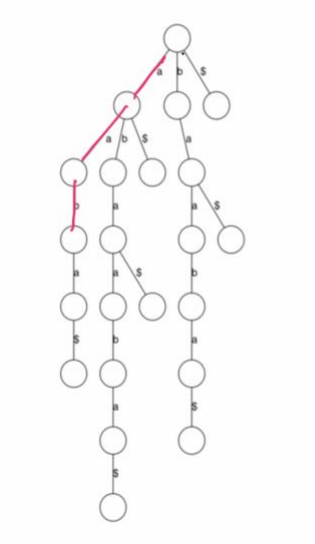
Suffix Trees

Wednesday, 13 September 2023 8:05 PM



aba aba $

aab

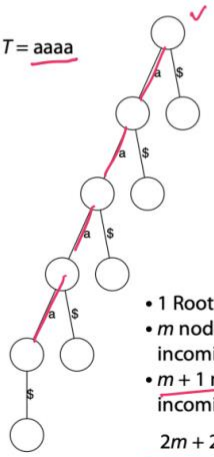
Suffix trie

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

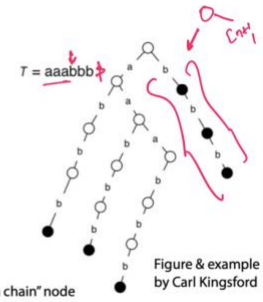
$2m + 2$ nodes

Suffix trie

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

Yes: $a^n b^n$

T = aaabbb



- 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)

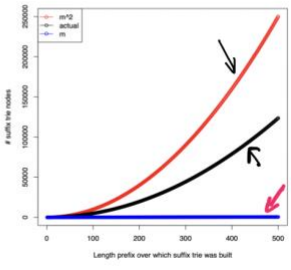
$n^2 + 4n + 2$ nodes, where $m = 2n$

Figure & example by Carl Kingsford

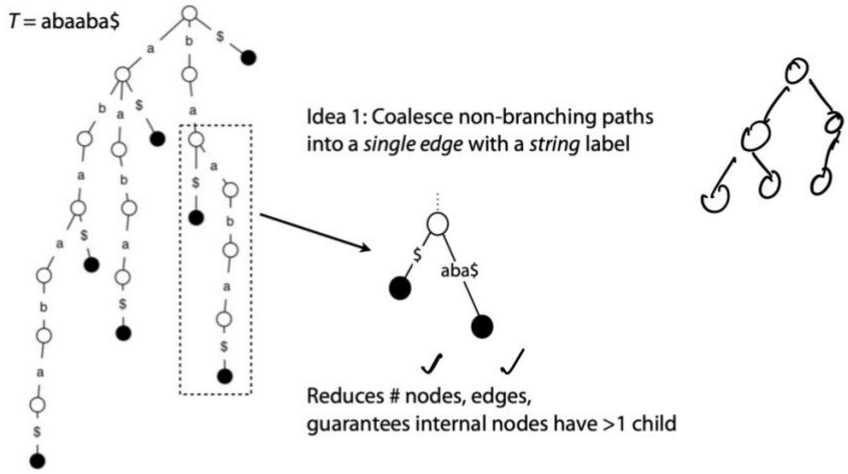
Suffix trie: actual growth

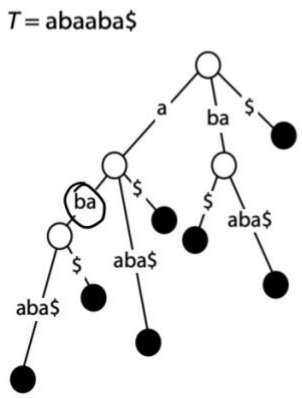
Build suffix tries for the first 500 prefixes of the lambda phage virus genome

Black curve shows how # nodes increases with prefix length



Suffix trie: making it smaller





a ba a

With respect to m:

How many leaves? m ✓

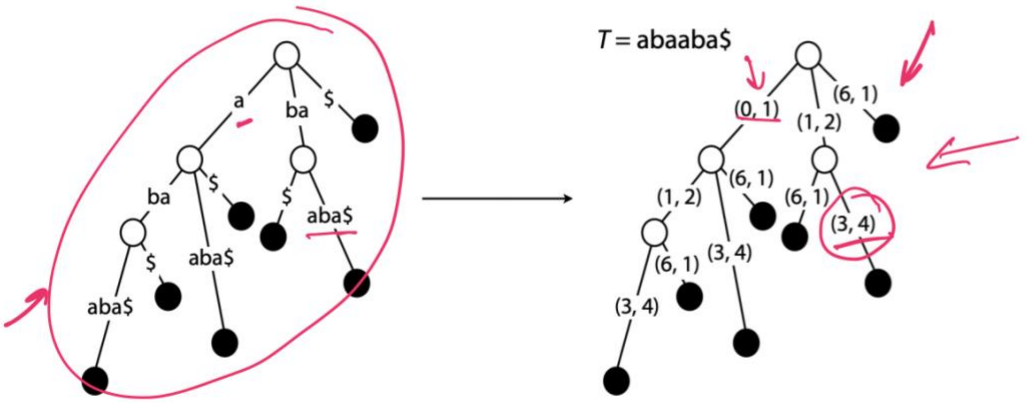
How many non-leaf nodes? $\leq m - 1$ ✓

$\leq 2m - 1$ nodes total, or $O(m)$ nodes

Suffix tree

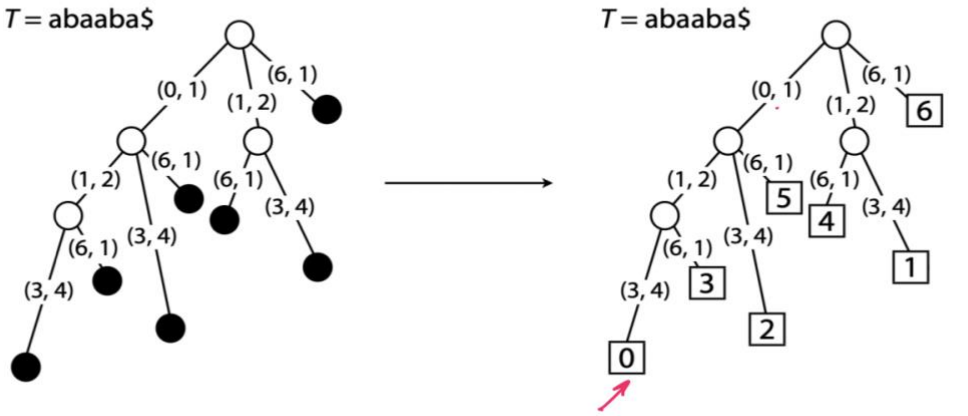
T = abaaba\$

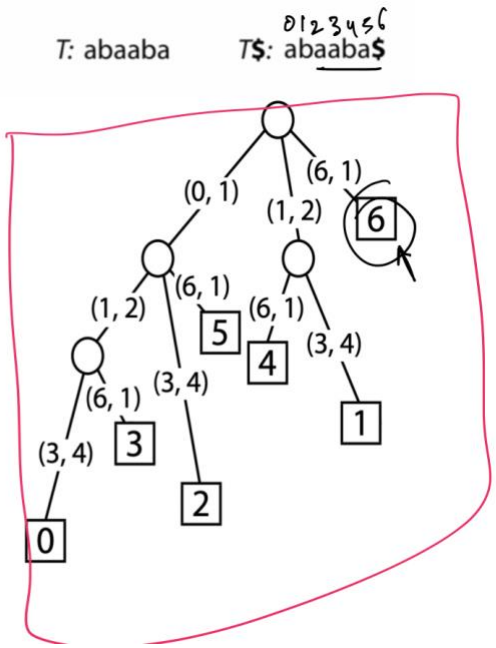
Idea 2: Store T itself in addition to the tree. Convert tree's edge labels to (offset, length) pairs with respect to T.

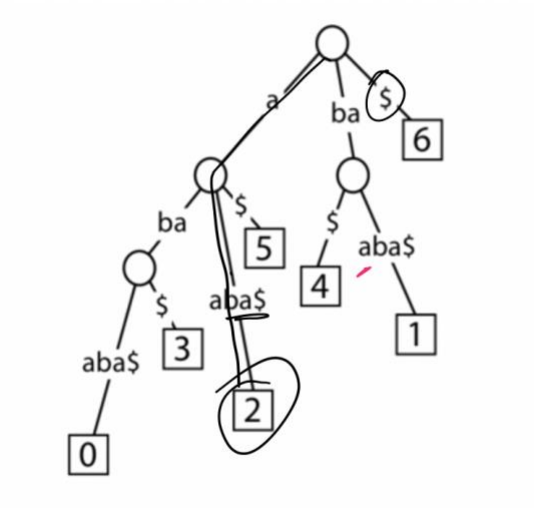


Space required for suffix tree is now $O(m)$

Suffix tree: leaves hold offsets



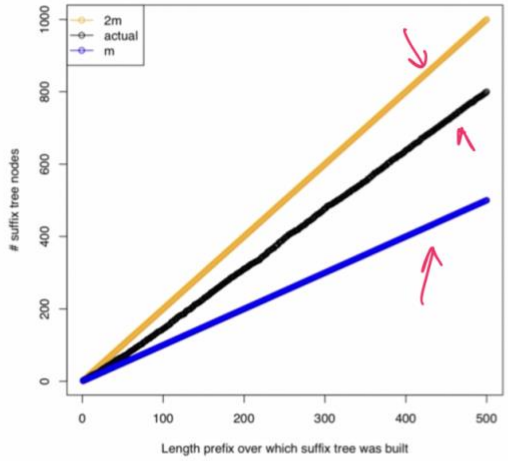

T: abaaba
T$: abaaba$
0123456
(0, 1)
(6, 1)
(1, 2)
6
(6, 1)
(1, 2)
5
(6, 1)
4
(3, 4)
(3, 4)
(6, 1)
3
1
(3, 4)
2
0

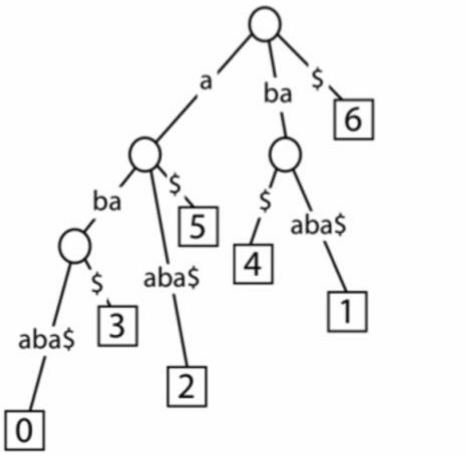

a
ba
$
6
ba
$
5
$
aba$
aba$
$
4
1
3
aba$
2
0

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

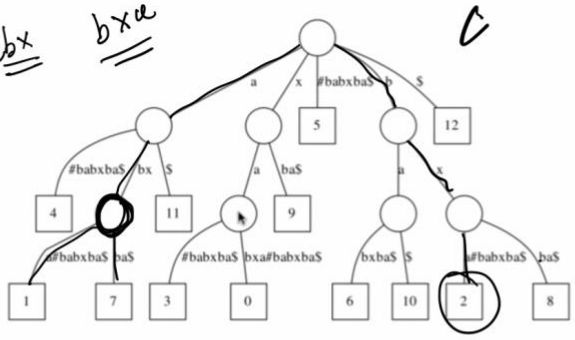




To find the longest common substring (LCS) of X and Y, make a new string $X\#Y\$$ where $\#$ and $\$$ are both terminal symbols. Build a suffix tree for $X\#Y\$$.

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

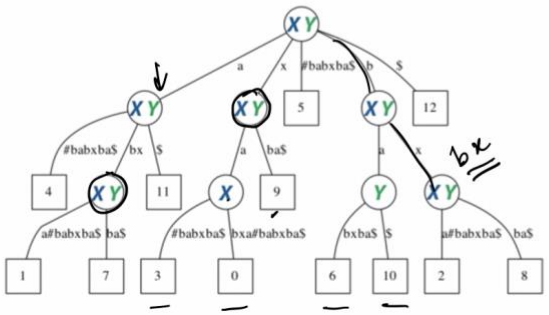
Consider leaves: offsets in [0, 4] are suffixes of X, offsets in [6, 11] are suffixes of Y



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



Suffix trees in the real world

Alignment of whole genomes (MUMmer):

Delcher, Arthur L., et al. "Alignment of whole genomes." *Nucleic Acids Research* 27.11 (1999): 2369-2376.

Delcher, Arthur L., et al. "Fast algorithms for large-scale genome alignment and comparison." *Nucleic Acids Research* 30.11 (2002): 2478-2483.

Kurtz, Stefan, et al. "Versatile and open software for comparing large genomes." *Genome Biol* 5.2 (2004): R12.

~ 2,000 citations `http://mummer.sourceforge.net`

Computing and visualizing repeats in whole genomes (REPuter):

Kurtz, Stefan, and Chris Schleiermacher. "REPuter: Fast computation of maximal repeats in complete genomes." *Bioinformatics* 15.5 (1999): 426-427.

Kurtz, Stefan, et al. "REPuter: the manifold applications of repeat analysis on a genomic scale." *Nucleic acids research* 29.22 (2001): 4633-4642.

~ 740 citations `http://bibiserv.techfak.uni-bielefeld.de/reputer`

Identifying sequence motifs

Marsan, Laurent, and Marie-France Sagot. "Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification." *Journal of Computational Biology* 7.3-4 (2000): 345-362.

Sagot, Marie. "Spelling approximate repeated or common motifs using a suffix tree." *LATIN'98: Theoretical Informatics* (1998): 374-390.

~ 550 citations

Also used in: multiple alignment