Suffix Trees: definition & size

Ben Langmead

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Department of Computer Science

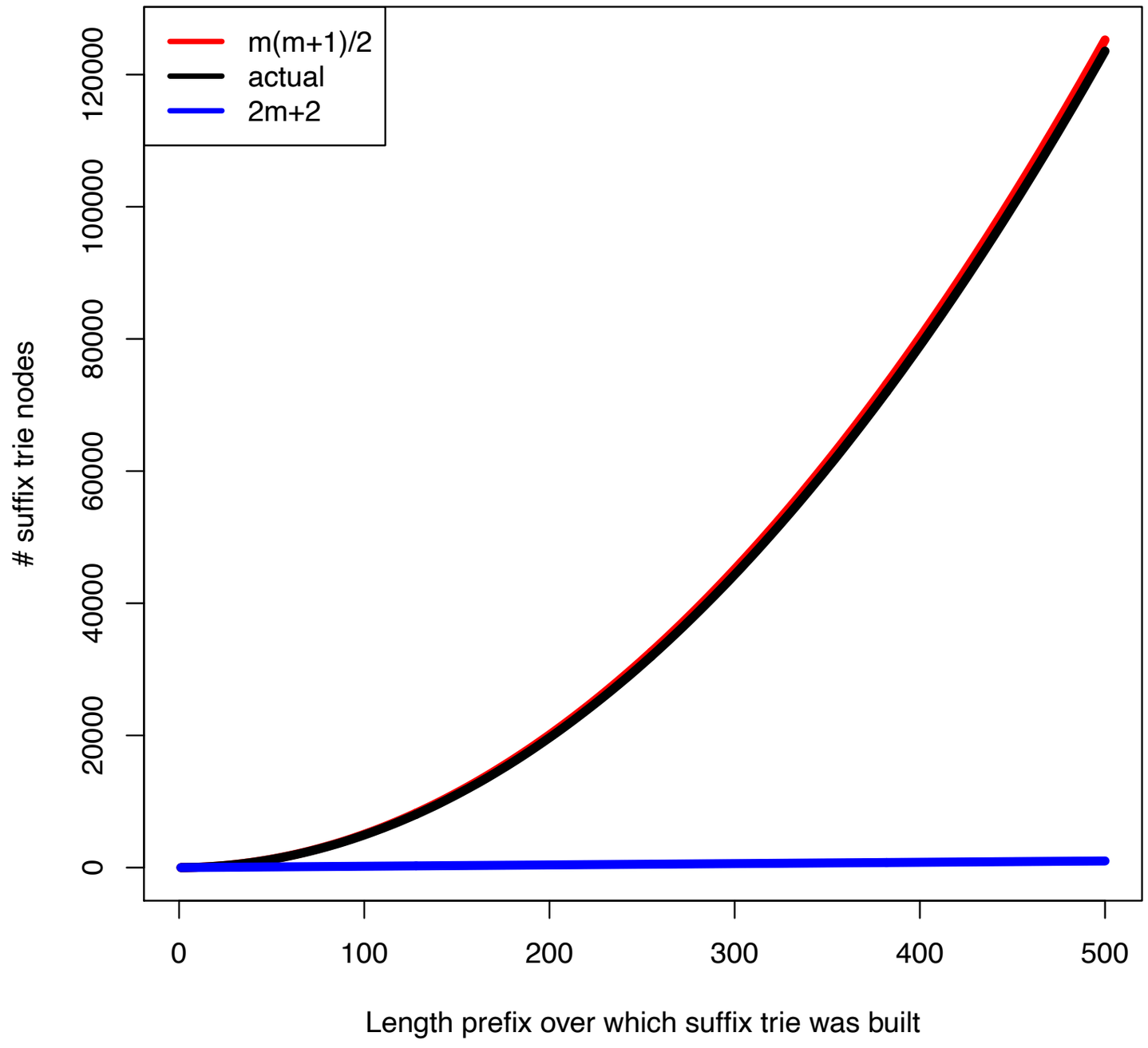
Please sign guestbook (www.langmead-lab.org/teaching-materials) to tell me briefly how you are using the slides. For original Keynote files, email me (ben.langmead@gmail.com).

Suffix trie

Suffix trie grows *quadratically* with string

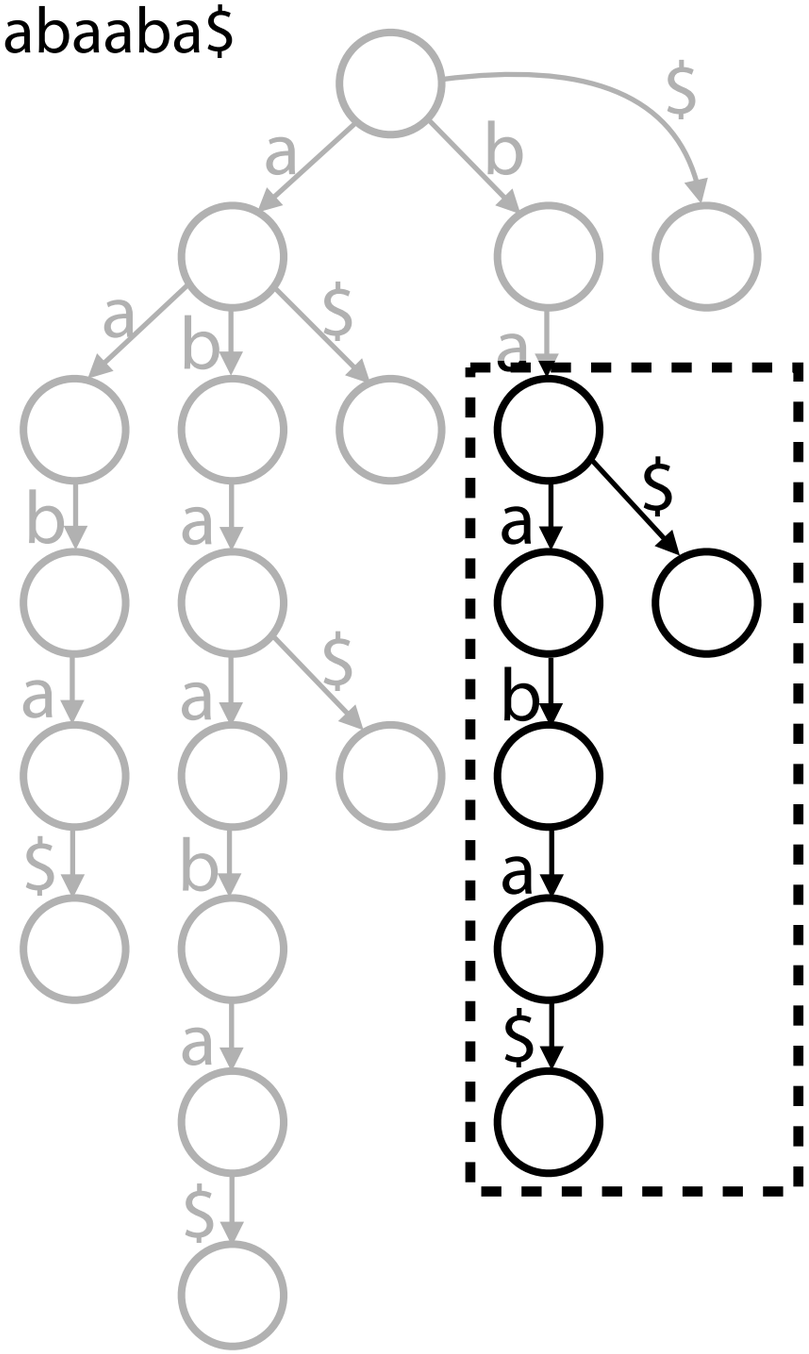
Human genome is $3 \cdot 10^9$ bases long

If $m = 3 \cdot 10^9$, m^2 is far beyond what we can store in memory



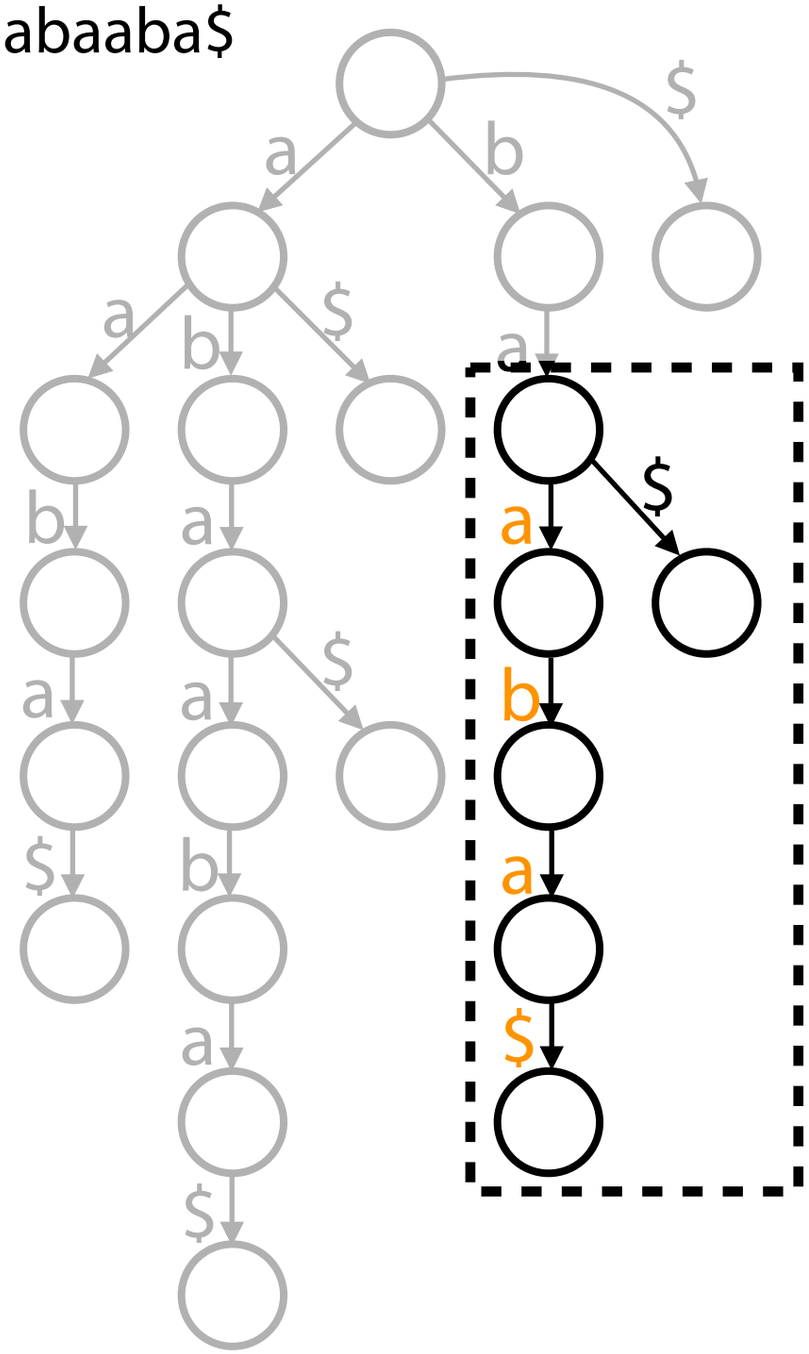
Suffix trie: making it smaller

T = abaaba$



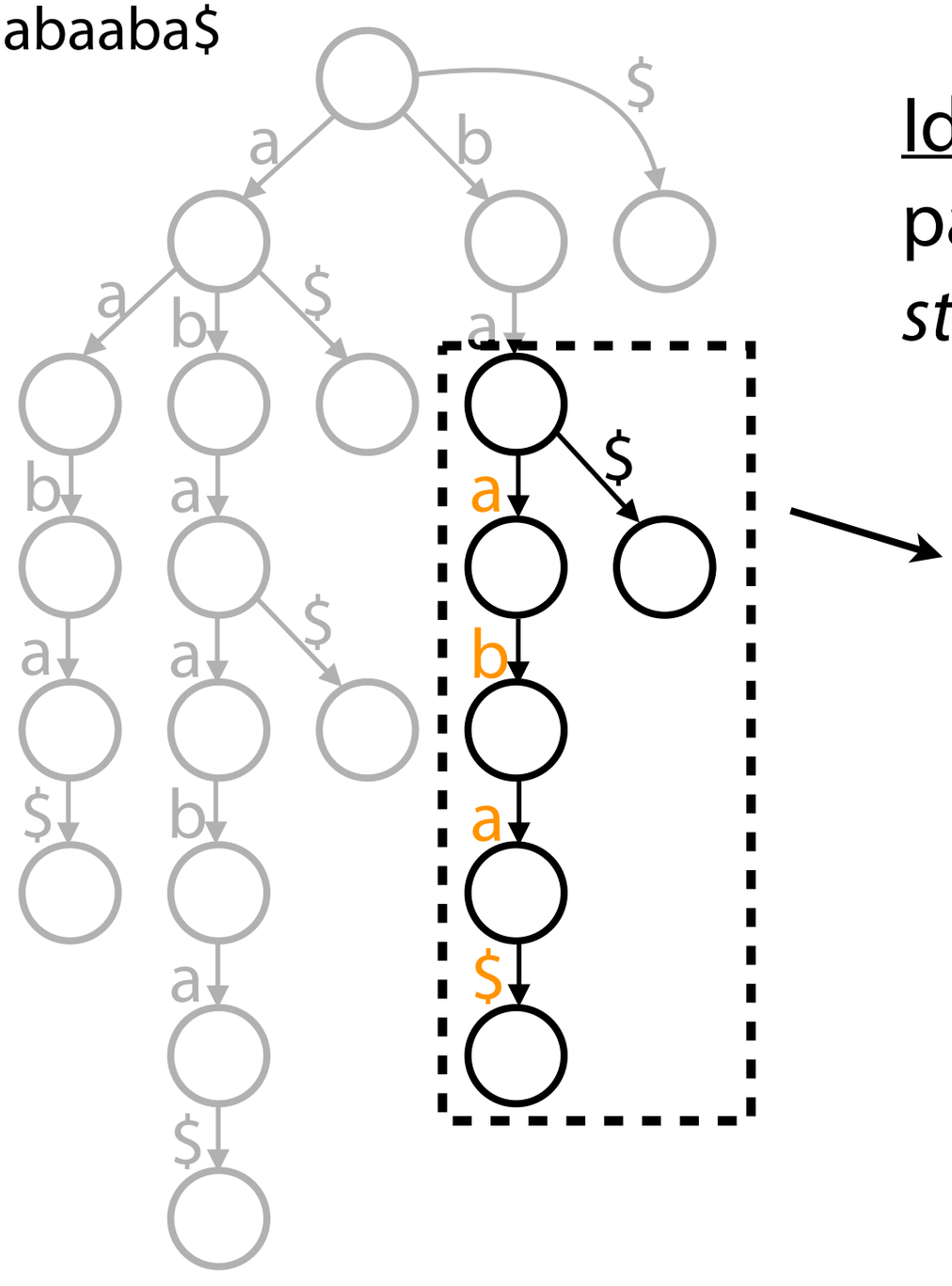
Suffix trie: making it smaller

T = abaaba$



Suffix trie: making it smaller

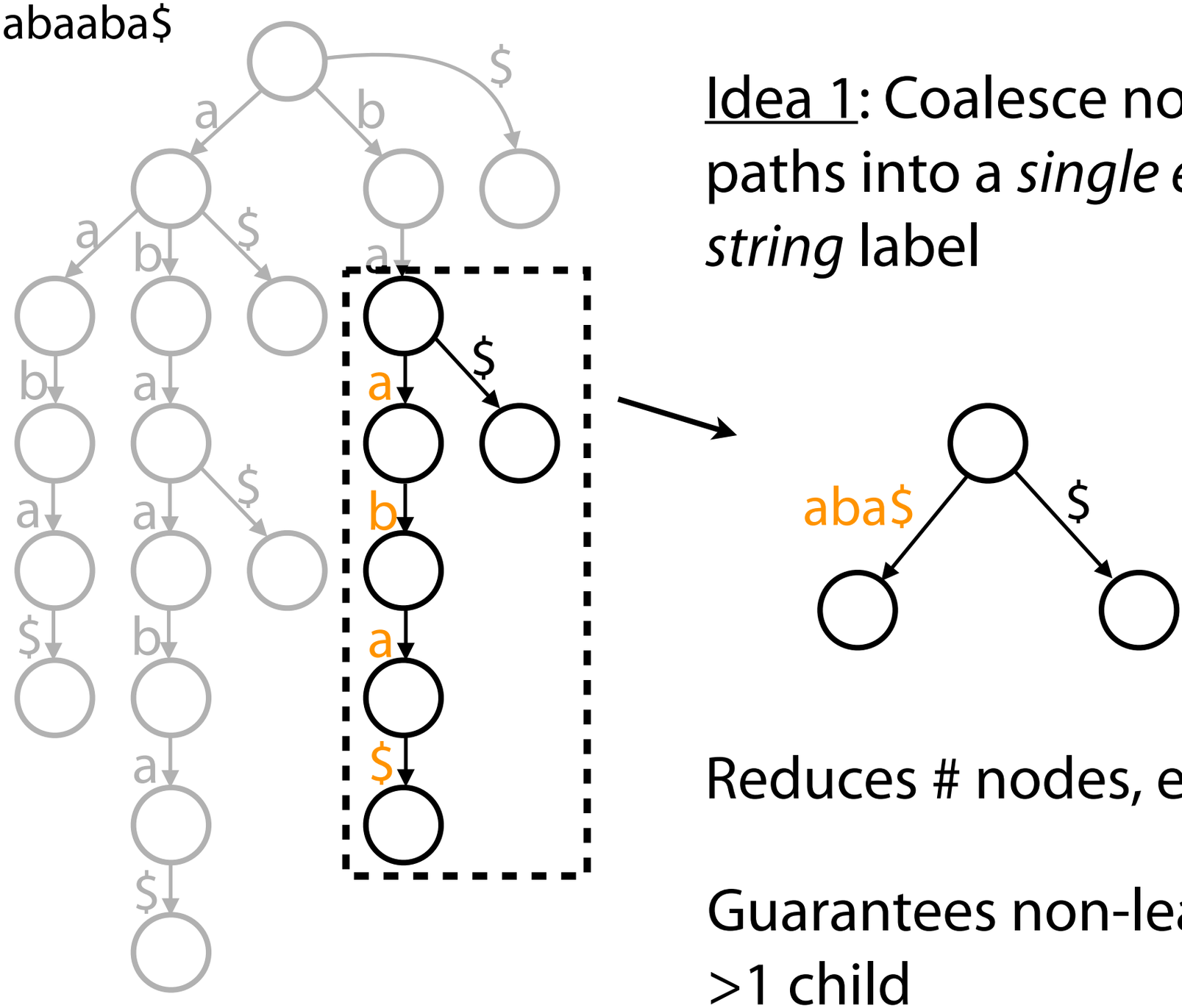
T = abaaba$



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

Suffix trie: making it smaller

T = abaaba$



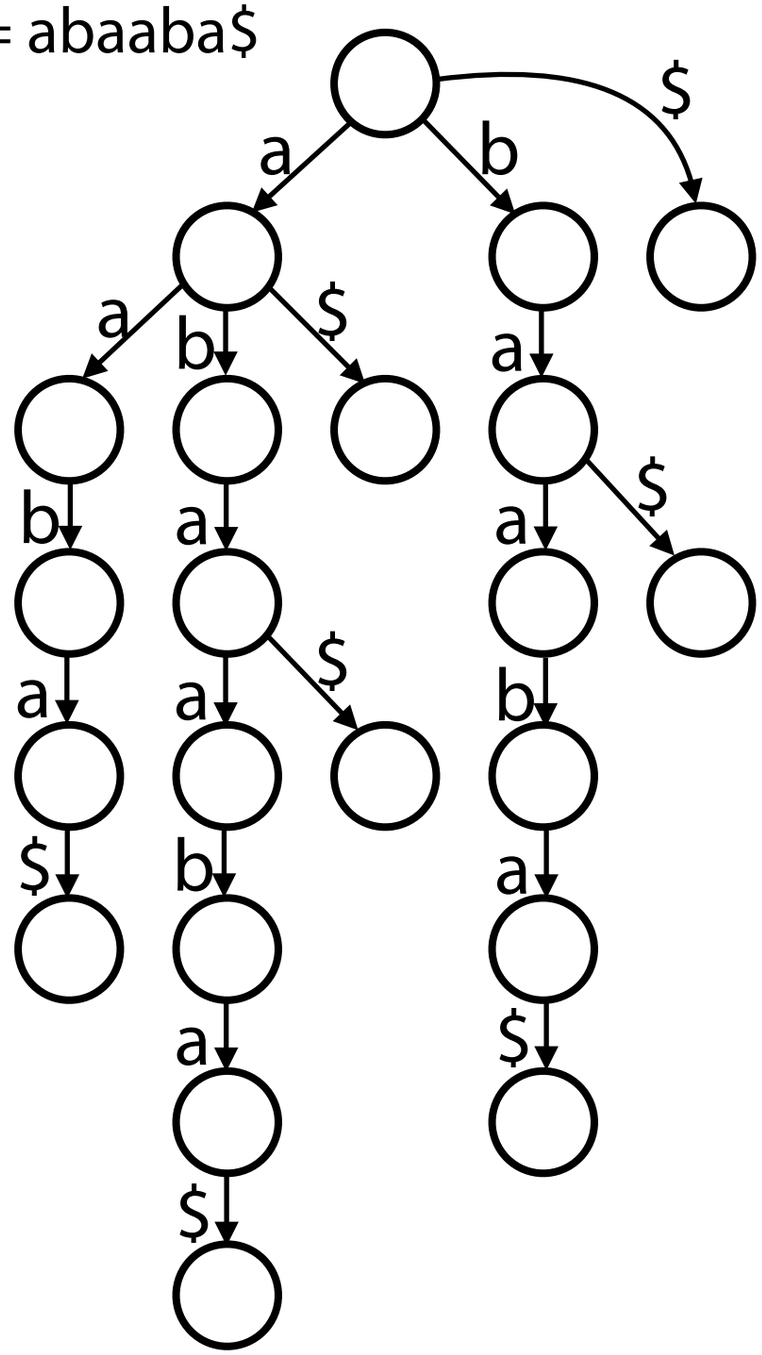
Idea 1: Coalesce non-branching paths into a *single edge* with a *string* label

Reduces # nodes, edges

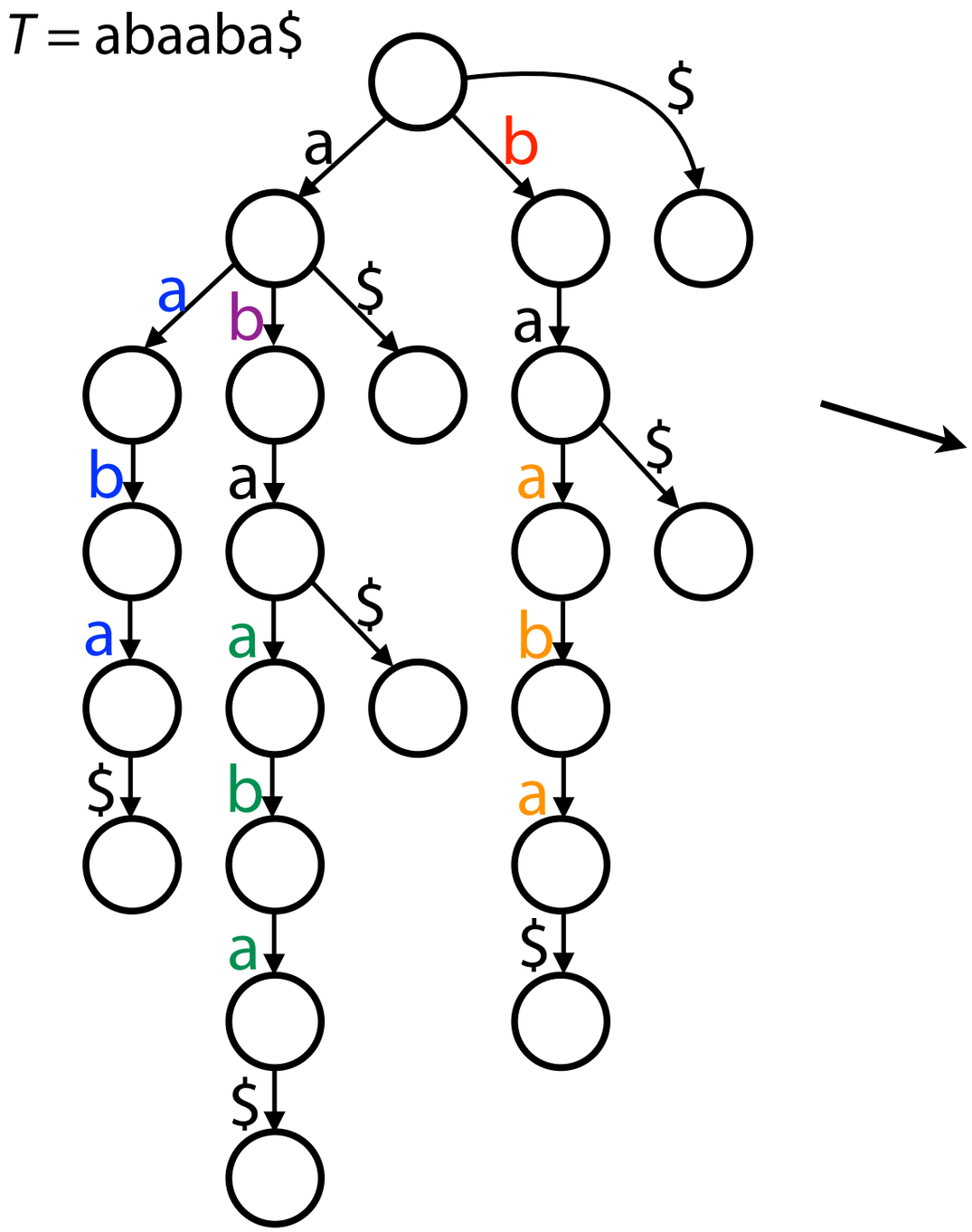
Guarantees non-leaf nodes have >1 child

Suffix trie: making it smaller

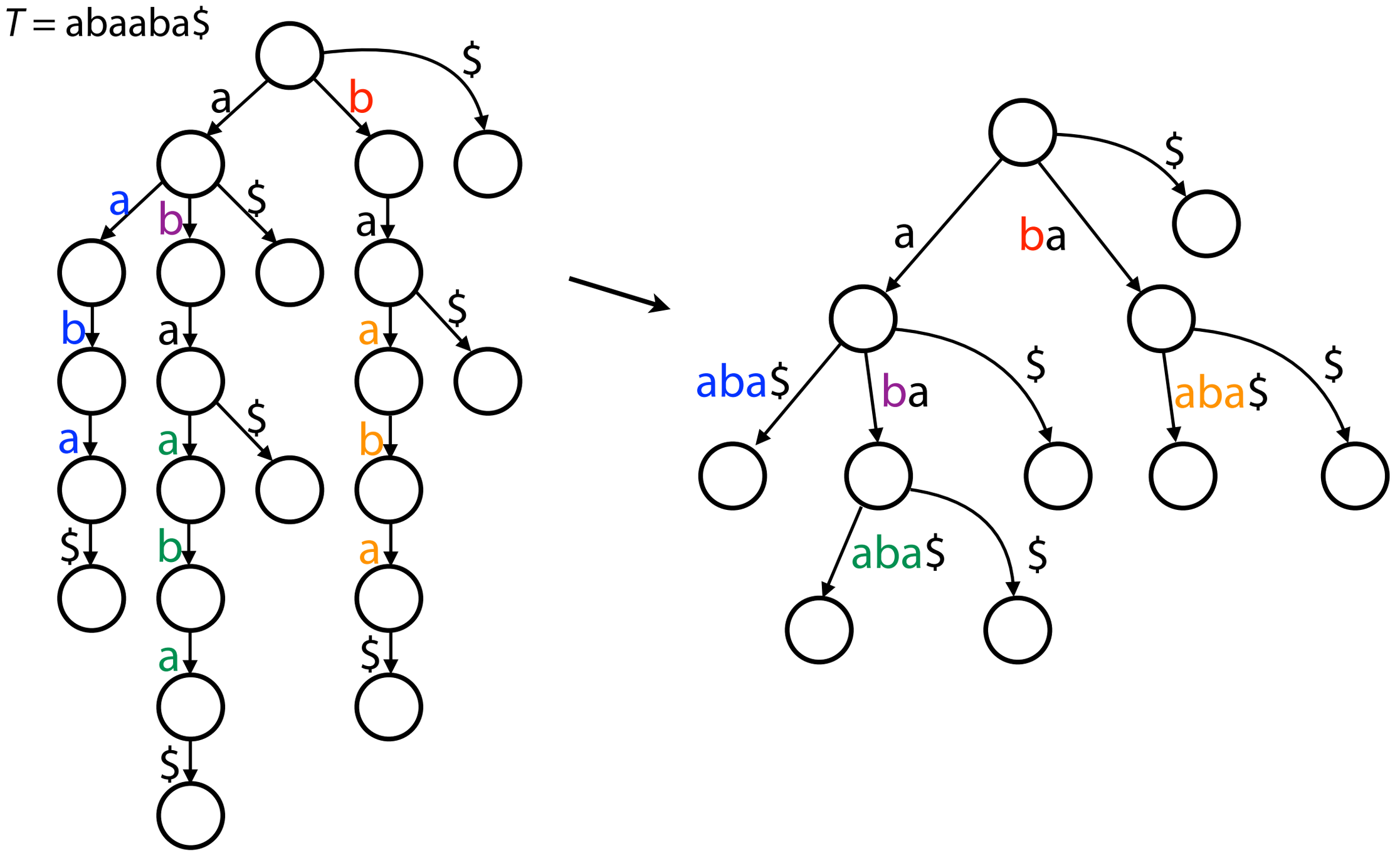
T = abaaba$



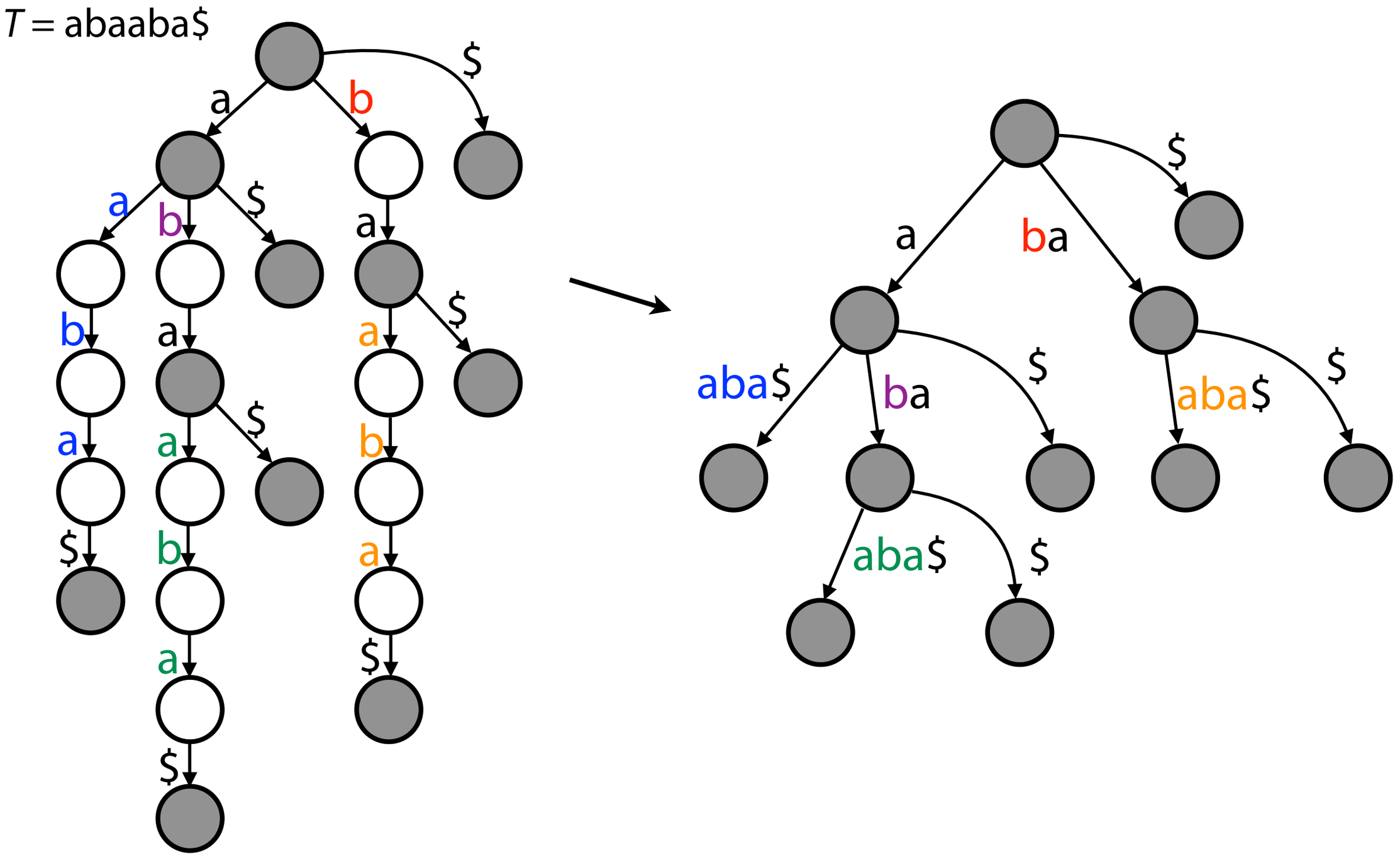
Suffix trie: making it smaller



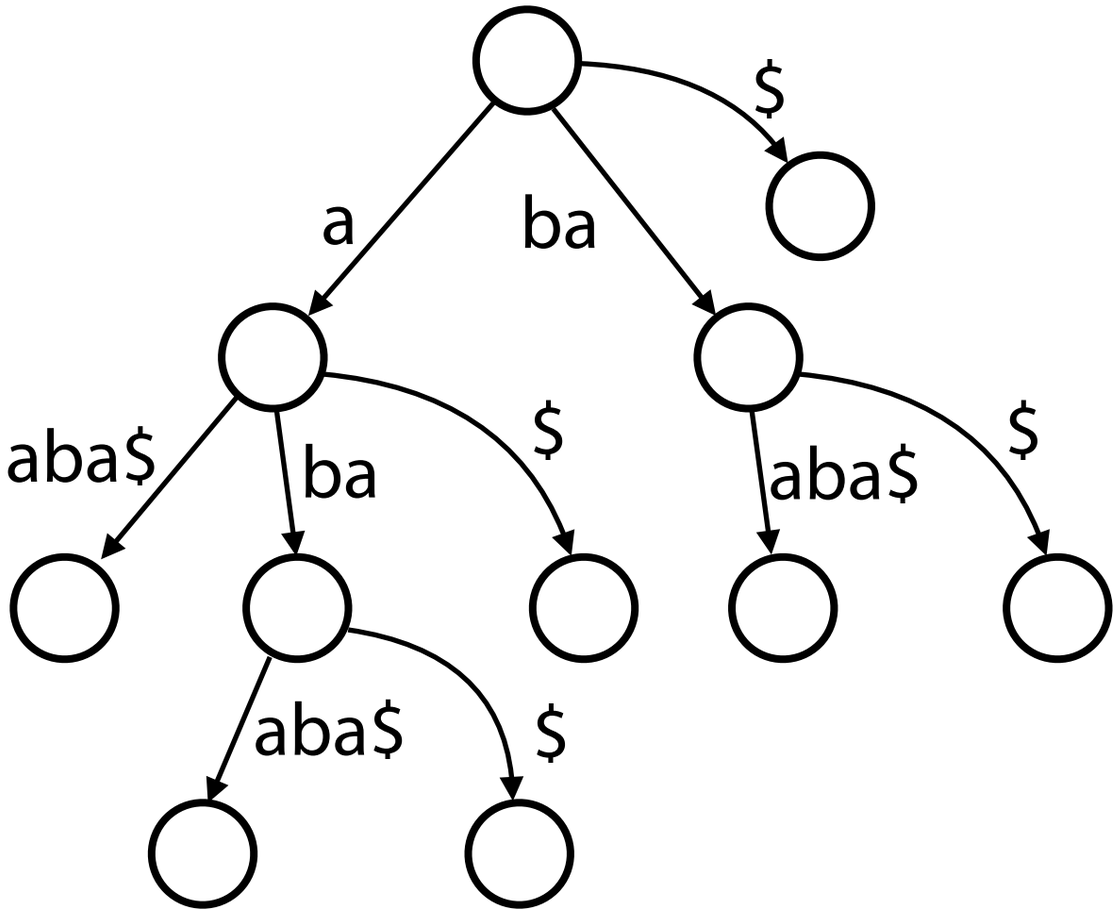
Suffix trie: making it smaller



Suffix trie: making it smaller



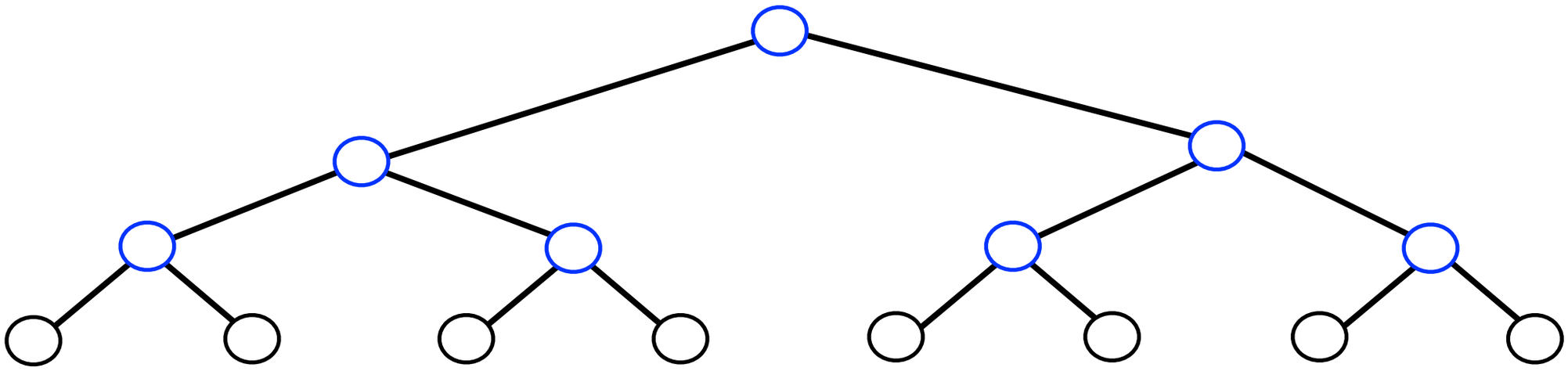
Suffix tree



When no node has an "only child," we can bound the total # nodes in terms of the # leaves, *m*

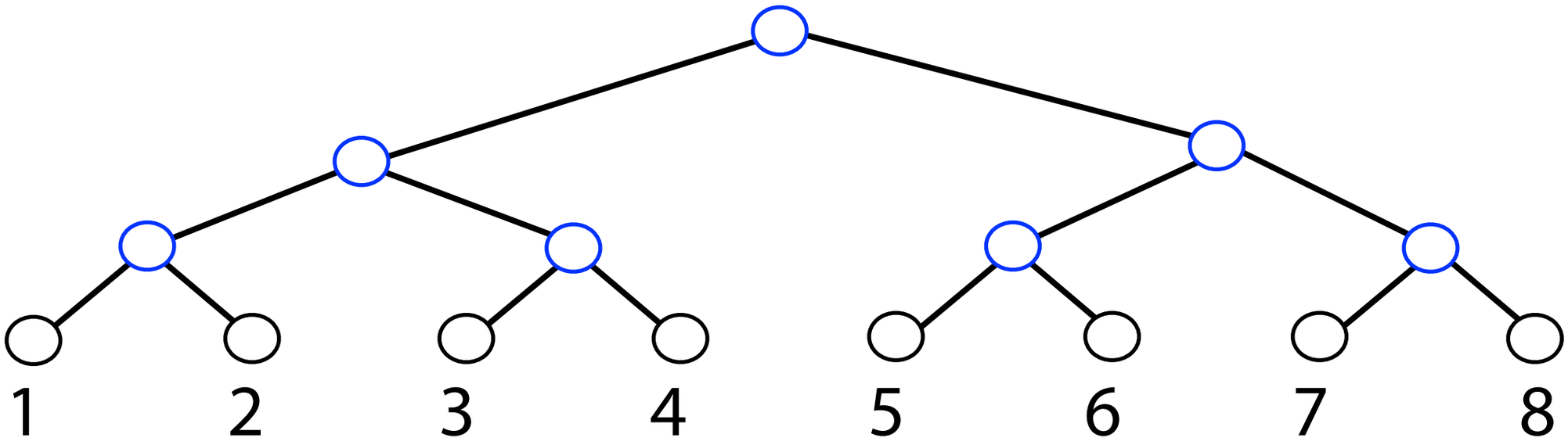
Suffix tree

In a **binary** tree, where each non-leaf has **exactly** 2 children, the # of non-leaves = m - 1



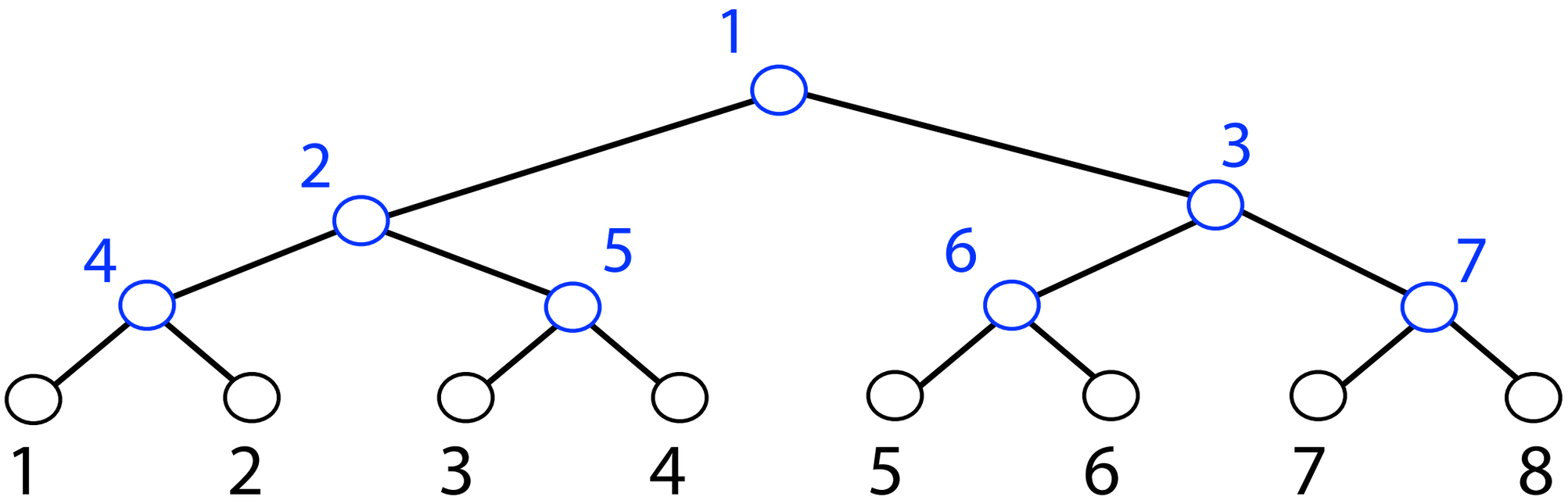
Suffix tree

In a **binary** tree, where each non-leaf has **exactly** 2 children, the # of non-leaves = m - 1



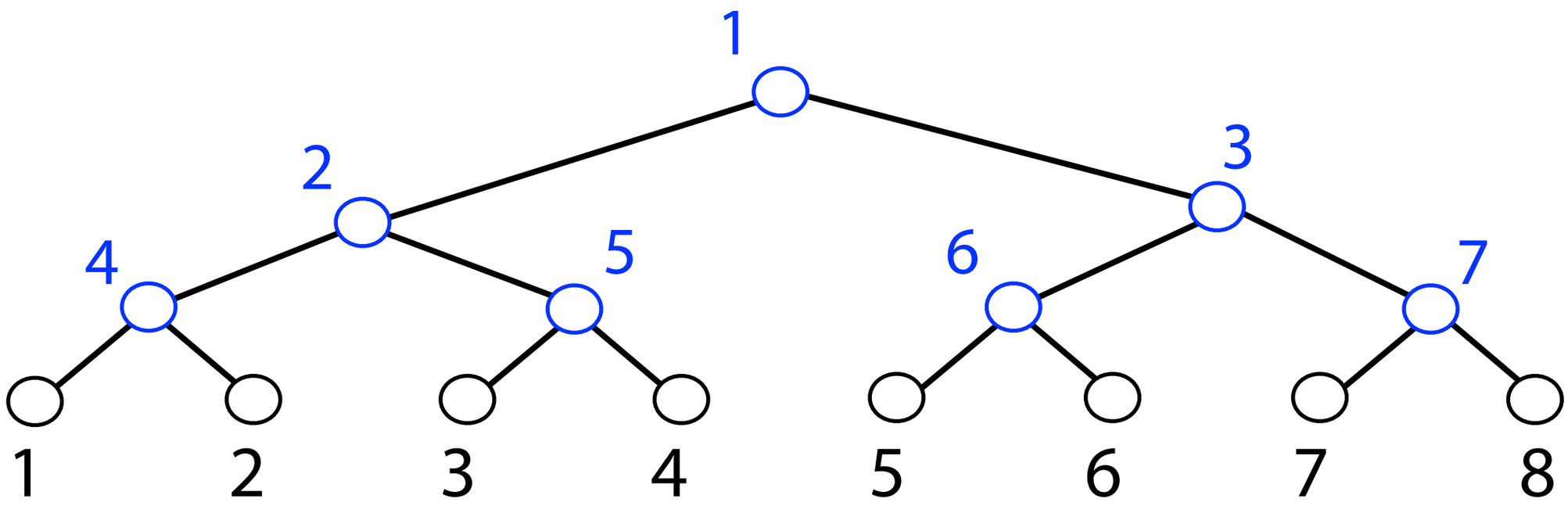
Suffix tree

In a **binary** tree, where each non-leaf has **exactly** 2 children, the # of non-leaves = m - 1



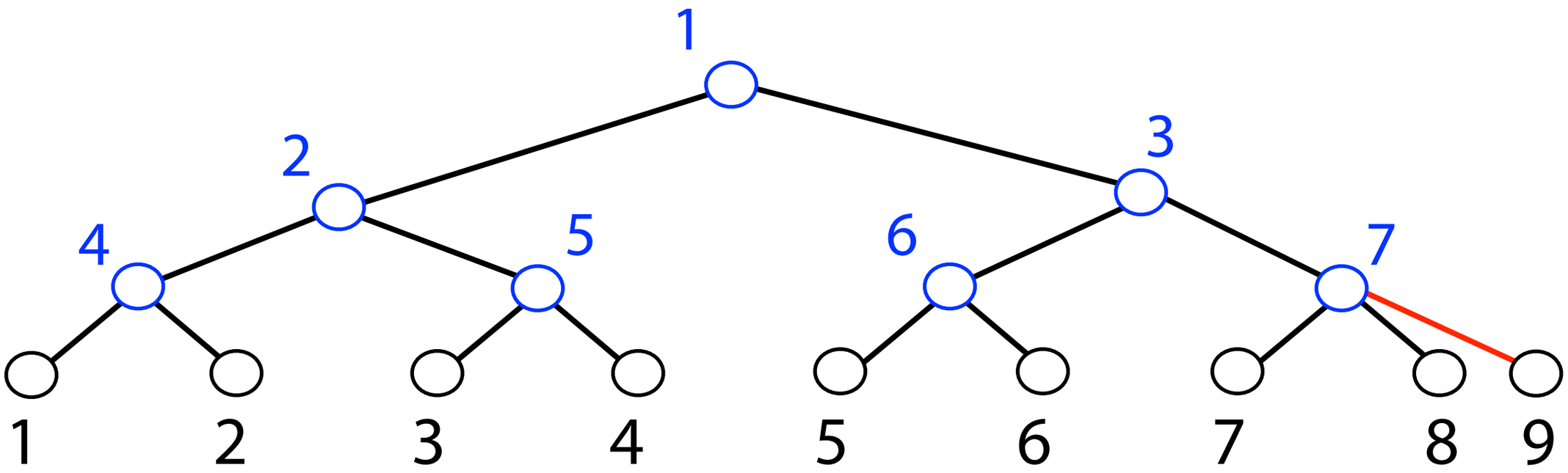
Suffix tree

If we allow non-leaves to have >2 children, then the number of leaves only increases relative to non-leaves



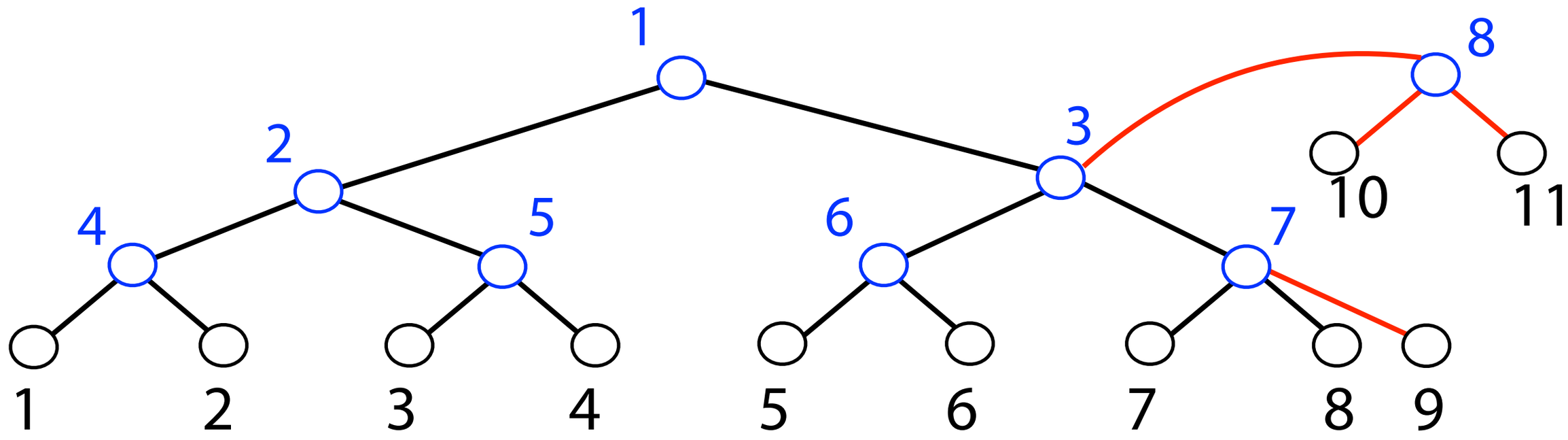
Suffix tree

If we allow non-leaves to have >2 children, then the number of leaves only increases relative to non-leaves



Suffix tree

If we allow non-leaves to have >2 children, then the number of leaves only increases relative to non-leaves

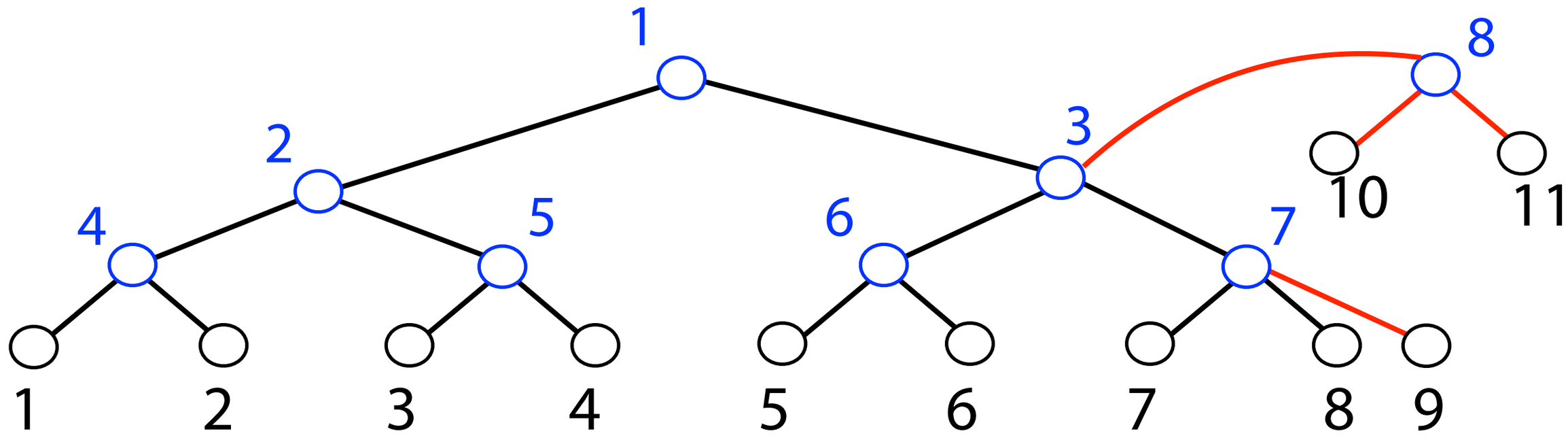


leaves *m*

non-leaf nodes

Suffix tree

If we allow non-leaves to have >2 children, then the number of leaves only increases relative to non-leaves



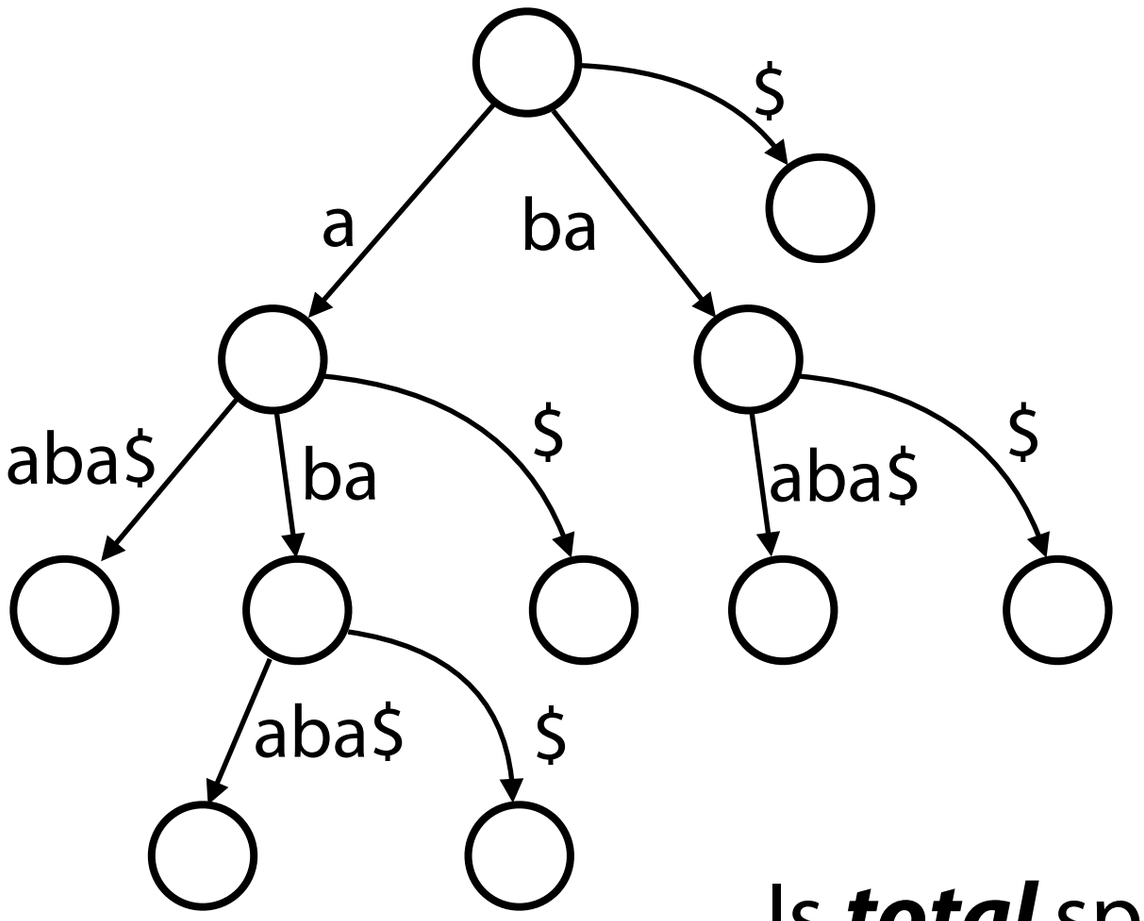
leaves m

non-leaf nodes $\leq m - 1$

Call this the "no-only-child" principle

Suffix tree

$T = \text{abaaba\$} \quad |T| = m$



No-only-child principle

leaves? ***m***

non-leaf nodes ***≤ m - 1***

≤ 2*m* -1 nodes total — $O(m)$

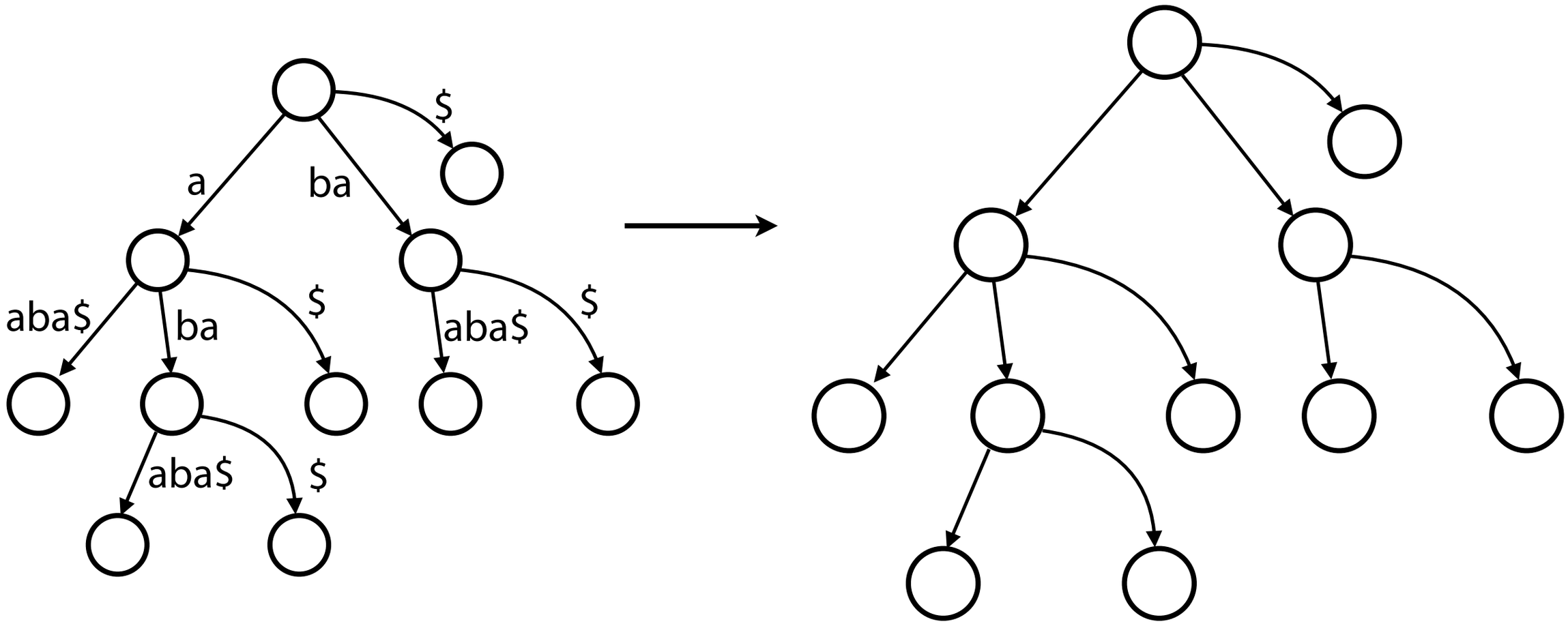
✅

Is ***total*** space *O*(*m*) now?

No: total length of **edge labels** grows with m^2

Suffix tree

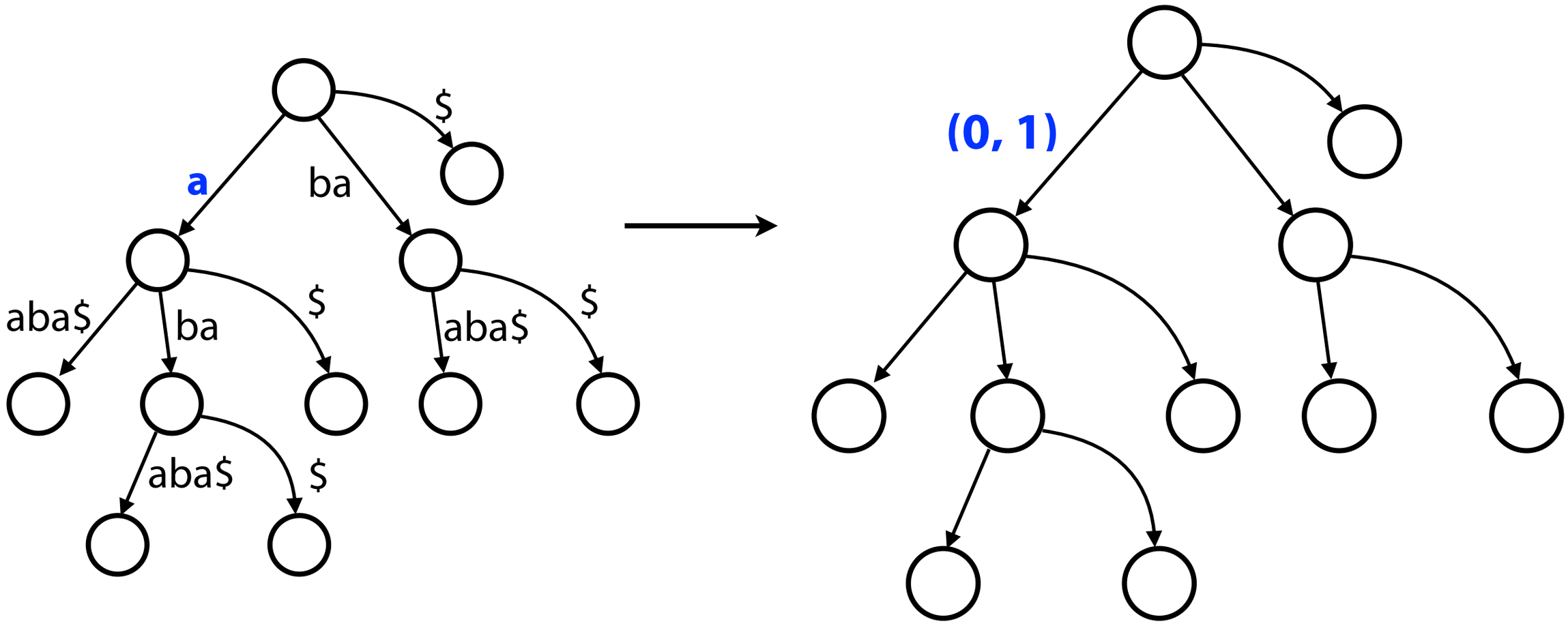
Idea 2: Store *T* itself in addition to the tree.
Convert edge labels to (offset, length) pairs with respect to *T*.



T = abaaba$

Suffix tree

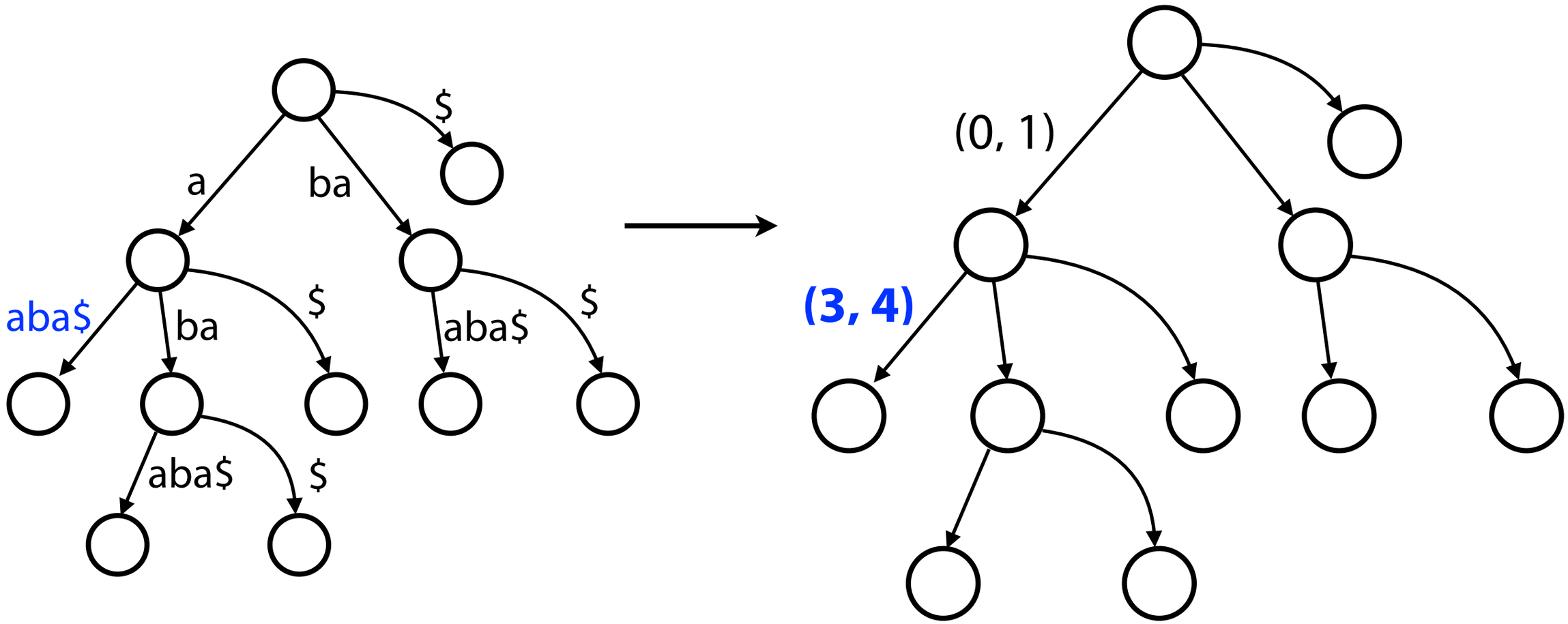
Idea 2: Store *T* itself in addition to the tree.
Convert edge labels to (offset, length) pairs with respect to *T*.



T = **a**baaba$

Suffix tree

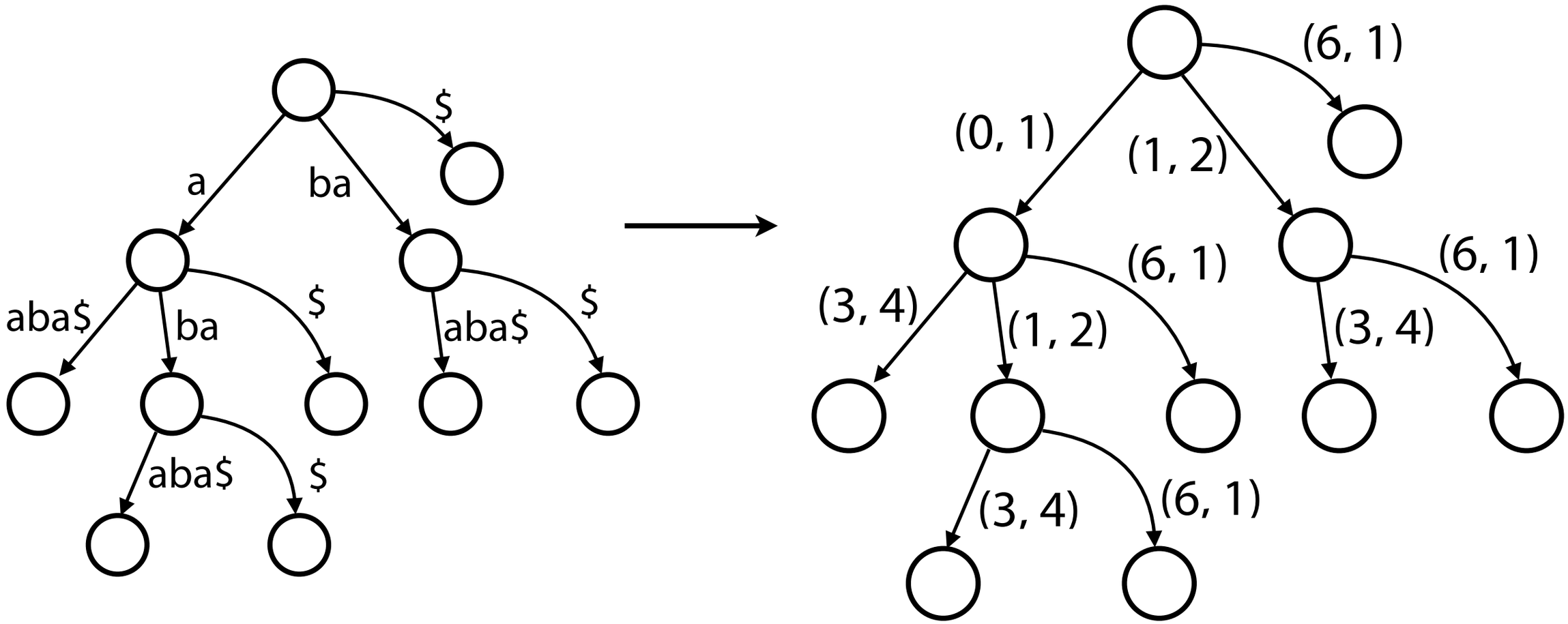
Idea 2: Store *T* itself in addition to the tree.
Convert edge labels to (offset, length) pairs with respect to *T*.



T = aba**aba$**

Suffix tree

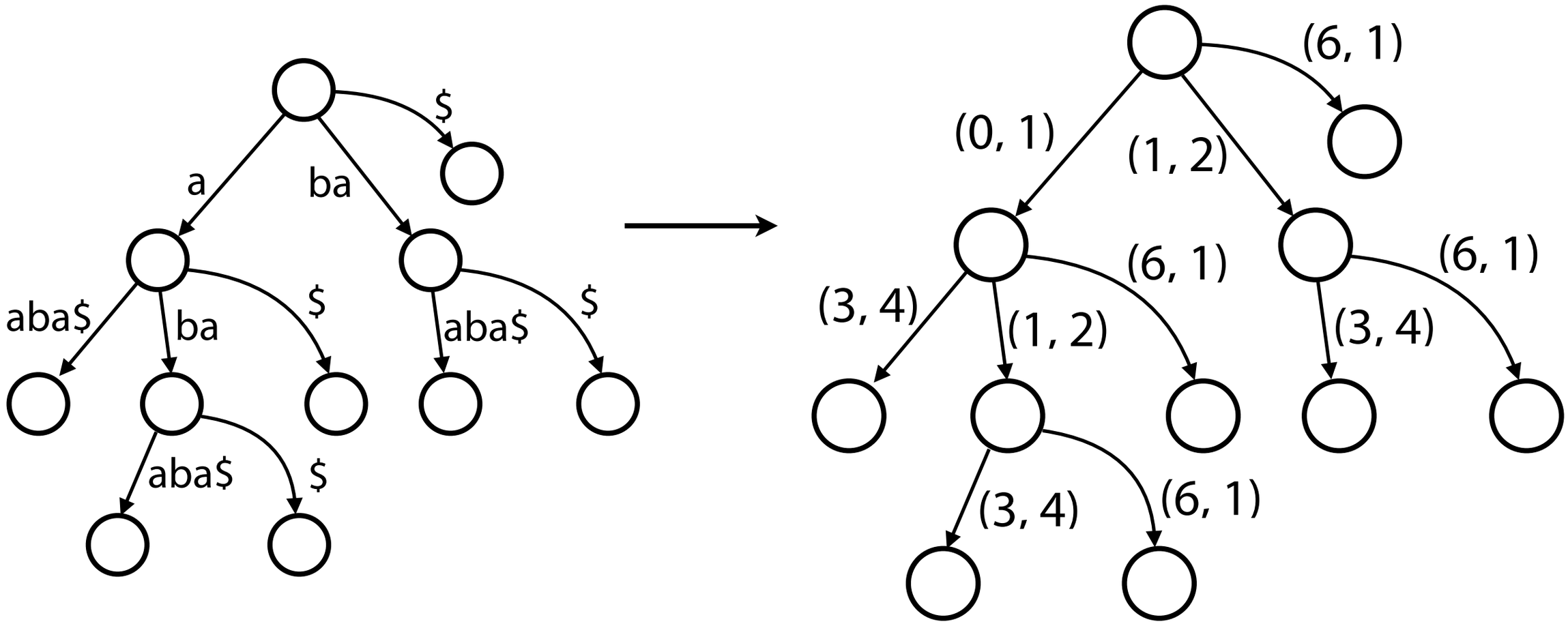
Idea 2: Store *T* itself in addition to the tree.
Convert edge labels to (offset, length) pairs with respect to *T*.



T = abaaba$

Suffix tree

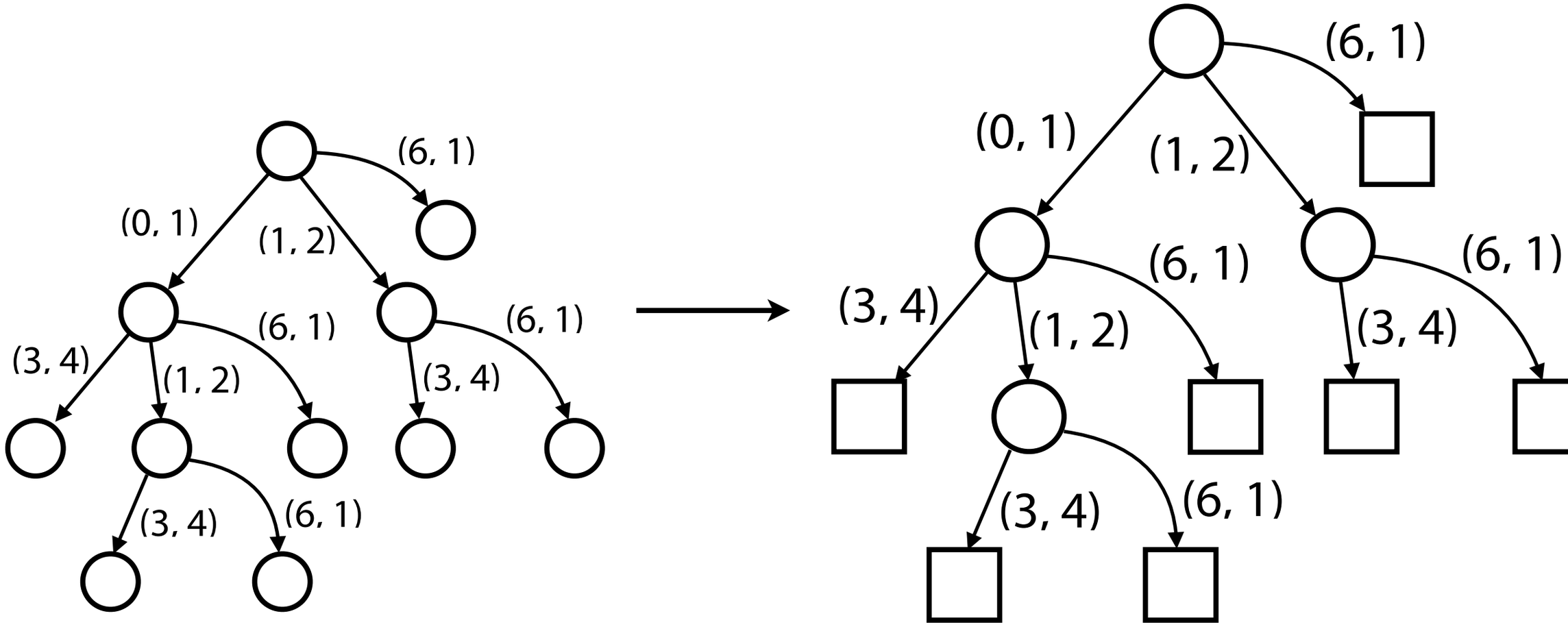
Idea 2: Store *T* itself in addition to the tree.
Convert edge labels to (offset, length) pairs with respect to *T*.



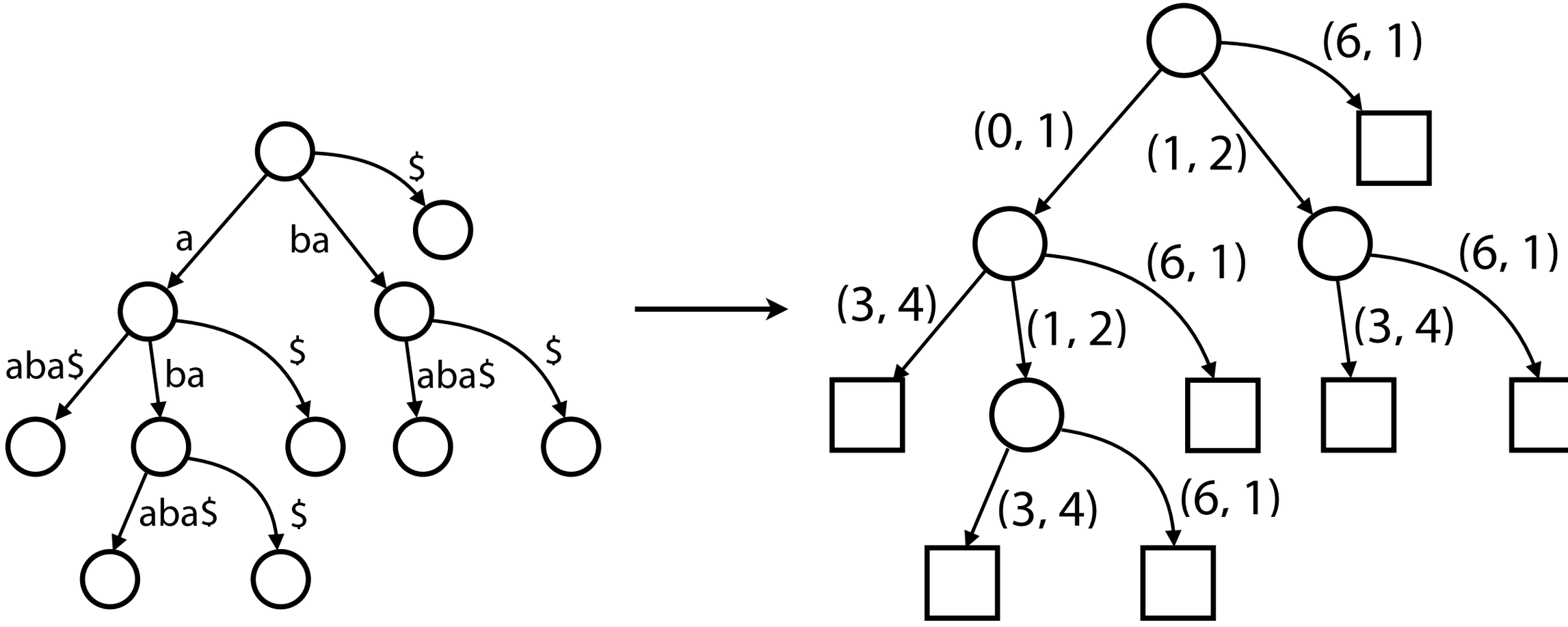
Total space is now $O(m)$ ✅
Suffix trie was $O(m^2)$!

T = abaaba$

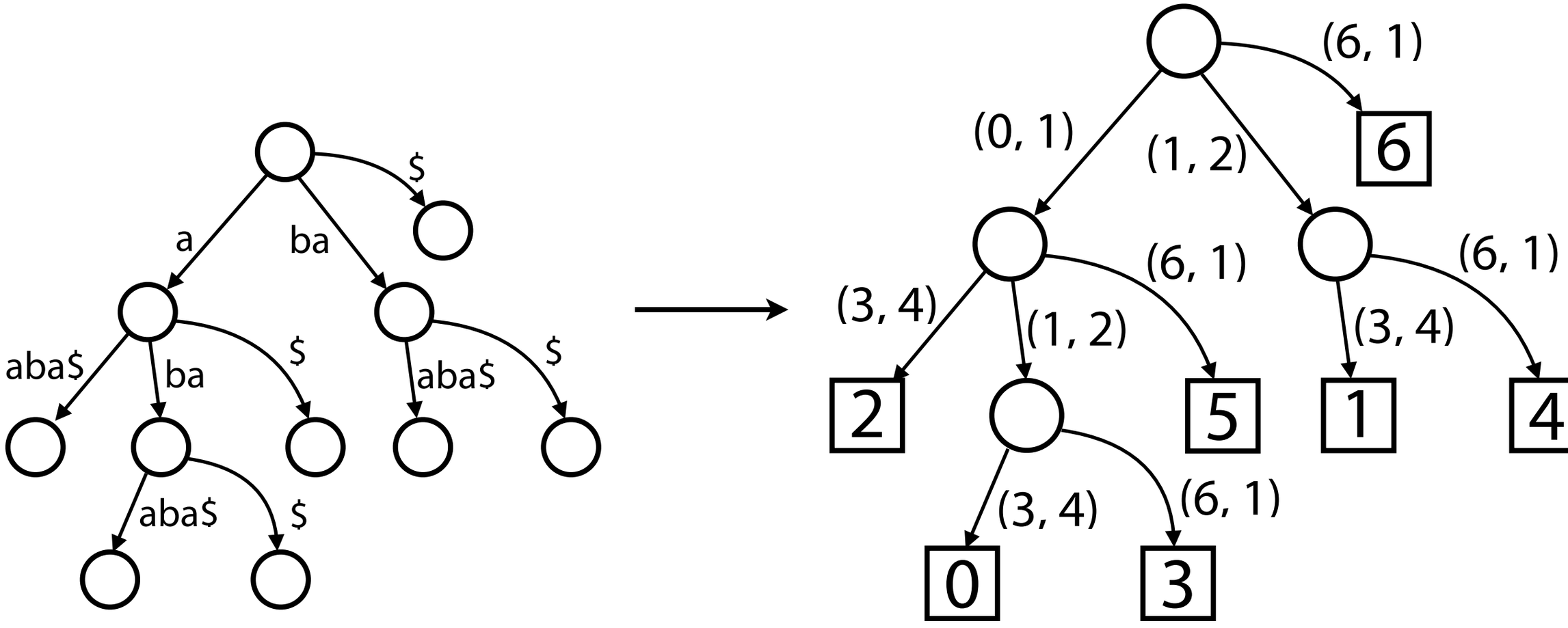
Suffix tree: leaves hold offsets



Suffix tree: leaves hold offsets

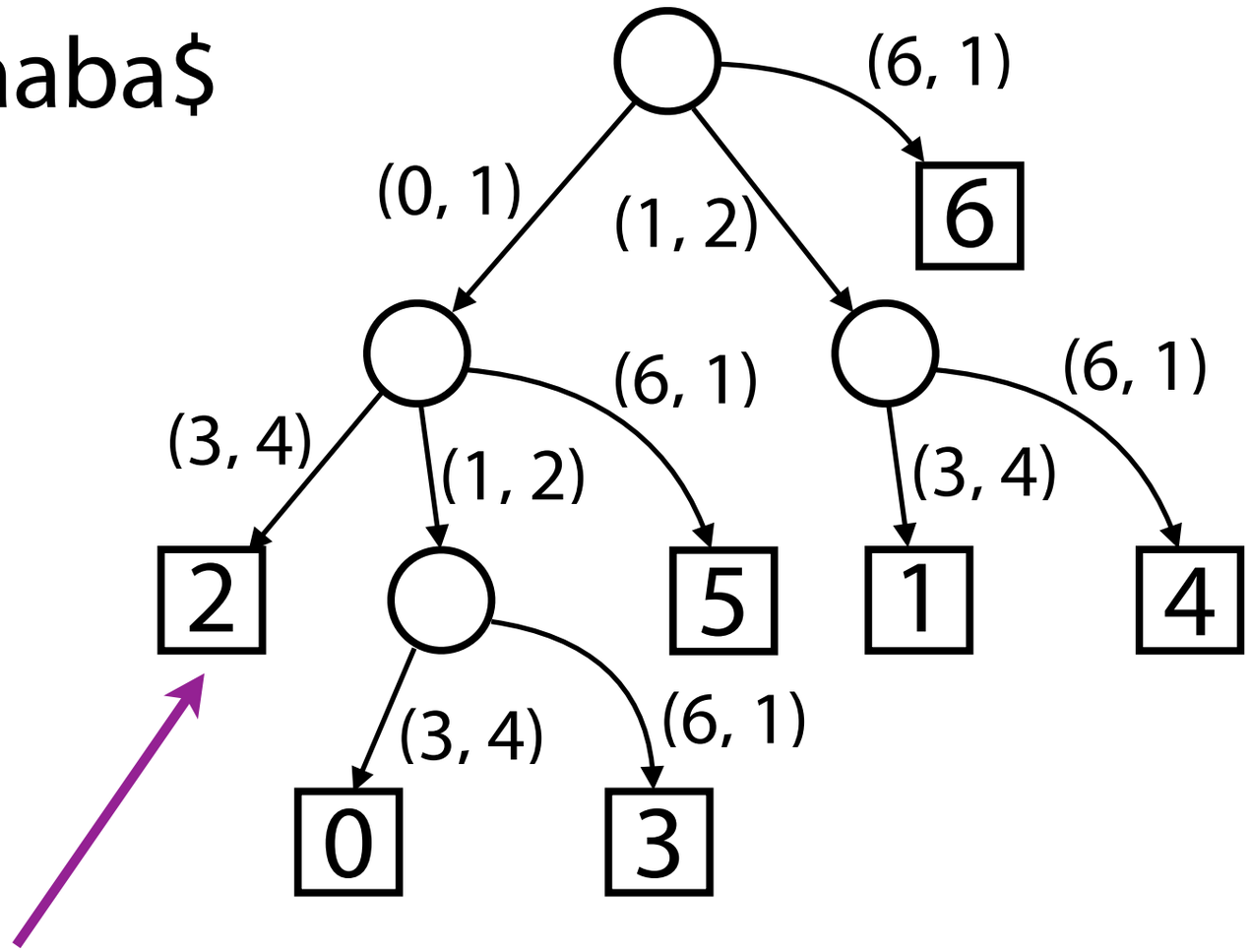


Suffix tree: leaves hold offsets

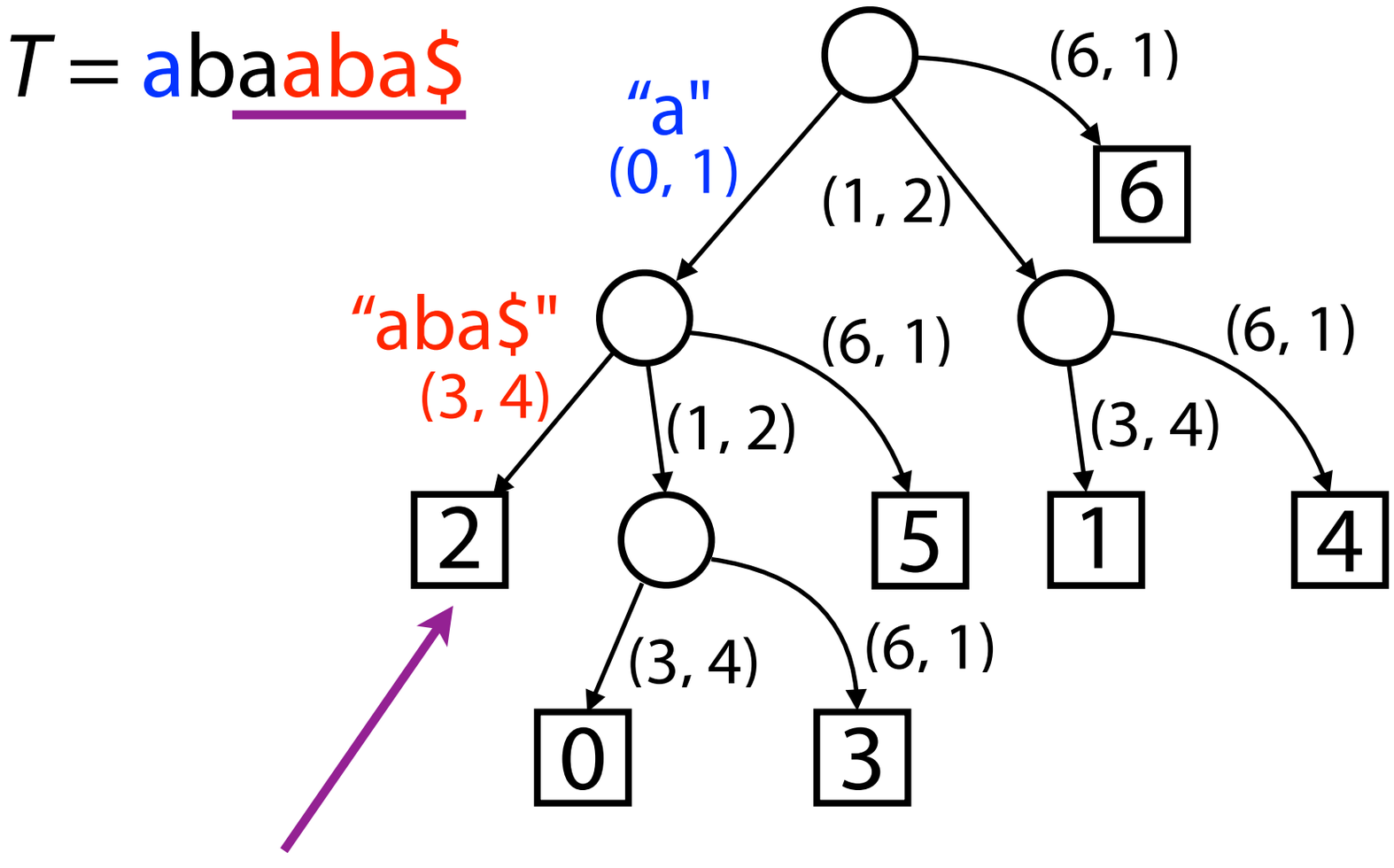


Suffix tree

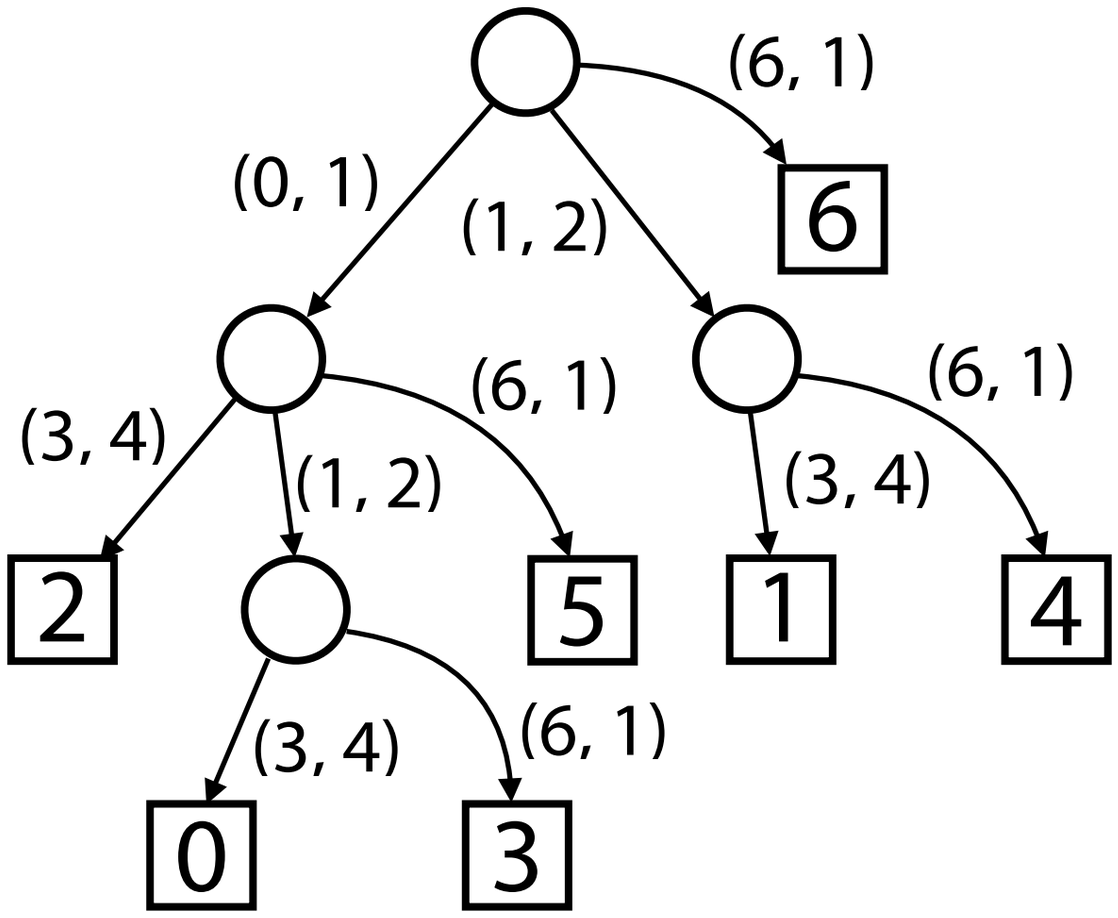
T = abaaba$



Suffix tree


T = ababaa$
"a"
(0, 1)
(6, 1)
(1, 2)
6
"aba$"
(3, 4)
(6, 1)
(6, 1)
(1, 2)
(3, 4)
2
5
1
4
(3, 4)
(6, 1)
0
3
Label = "aaba$", the suffix at offset 2

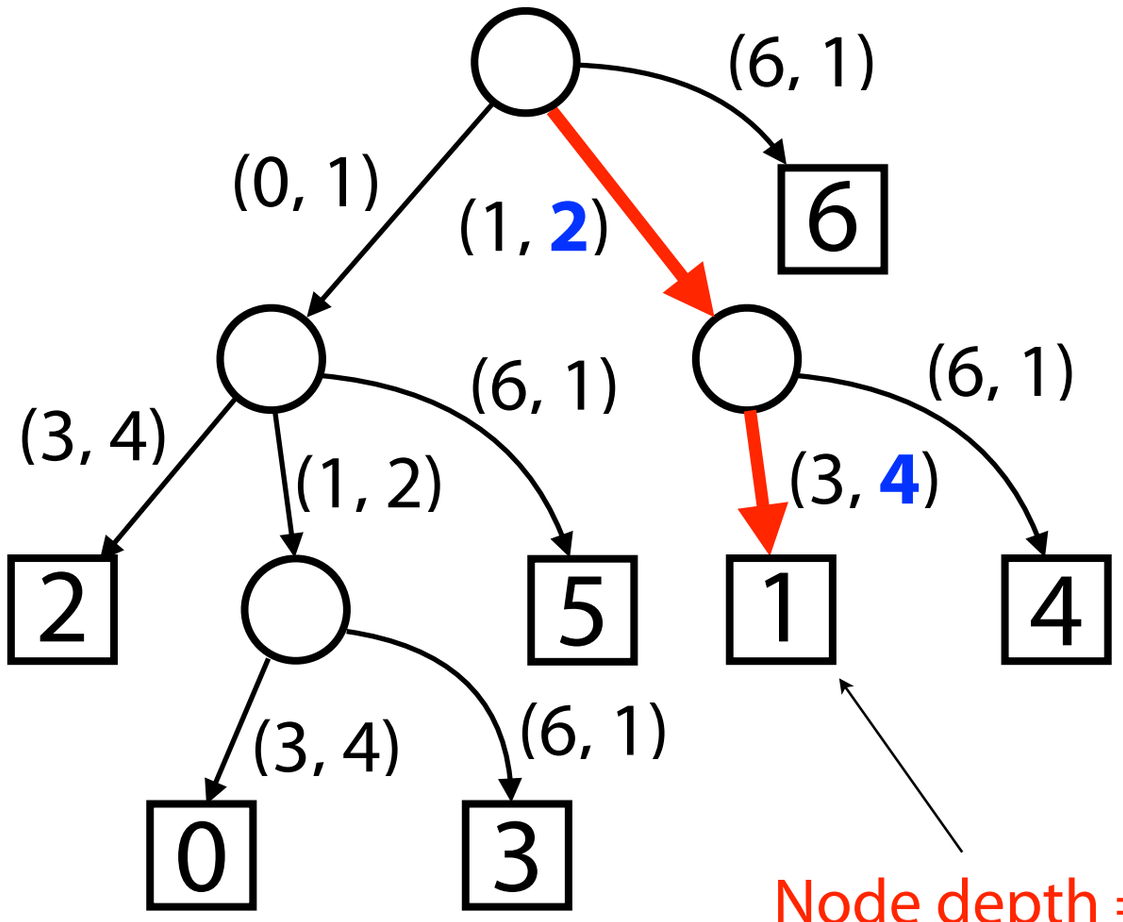
Suffix tree: depth & labels



Two notions of depth:

- **Node** depth: # edges from root to node
- **Label** depth: total length of edge labels from root to node

Suffix tree: depth & labels

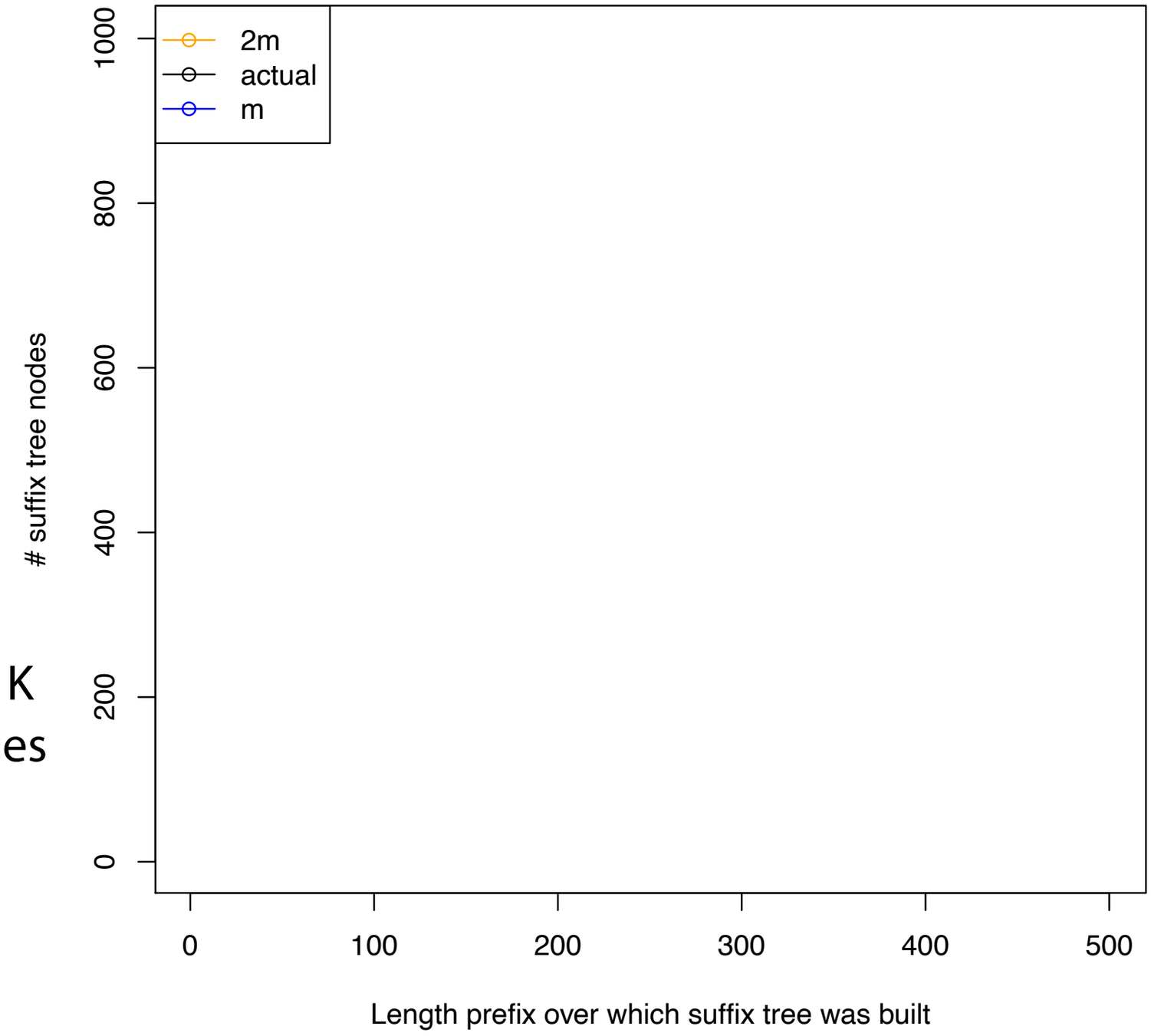


Two notions of depth:

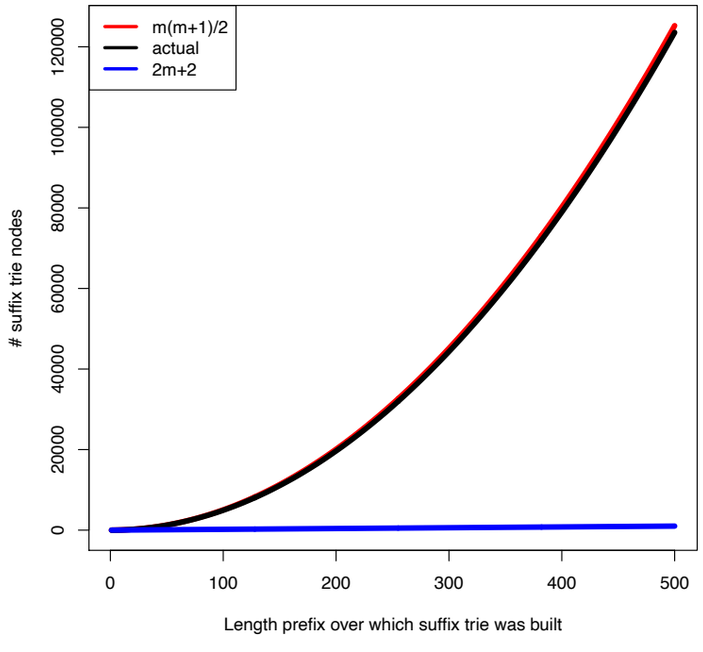
- **Node** depth: # edges from root to node
- **Label** depth: total length of edge labels from root to node

Suffix tree: actual growth

Suffix trees built for first 500 prefixes of the lambda phage virus genome



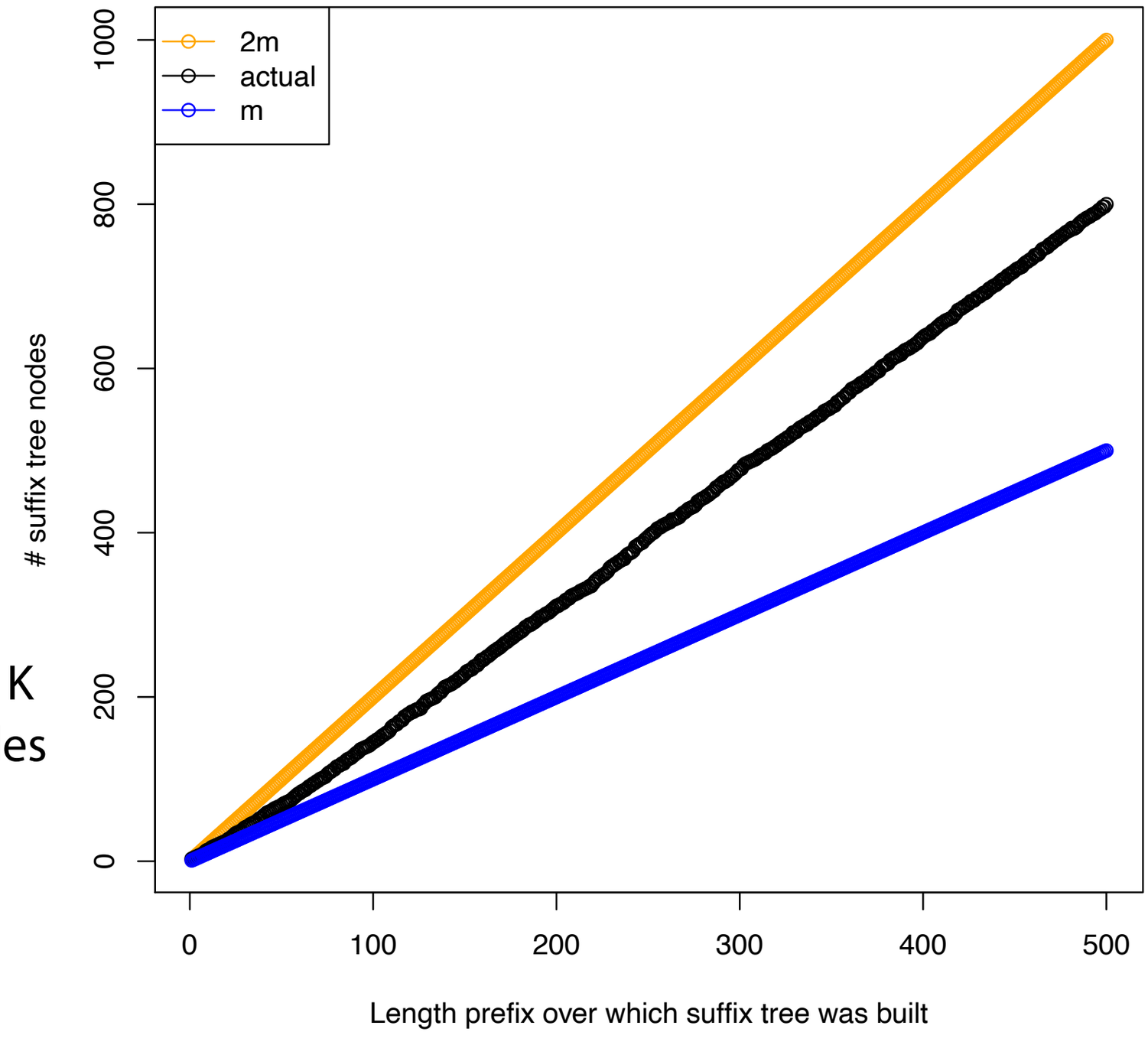
Remember suffix trie plot:



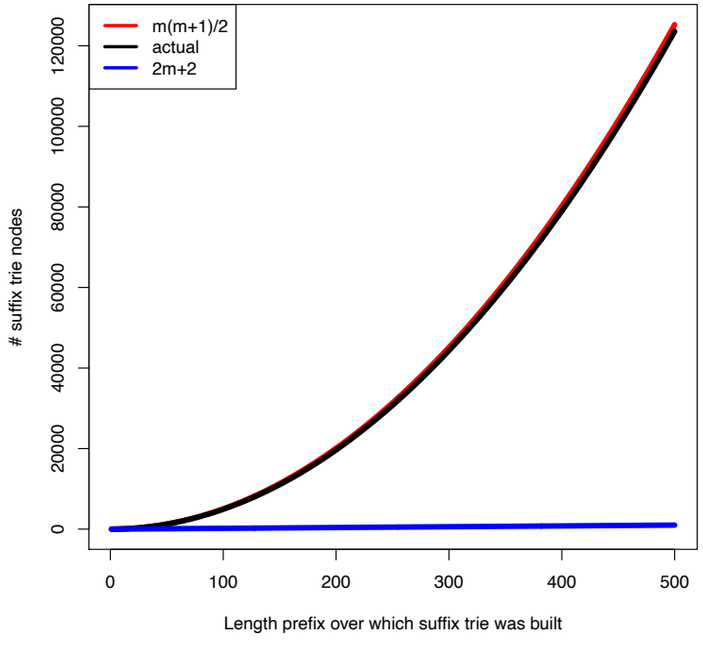
123 K nodes

Suffix tree: actual growth

Suffix trees built for first 500 prefixes of the lambda phage virus genome



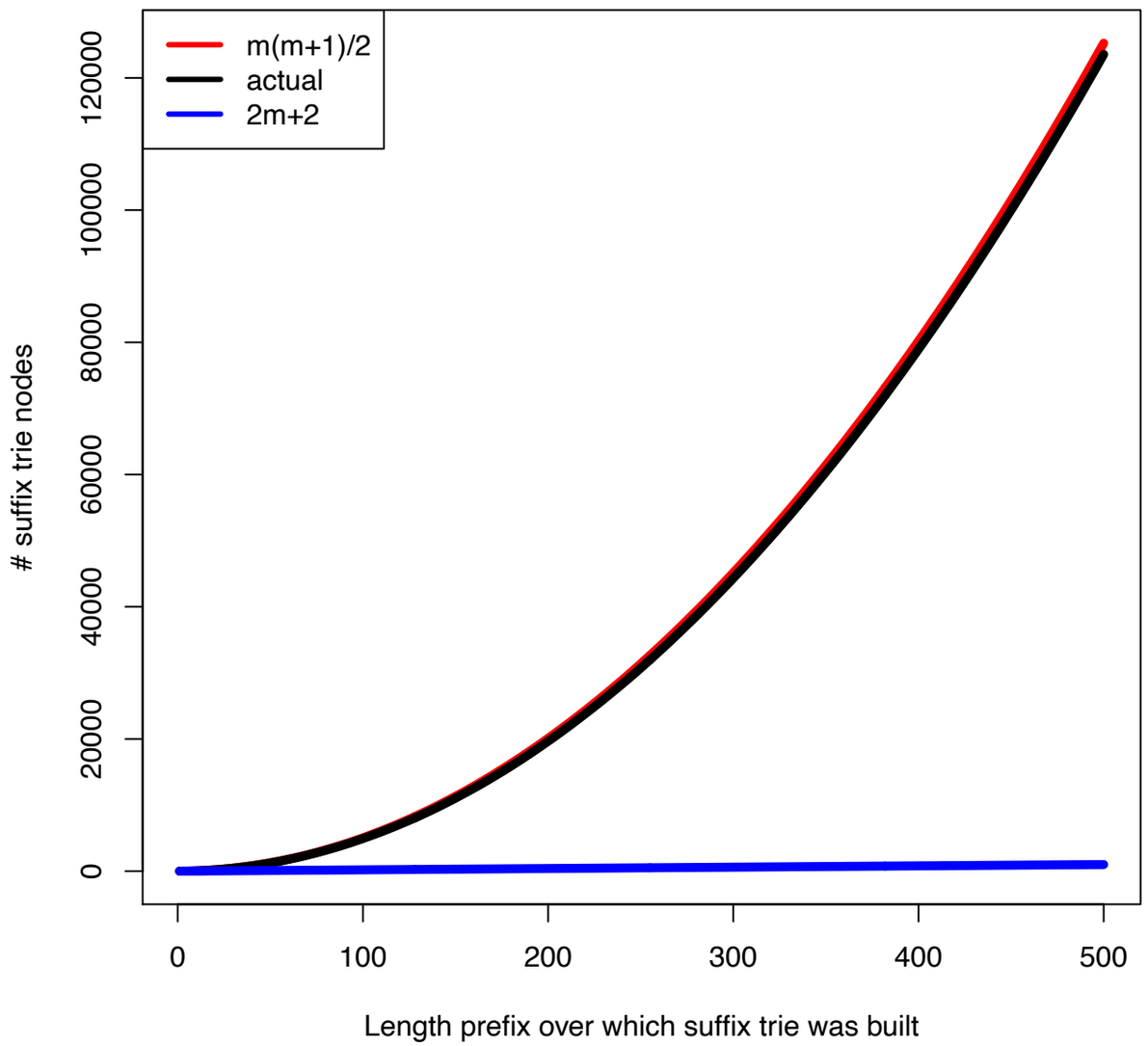
Remember suffix trie plot:



123 K nodes

Suffix trie

>100K nodes



Suffix tree

<1K nodes