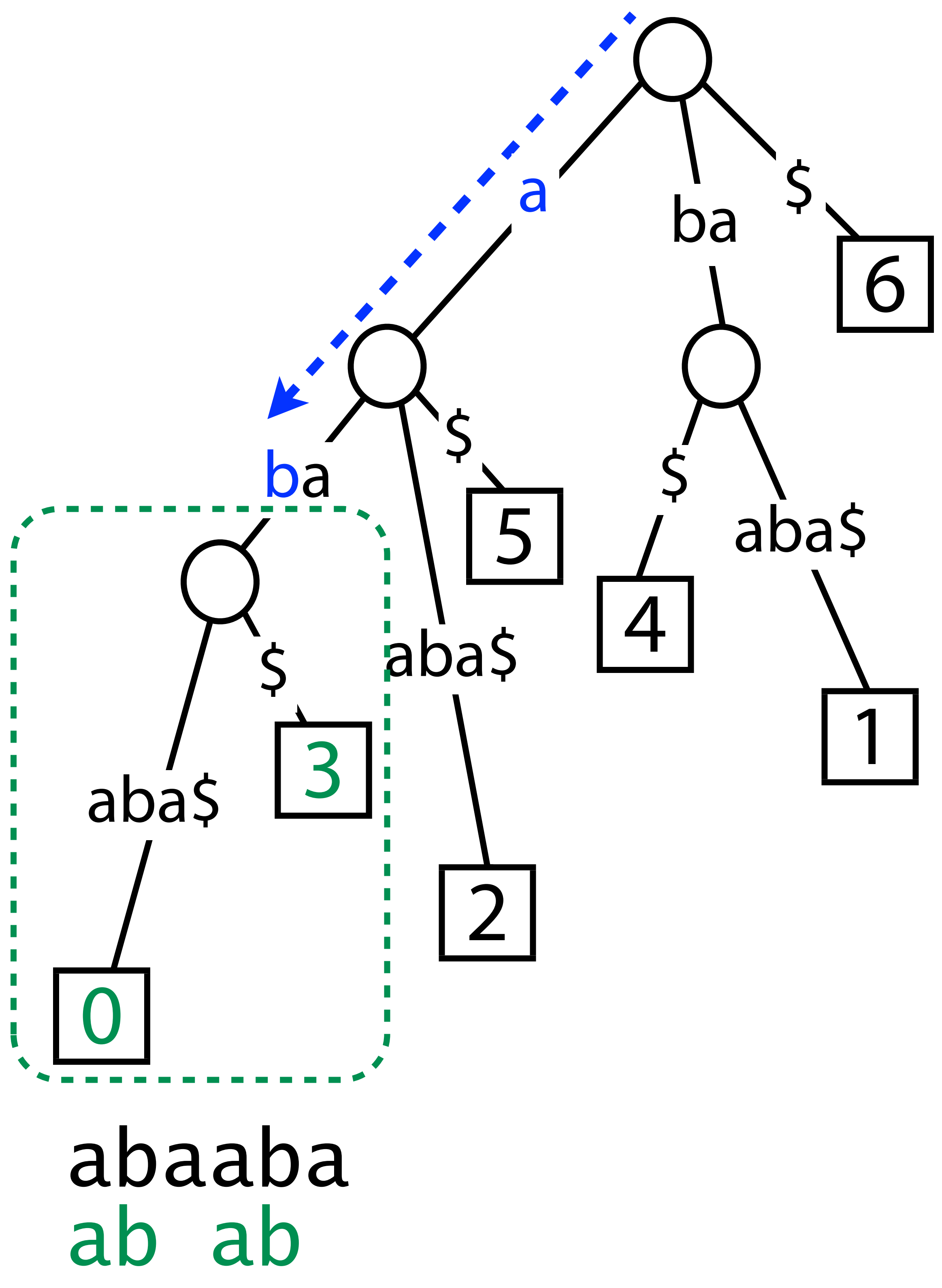
$O(k)$

Step 2: visit all leaf nodes below

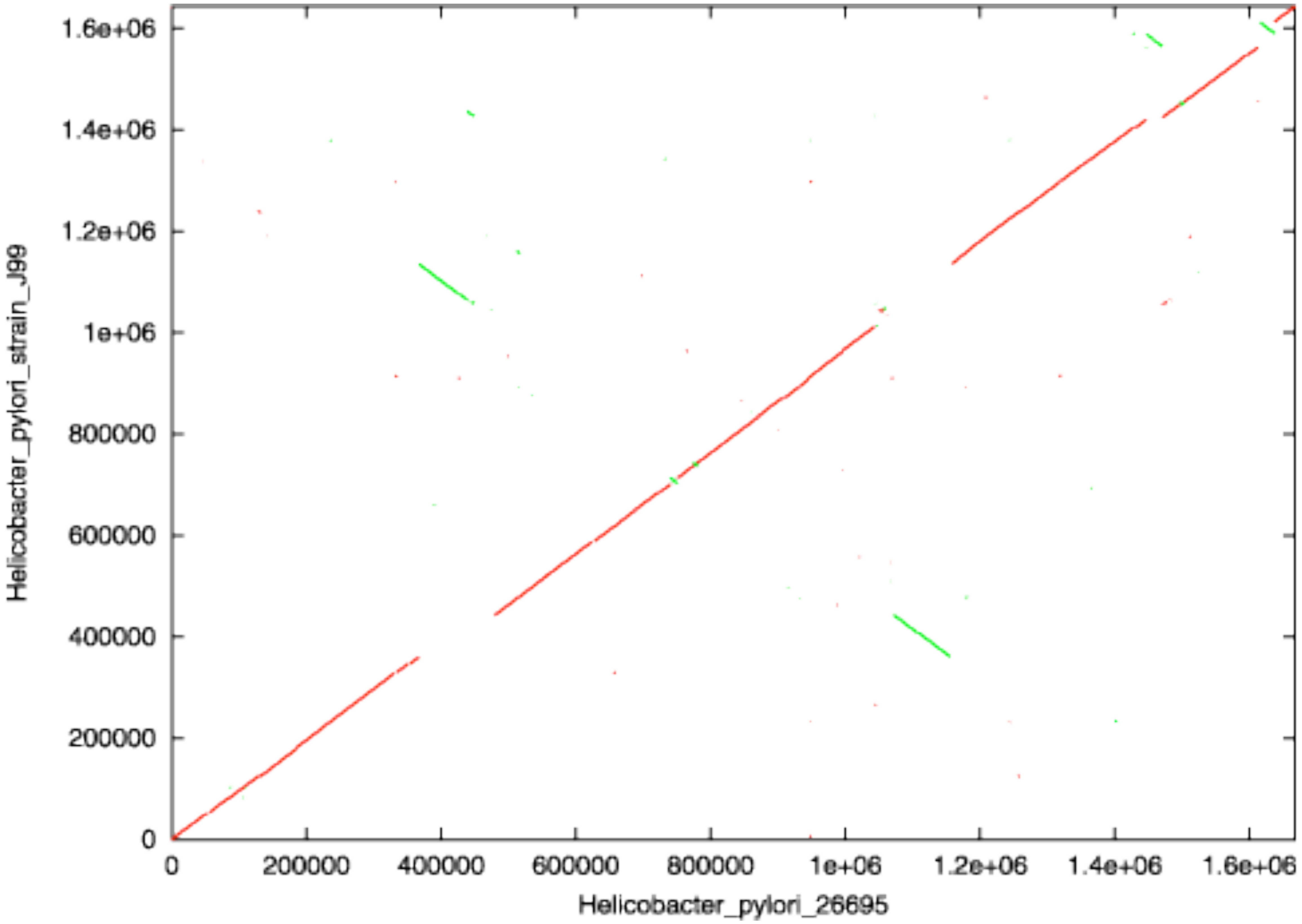
Report each leaf offset as match offset

With proper tree modifications to access leaves of a node in O(1) each

$O(n + k)$ time



Suffix tree application: find long common substrings



Dots are *maximal unique matches* (*MUMs*), a kind of long substring shared by two sequences

Red = match was between like strands, green = different strands

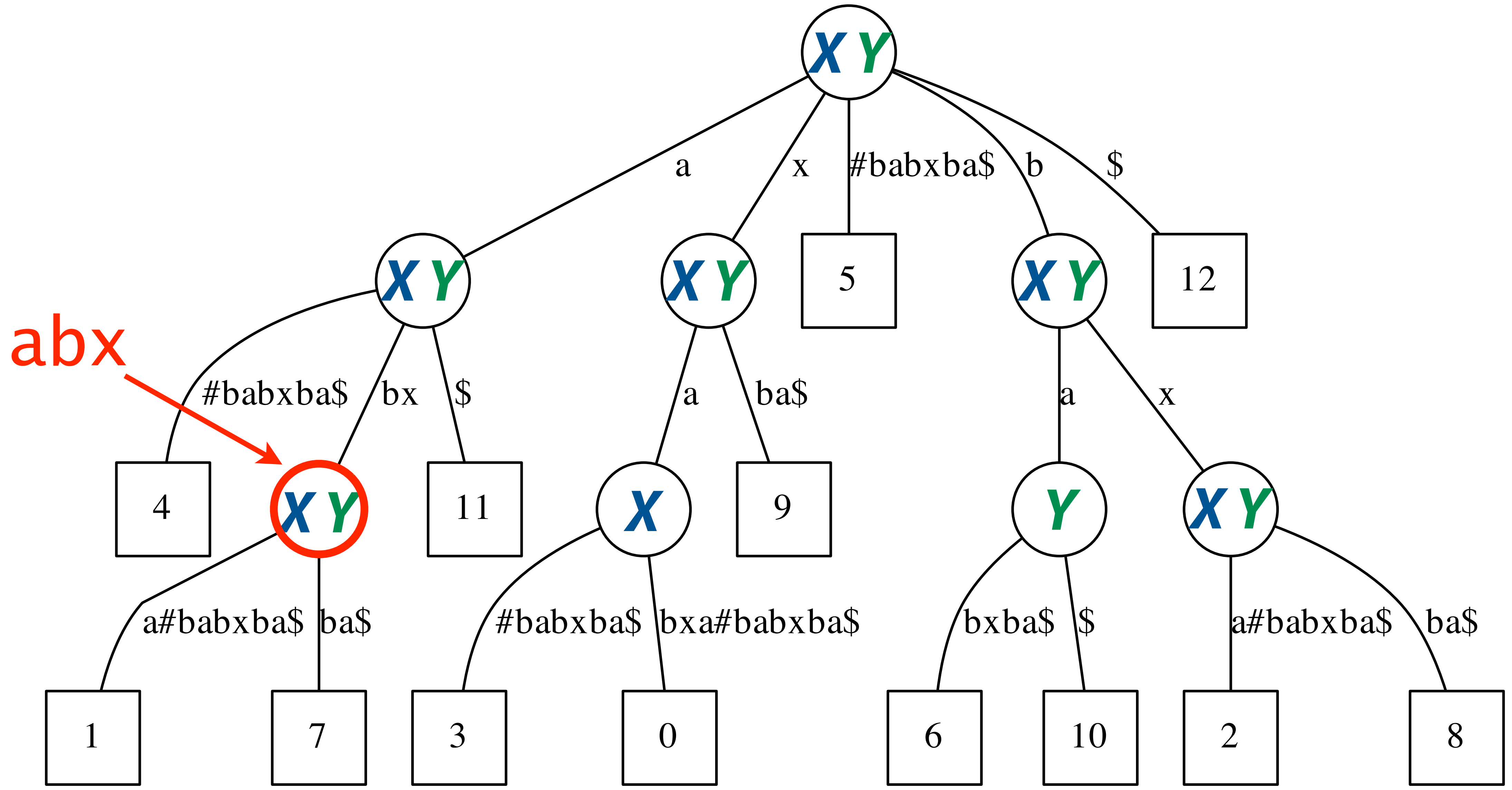
Axes show different strains of Helicobacter pylori, a bacterium found in the stomach and associated with gastric ulcers

Suffix tree application: find longest common substring

To find the longest common substring (LCS) of *X* and *Y*, make a new string $X\#Y\$$ where $\#$ and $\$$ are both terminal symbols. Build a suffix tree for $X\#Y\$$.

X = xabxa
Y = babxba
***X*#*Y*$** = xabxa#babxba$

Consider leaves: offsets in [0, 4] are suffixes of ***X***, offsets in [6, 11] are suffixes of ***Y***



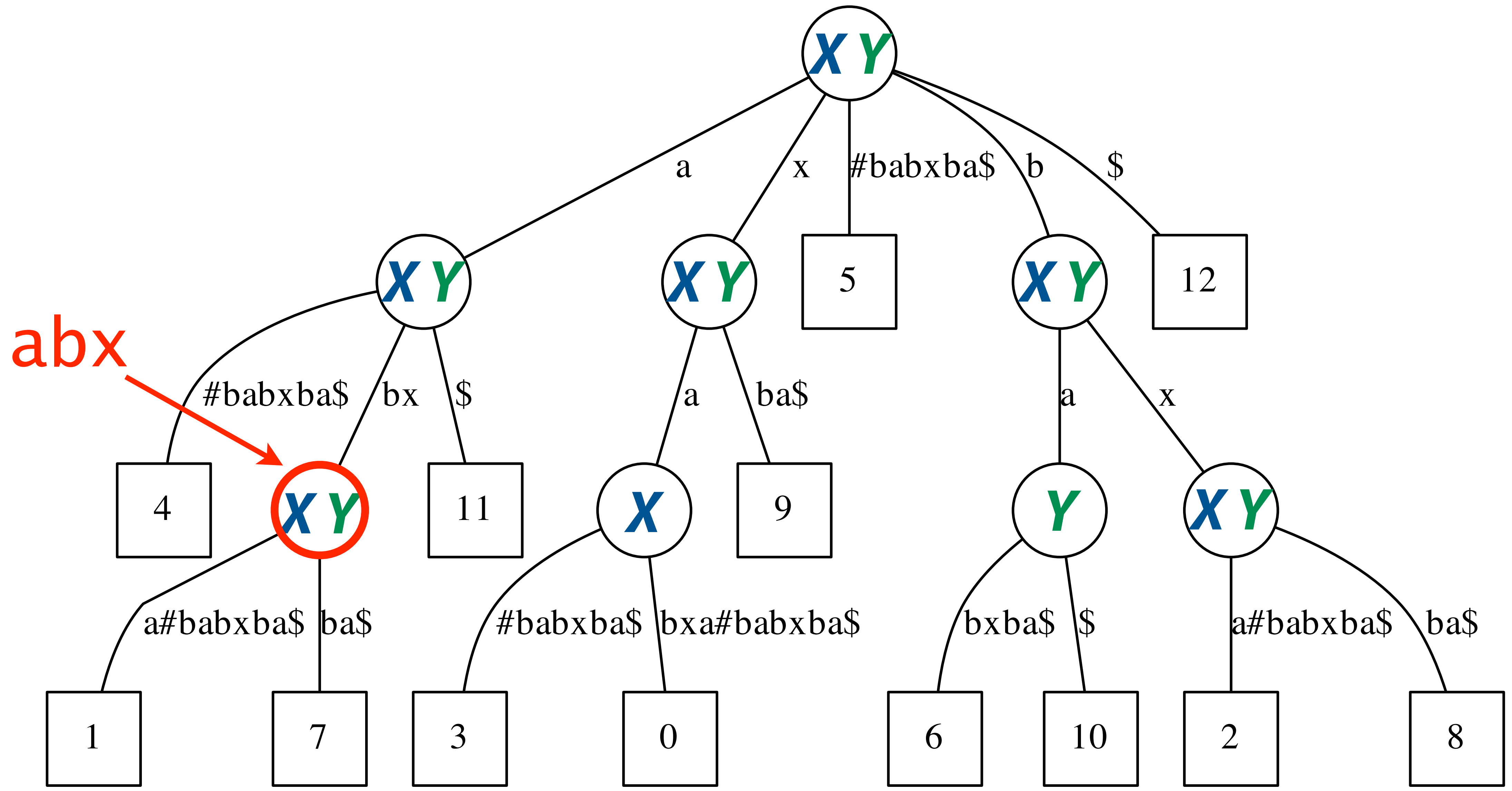
Traverse the tree and annotate each node according to whether leaves below it include suffixes of ***X***, ***Y*** or both

The deepest node annotated with both ***X*** and ***Y*** has LCS as its label. $O(|\,X\,|+|\,Y\,|)$ time and space.

Suffix tree application: generalized suffix trees

This is one example of many applications where it is useful to build a suffix tree over many strings at once

Such a tree is called a *generalized suffix tree*.



Longest Common Extension

Longest common extension: We are given strings S and T. In the future, many pairs (i,j) will be provided as queries, and we want to quickly find:

the longest substring of S starting at i that matches a substring of T starting at j.

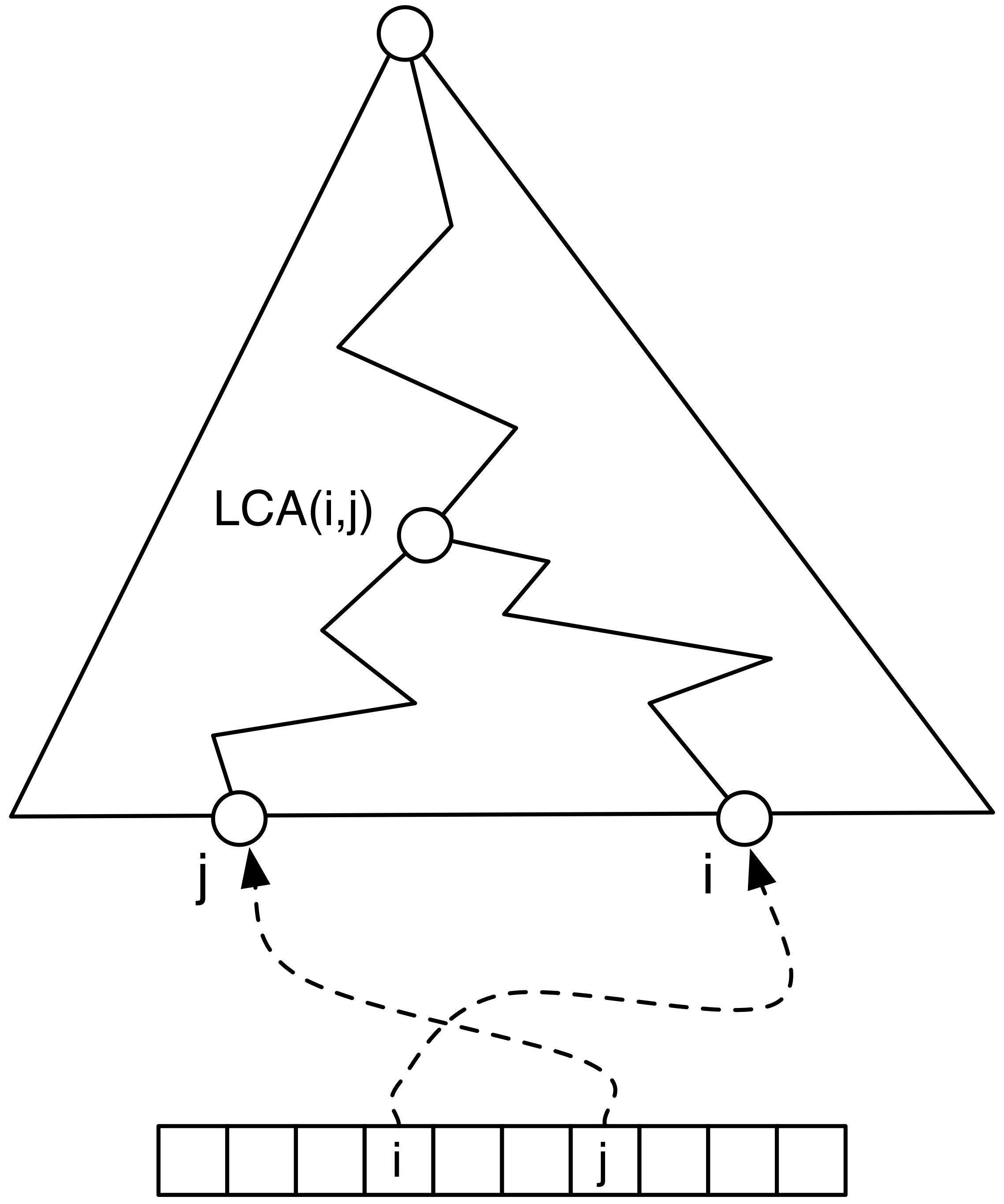
Build generalized suffix tree for S and T.

Preprocess tree so that lowest common ancestors (LCA) can be found in constant time. This can be

done using range-minimum queries (RMQ)

LCA in O(1) time

https://www.ics.uci.edu/~eppstein/261/BenFar-LCA-00.pdf



Create an array mapping suffix numbers to leaf nodes.

Given query (i,j):
- Find the leaf nodes for i and j
- Return string of LCA for i and j

Longest Common Extension

Longest common extension: We are given strings S and T. In the future, many pairs (i,j) will be provided as queries, and we want to quickly find:

the longest substring of S starting at i that matches a substring of T starting at j.

Build generalized suffix tree for S and T. O(|S| + |T|)

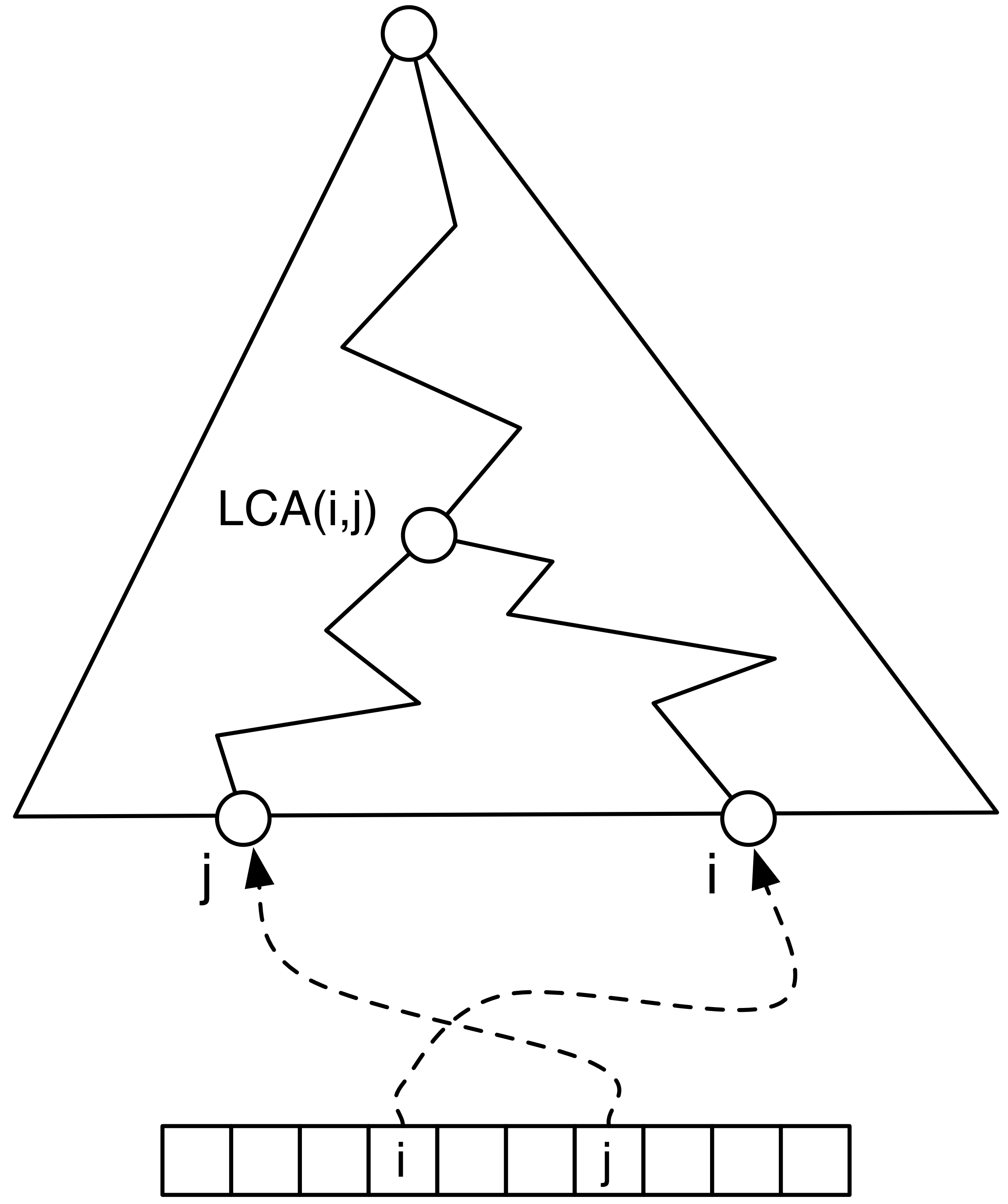
Preprocess tree so that lowest common ancestors (LCA) can be found in constant time. This can be done using range-minimum queries (RMQ) O(|S| + |T|)

LCA in O(1) time

https://www.ics.uci.edu/~eppstein/261/BenFar-LCA-00.pdf

Create an array mapping suffix numbers to leaf nodes. O(|S| + |T|)

Given query (i,j):
  Find the leaf nodes for i and j O(1)
  Return string of LCA for i and j O(LCA(i,j))



Indexed search in practice : MUMMER

“Text” (human chr1)

“Pattern” (LINE-1 Long interspersed nuclear element)

```
(base) rob@fermi mummer-4.0.0rc1 % ./mummer -maxmatch data/chr1.fa data/line1.fa | head -n100
> NC_000001.11:62194849-62212328
62194849        1    17480
62195013      125       21
62194973      165       21
62195059      171       35
62195189      181       23
62195019      211       35
62195189      221       23
62195029      341       23
62195069      341       23
227144292      621       22
226080010      623       45
150924429      623       29
109966311      624       44
160754930      624       44
13373677      624       44
28521220      624       44
195870013      624       44
155102021      624       31
39601269      624       31
35187449      624       30
168153630      624       30
155228268      624       28
117635680      624       24
37579266      624       21
155399348      624       20
176178570      624       20
174043843      624       20
23923078      624       20
169606035      624       20
28188030      624       20
232504257      624       20
27791535      625       37
112543271      627       34
150082229      628       40
```

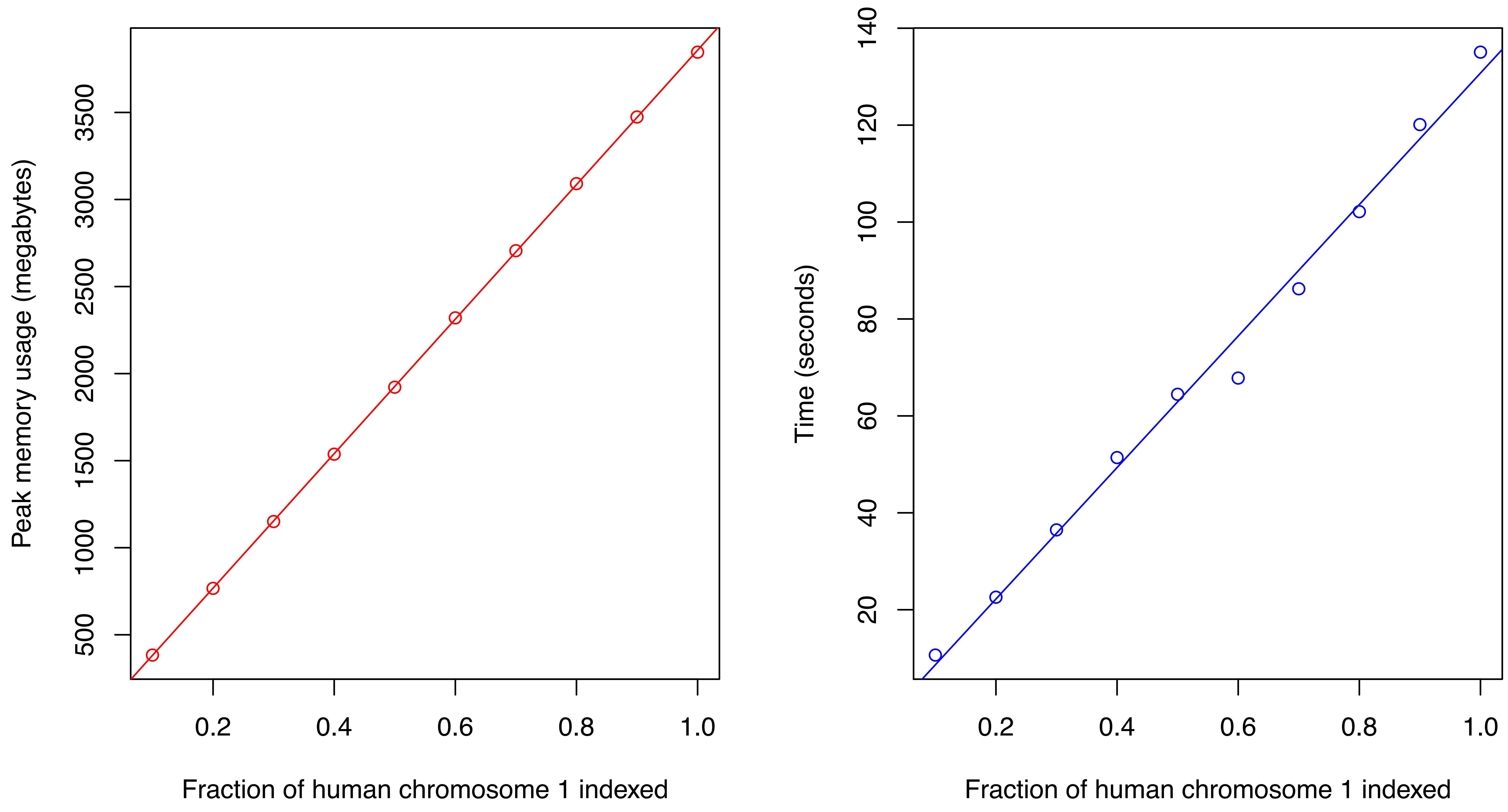
Output format (stdout):
column 1: offset in T
column 2: offset in P
column 3: length of match

MUMMER first builds an index on T (this takes about a minute for chr1 on my machine)

Then is searches for maximal matches shared between T and P — these are output very fast (a second or so; and there are **many**).

Suffix trees in the real world: MUMmer

MUMmer v3.32 time and memory scaling when indexing increasingly larger fractions of human chromosome 1



For whole chromosome 1, took 2m:14s and used 3.94 GB memory

Suffix trees in the real world: MUMmer

Attempt to build index for whole human genome reference:

```
mummer: suffix tree construction failed: textlen=3101804822
larger than maximal textlen=536870908
```

We can predict it would have taken about 47 GB of memory

Suffix trees in the real world: the constant factor

While $O(m)$ is desirable, the constant in front of the m limits wider use of suffix trees in practice

Constant factor varies depending on implementation:

Estimate of MUMmer's constant factor = 3.94 GB / 250 million nt ≈ **15.75 bytes per node**

Literature reports implementations achieving as little as 8.5 bytes per node, but no implementation used in practice that I know of is better than ≈ **12.5 bytes per node**

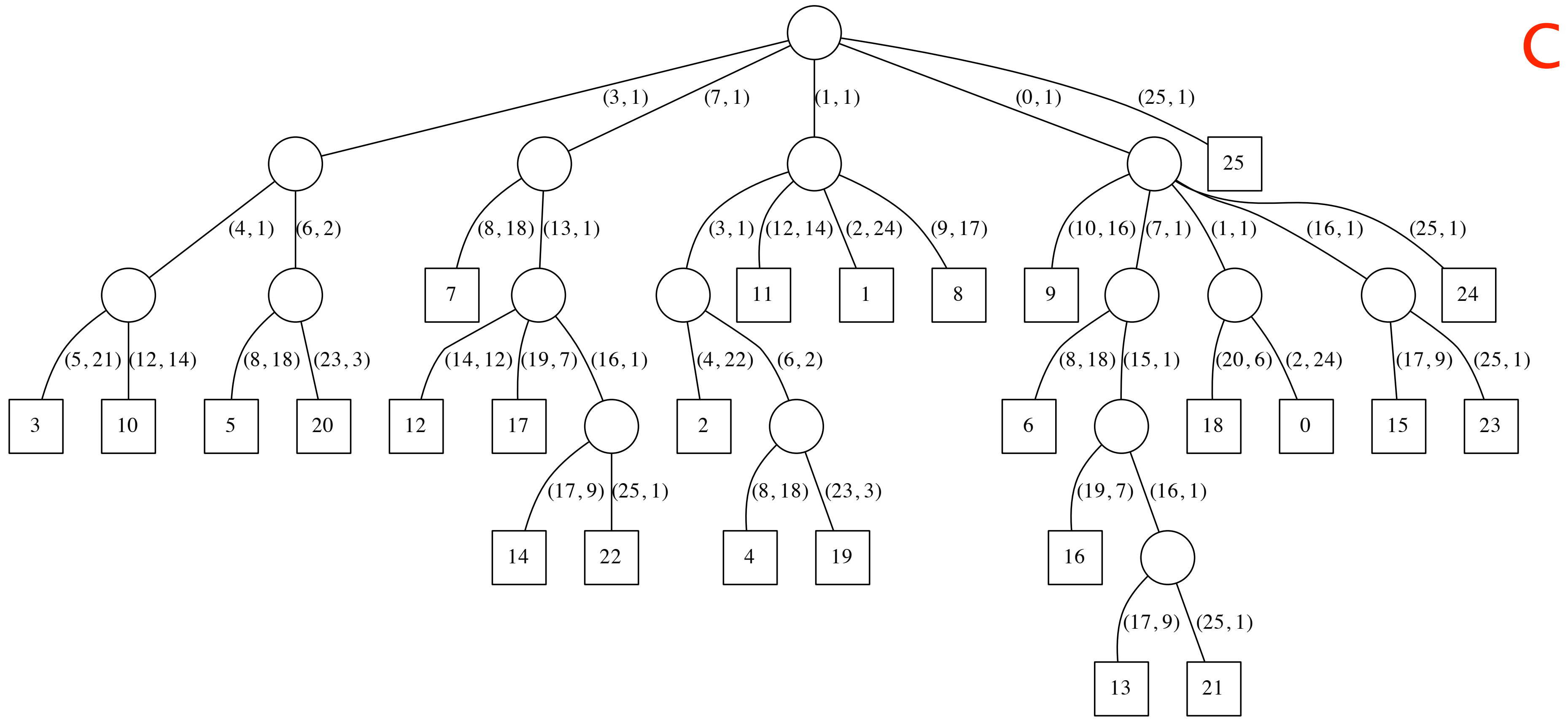
Kurtz, Stefan. "Reducing the space requirement of suffix trees." *Software Practice and Experience* 29.13 (1999): 1149-1171.

Suffix tree: summary

Organizes all suffixes into an incredibly useful, flexible data structure, in $O(m)$ time and space

A naive method (e.g. suffix trie) could easily be quadratic or worse

Used in practice for whole genome alignment, repeat identification, etc



Actual memory footprint (bytes per node) is quite high, limiting usefulness

GTTATAGCTGATCGCGGCGTAGCGG$ *m* chars

GTTATAGCTGATCGCGGCGTAGCGG$
TTATAGCTGATCGCGGCGTAGCGG$
TATAGCTGATCGCGGCGTAGCGG$
ATAGCTGATCGCGGCGTAGCGG$
TAGCTGATCGCGGCGTAGCGG$
AGCTGATCGCGGCGTAGCGG$
GCTGATCGCGGCGTAGCGG$
CTGATCGCGGCGTAGCGG$
TGATCGCGGCGTAGCGG$
GATCGCGGCGTAGCGG$
ATCGCGGCGTAGCGG$
TCGCGGCGTAGCGG$
CGCGGCGTAGCGG$
GCGGCGTAGCGG$
CGGCGTAGCGG$
GGCGTAGCGG$
GCGTAGCGG$
CGTAGCGG$
GTAGCGG$
TAGCGG$
AGCGG$
GCGG$
CGG$
GG$
G$
$

m(m+1)/2 chars

Bonus Content (not required) :
Suffix tree construction

WOTD (Write-Only Top-Down) Construction

Giegerich, Robert, and Stefan Kurtz. "A comparison of imperative and purely functional suffix tree constructions." Science of Computer Programming 25.2 (1995): 187-218.

Build a suffix tree for string s$

Recursive construction:

For every branching node **node**(u), subtree of **node**(u) is determined by all suffixes of s$ where u is a prefix.

Recursively construct subtree for all suffixes where u is a prefix.

Definition: *remaining suffixes* of u

R(node(u)) = { v | uv is a suffix of s$ }

*

WOTD (Write-Only Top-Down) Construction

Build a suffix tree for string s$

Recursive construction:

For every branching node **node**(u), subtree of **node**(u) is determined by all suffixes of s$ where u is a prefix.

Recursively construct subtree for all suffixes where u is a prefix.

Definition: *remaining suffixes* of u

$$R(\text{node}(u)) = \{ v \mid uv \text{ is a suffix of } s\$ \}$$

Definition: *c-group* of node(u)

$$\text{group}(\text{node}(u), c) = \{ w \in \Sigma^* \mid cw \in R(\text{node}(u))\}$$

*

WOTD (Write-Only Top-Down) Construction

```
def WOTD(T : tree, node(u): node):
   for each c ∈ Σ ∪ {$}:
      G = group(node(u), c)
      ucv = lcp(G)
      if |G| == 1:
         add leaf node(ucv) as a child of node(u)
      else:
         add inner node(ucv) as a child of node(u)
         WOTD(T, node(ucv))
```

non-branching suffix

branching suffix

Start the algorithm by calling WOTD(T, node(ϵ))

root node

*

WOTD Example

s = ttatctctta$

suffixes are
read top-to-bottom

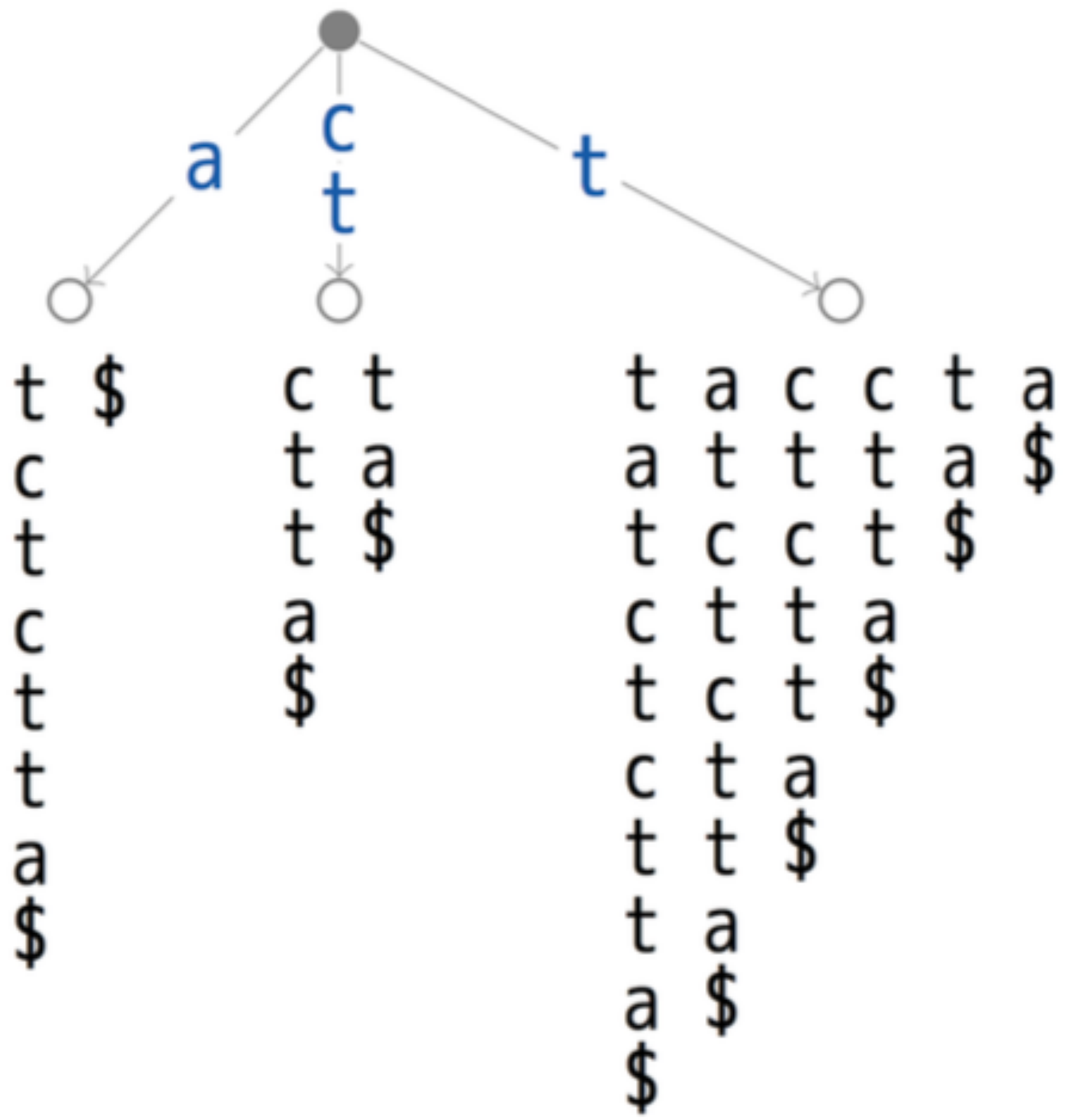
```
t t a t c t c t t a $
t a t c t c t t a $
a t c t c t t a $
t c t c t t a $
c t c t t a $
t c t t a $
c t t a $
t t a $
t a $
a $
$
```

*

example from: http://www.mi.fu-berlin.de/wiki/pub/ABI/SS13Lecture4Materials/wotd.pdf

WOTD Example

s = ttatctctta

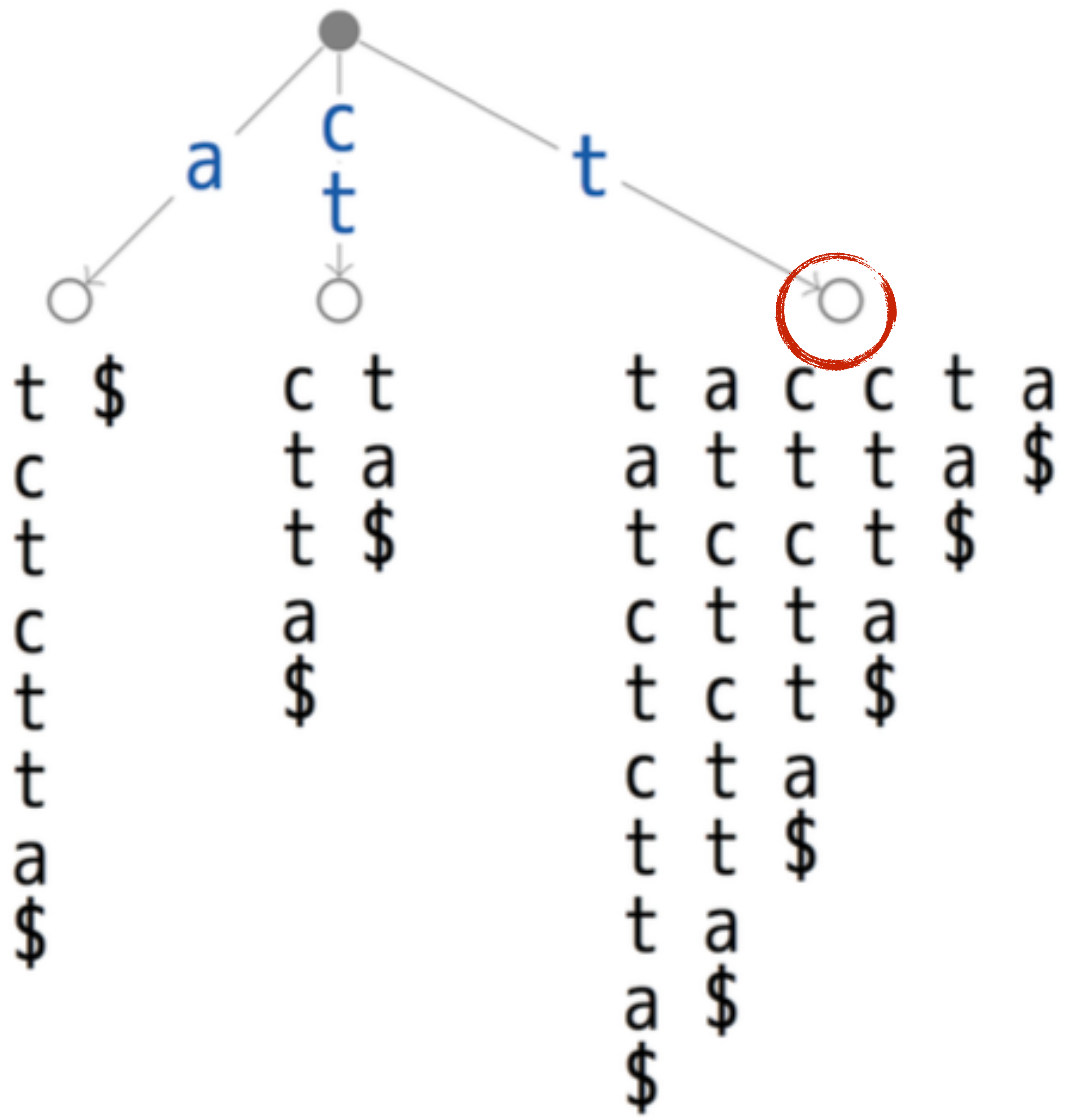


*

example from: http://www.mi.fu-berlin.de/wiki/pub/ABI/SS13Lecture4Materials/wotd.pdf

WOTD Example

s = ttatctctta

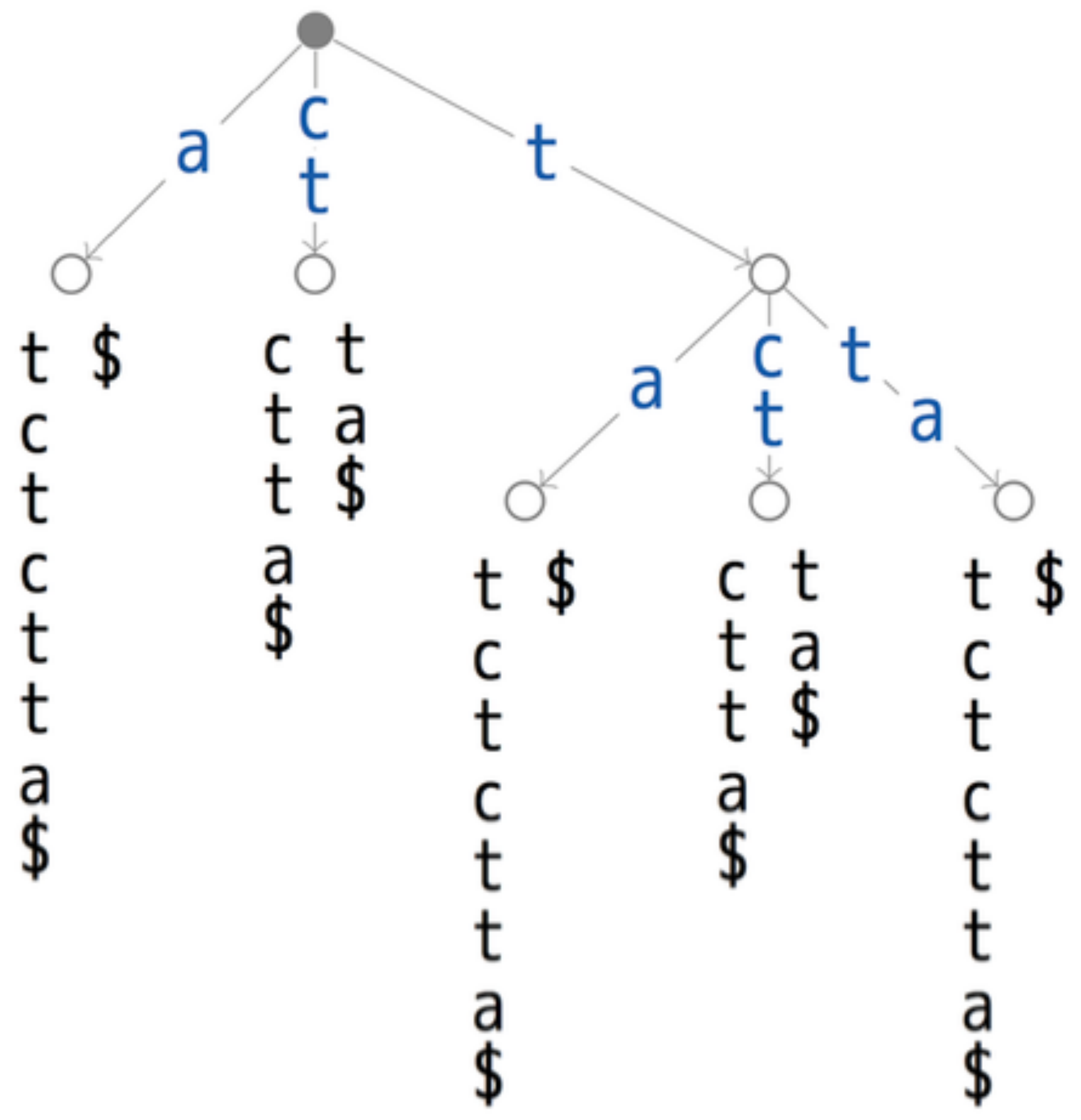


*

example from: http://www.mi.fu-berlin.de/wiki/pub/ABI/SS13Lecture4Materials/wotd.pdf

WOTD Example

s = ttatctctta

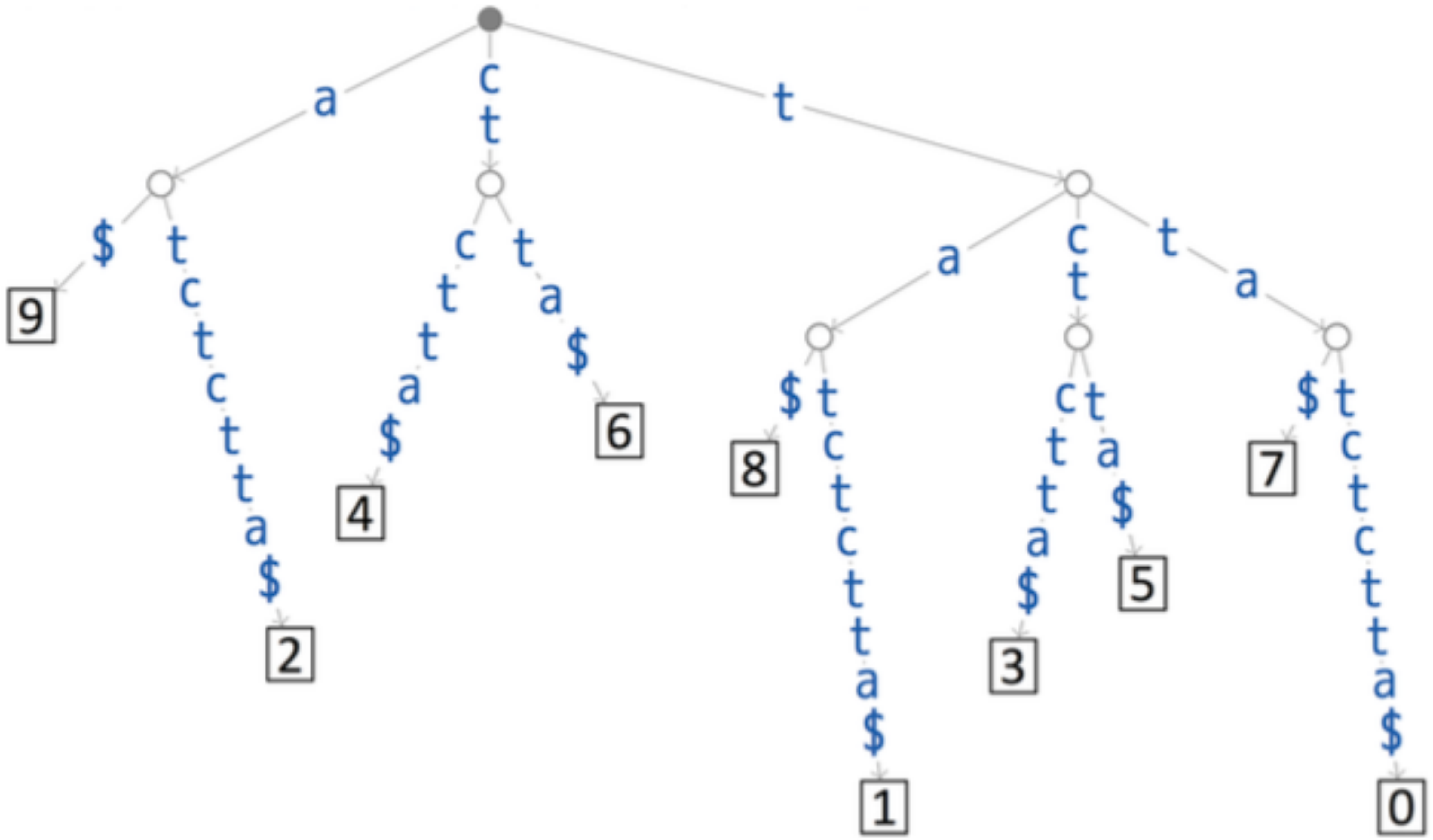


example from: http://www.mi.fu-berlin.de/wiki/pub/ABI/SS13Lecture4Materials/wotd.pdf

*

WOTD Example

s = ttatctctta



example from: http://www.mi.fu-berlin.de/wiki/pub/ABI/SS13Lecture4Materials/wotd.pdf

WOTD Properties

- Worst case time still $\in O(|T|^2)$

- Expected case time $\in O(|T| \log |T|)$

- Write-only property & recursive construction lends itself well to parallelism

- Good caching properties (locality of reference for substrings belonging to a subtree)

- Top-down construction order allows lazy construction as discussed in:

Giegerich, Robert, Stefan Kurtz, and Jens Stoye. "Efficient implementation of lazy suffix trees." Software: Practice and Experience 33.11 (2003): 1035-1049.