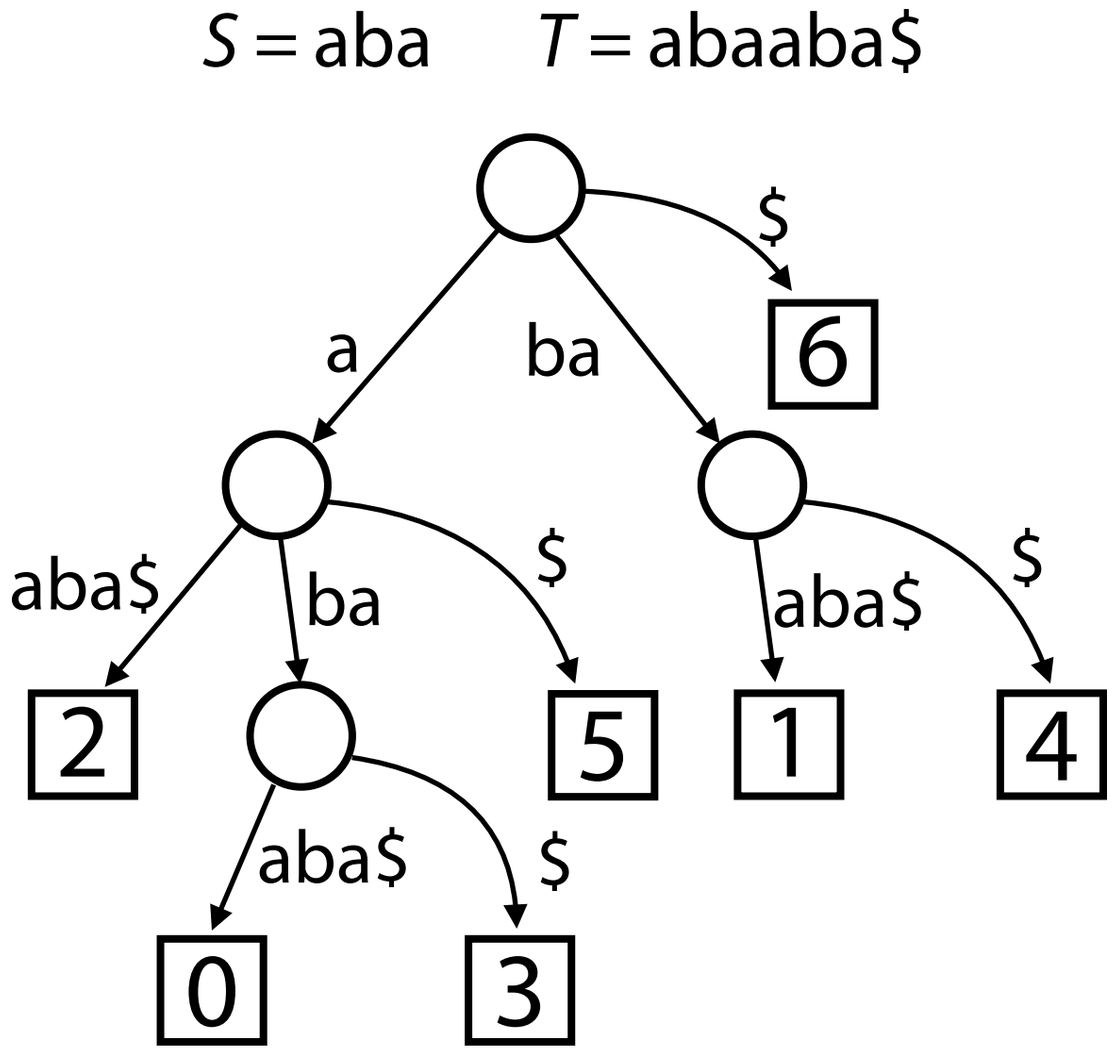
Suffix Trees: suffix links

Ben Langmead

Suffix tree

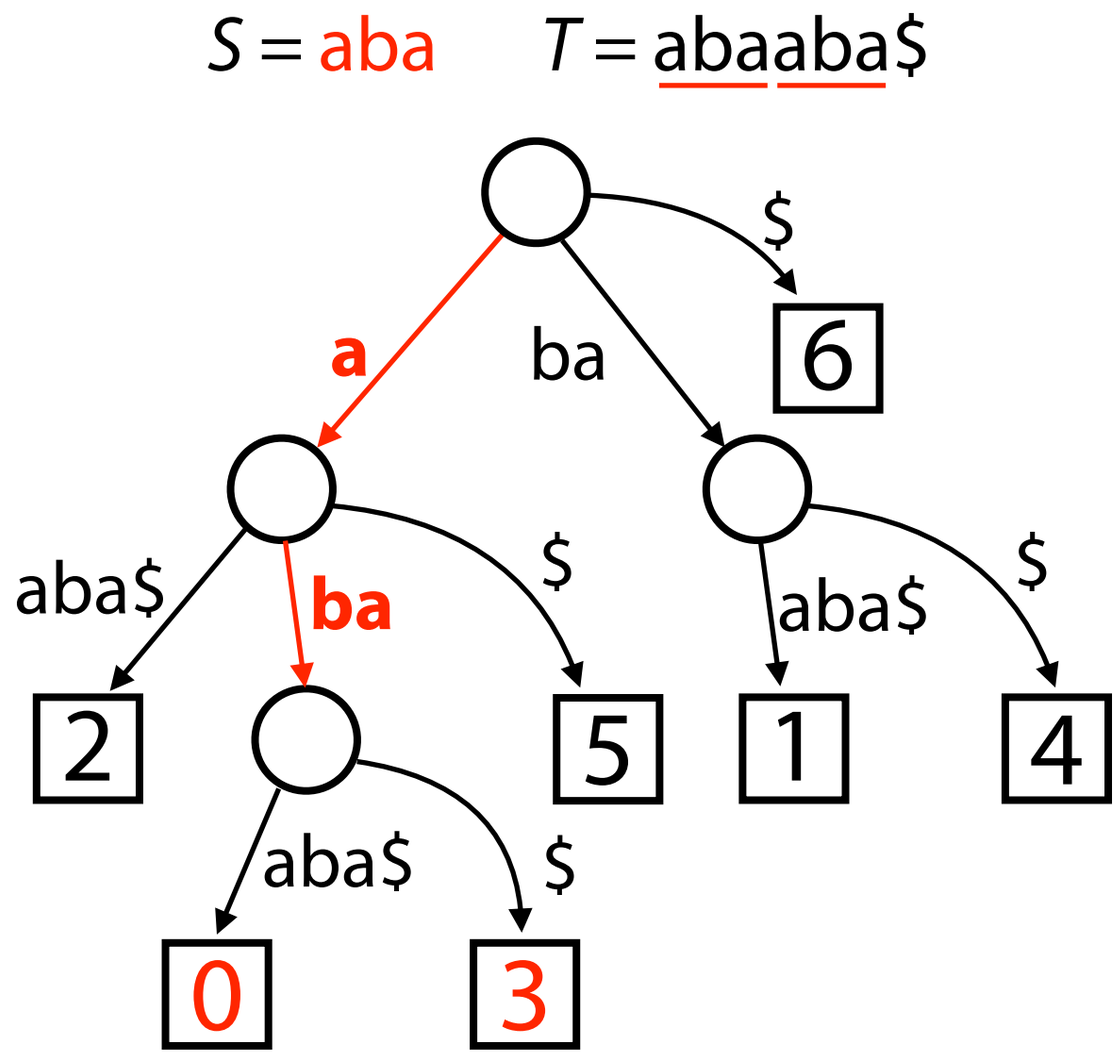
We know a query that solves the exact matching problem



Suffix tree

We know a query that solves the exact matching problem

What about other problems, like finding **substrings** of *S* that match well?



Motivation

Hear the mellow wedding bells,
Golden bells!
What a world of happiness their harmony foretells!
Through the balmy air of night

Hear the sledges with the bells—
Silver bells!
What a world of merriment their melody foretells!
In the icy air of night

Motivation

Hear the mellow wedding bells,
Golden bells!
What a world of happiness their harmony foretells!
Through the balmy air of night

Hear the sledges with the bells—
Silver bells!
What a world of merriment their melody foretells!
In the icy air of night

Motivation: Whole Genome Alignment

2 strains of SARS-CoV-2

AACAACAGAGTTGTTATTTCTAGTGATGTTCTTGTTAACAACTAAACGAACAATGTTTGT
TTTTCTTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTACAACCAGAACTCA
ATTACCCCCTGCATACACTAATTCTTTCACACGTGGTGTTTATTACCCTGACAAAGTTTT
CAGATCCTCAGTTTTACATTCAACTCAGGACTTGTTCTTACCTTTCTTTTCCAATGTTAC
TTGGTTCCATGCTATCTCTGGGACCAATGGTACTAAGAGGTTTGATAACCCTGTCCTACC
ATTTAATGATGGTGTTTATTTTGCTTCCACTGAGAAGTCTAACATAATAAGAGGCTGGAT
TTTTGGTACTACTTTAGATTCGAAGACCCAGTCCCTACTTATTGTTAATAACGCTACTAA
TGTTGTTATTAAAGTCTGTGAATTTCAATTTTGTAATGATCCATTTTTGGGTGTTTACCA
CAAAAACAACAAAAGTTGGATGGAAAGTGAGTTCAGAGTTTATTCTAGTGCGAATAATTG
CACTTTTGAATATGTCTCTCAGCCTTTTCTTATGGACCTTGAAGGAAAACAGGGTAATTT
CAAAAATCTTAGGGAATTTGTGTTTAAGAATATTGATGGTTATTTTAAAAT

AACAACAGAGTTGTTATTTCTAGTGATGTTCTTGTTAACAACTAAACGAACAATGTTTGT
TTTTCTTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTAGAACCAGAACTCA
ATTACCCCCTGCATACACTAATTCTTTCACACGTGGTGTTTATTACCCTGACAAAGTTTT
CAGATCCTCAGTTTTACATTCAACTCAGGACTTGTTCTTACCTTTCTTTTCCAATGTTAC
TTGGTTCCATGCTATACATGTCTCTGGGACCAATGGTACTACGAGGTTTGATAACCCTGT
CCTACCATTTAATGATGGTGTTTATTTTGCTTCCACTGAGAAGTCTAACATAATAAGAGG
CTGGATTTTTGGTACTACTTTAGATTCGAAGACCCAGTCCCTACTTATTGTTAATAACGC
TACTAATGTTGTTATTAAAGTCTGTGAATTTCAATTTTGTAATGATCCATTTTTGGGTGT
TTATTACCACAAAAACAACAAAAGTTGGATGGAAAGTGGAGTTTATTCTAGTGCGAATAA
TTGCACTTTTGAATATGTCTCTCAGCCTTTTCTTATGGACCTTGAAGGAAAACAGGGTAA
TTTCAAAAATCTTAGGGAATTTGTGTTTAAGAATATTGATGGTTATTTTAAAAT

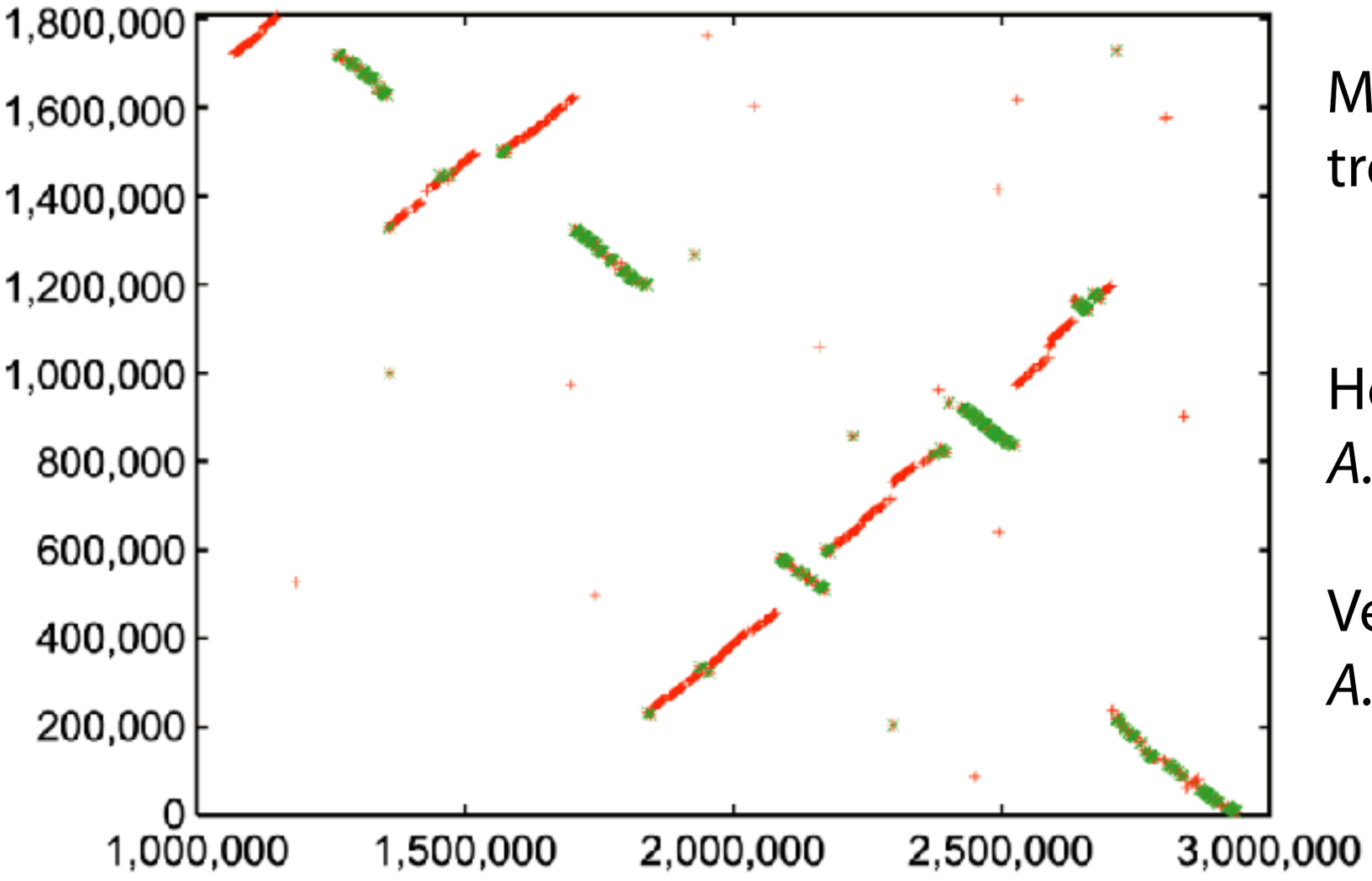
Motivation: Whole Genome Alignment

2 strains of SARS-CoV-2

AACAACAGAGTTGTTATTTCTAGTGATGTTCTTGTTAACAACTAAACGAACAATGTTTGT
TTTTCTTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTA**C**AACCAGAACTCA
ATTACCCCCTGCATACACTAATTCTTTCACACGTGGTGTTTATTACCCTGACAAAGTTTT
CAGATCCTCAGTTTTACATTCAACTCAGGACTTGTTCTTACCTTTCTTTTCCAATGTTAC
TTGGTTCCATGCTATCTCTGGGACCAATGGTACTA**A**GAGGTTTGATAACCCTGTCCTACC
ATTTAATGATGGTGTTTATTTTGCTTCCACTGAGAAGTCTAACATAATAAGAGGCTGGAT
TTTTGGTACTACTTTAGATTCGAAGACCCAGTCCCTACTTATTGTTAATAACGCTACTAA
TGTTGTTATTAAAGTCTGTGAATTTCAATTTTGTAATGATCCATTTTTGGGTGTTTACCA
CAAAAACAACAAAAGTTGGATGGAAAGTG**AGTTCA**GAGTTTATTCTAGTGCGAATAATTG
CACTTTTGAATATGTCTCTCAGCCTTTTCTTATGGACCTTGAAGGAAAACAGGGTAATTT
CAAAAATCTTAGGGAATTTGTGTTTAAGAATATTGATGGTTATTTTAAAAT

AACAACAGAGTTGTTATTTCTAGTGATGTTCTTGTTAACAACTAAACGAACAATGTTTGT
TTTTCTTGTTTTATTGCCACTAGTCTCTAGTCAGTGTGTTAATCTTA**G**AACCAGAACTCA
ATTACCCCCTGCATACACTAATTCTTTCACACGTGGTGTTTATTACCCTGACAAAGTTTT
CAGATCCTCAGTTTTACATTCAACTCAGGACTTGTTCTTACCTTTCTTTTCCAATGTTAC
TTGGTTCCATGCTAT**ACATGT**CTCTGGGACCAATGGTACTA**C**GAGGTTTGATAACCCTGT
CCTACCATTTAATGATGGTGTTTATTTTGCTTCCACTGAGAAGTCTAACATAATAAGAGG
CTGGATTTTTGGTACTACTTTAGATTCGAAGACCCAGTCCCTACTTATTGTTAATAACGC
TACTAATGTTGTTATTAAAGTCTGTGAATTTCAATTTTGTAATGATCCATTTTTGGGTGT
TTA**TTA**CCACAAAAACAACAAAAGTTGGATGGAAAGTGGAGTTTATTCTAGTGCGAATAA
TTGCACTTTTGAATATGTCTCTCAGCCTTTTCTTATGGACCTTGAAGGAAAACAGGGTAA
TTTCAAAAATCTTAGGGAATTTGTGTTTAAGAATATTGATGGTTATTTTAAAAT

Motivation: Whole Genome Alignment



MUMmer uses a suffix tree to make this plot

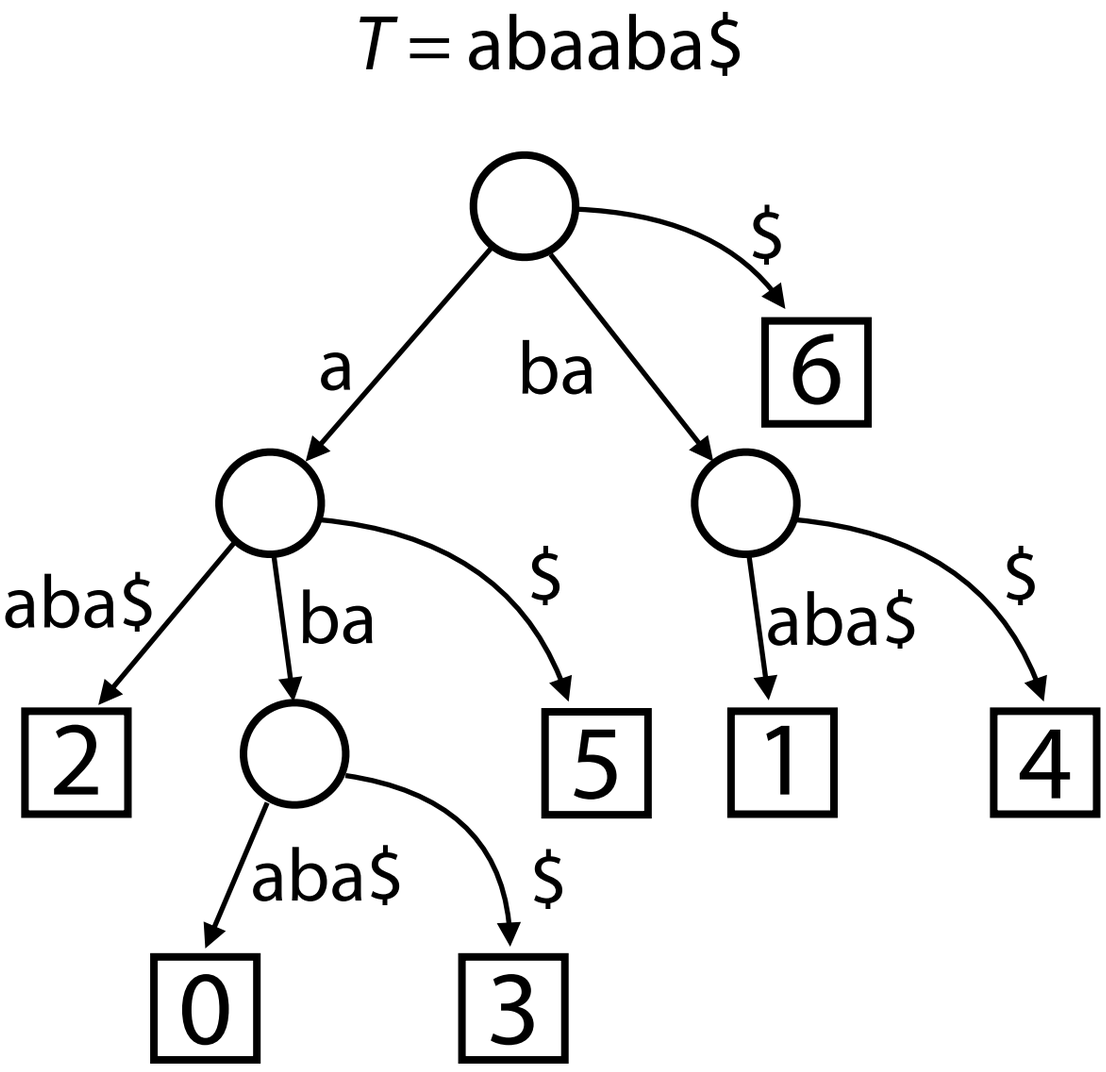
Horizontal: part of the *A. fumigatus* genome

Vertical: part of the *A. nidulans* genome

"Dots" mark similar substring between the genomes; red for "forward" matches, green for "backward"

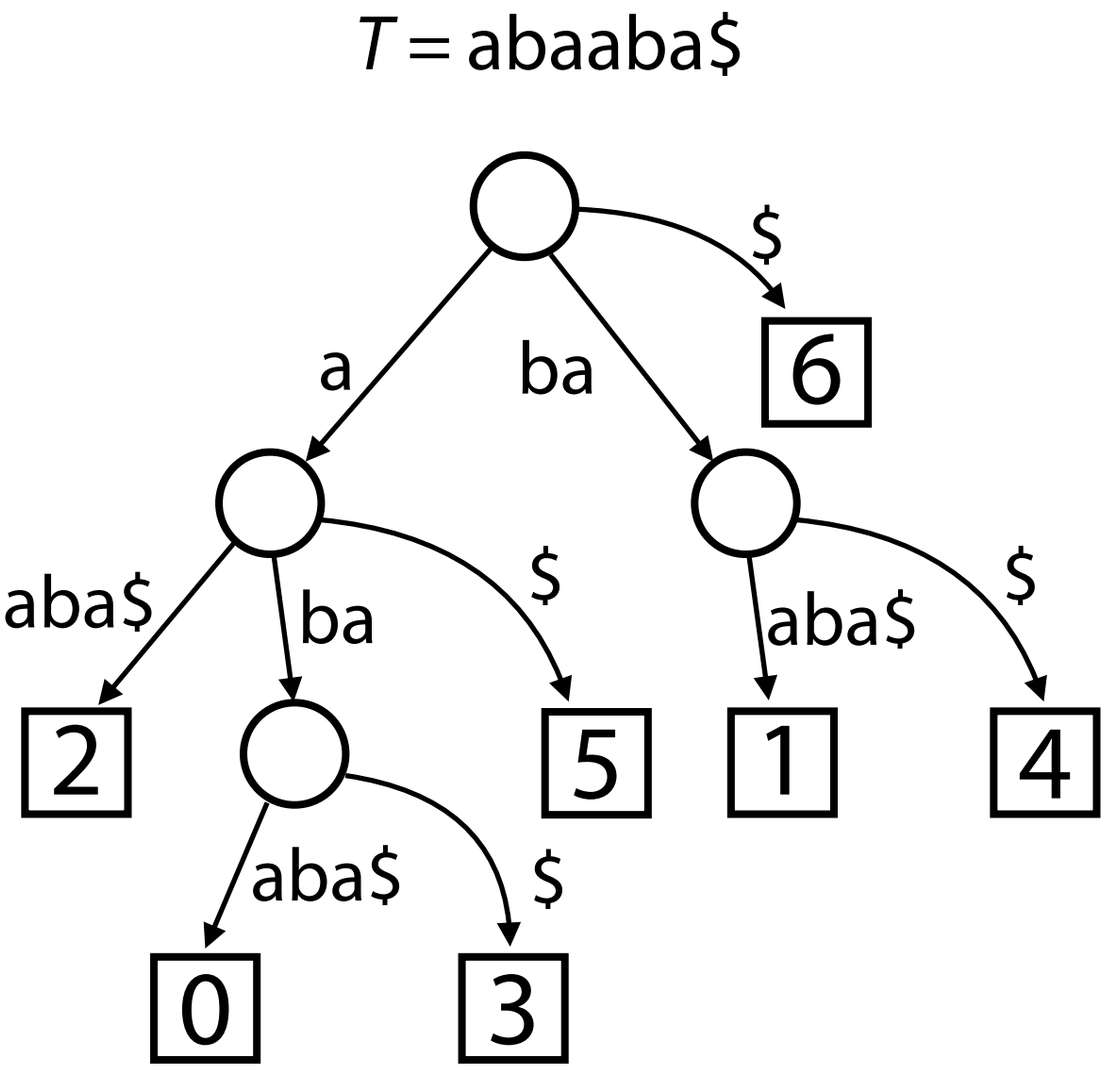
Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.

Suffix tree



S = abab

Suffix tree

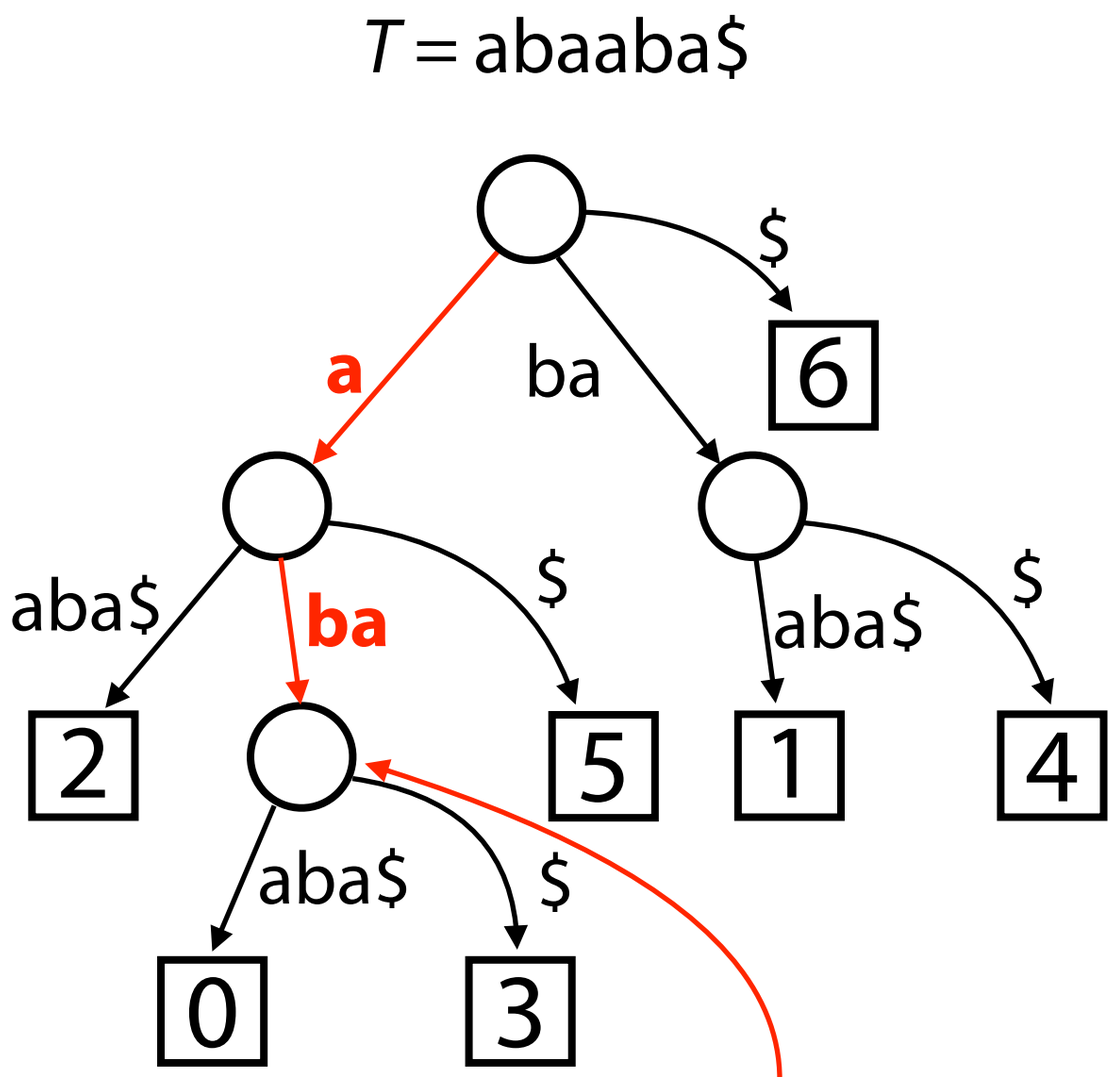


S = abab does not occur...

...but a chunk of it does

How do we discover this while traversing the tree?

Suffix tree



Consider where we fall off

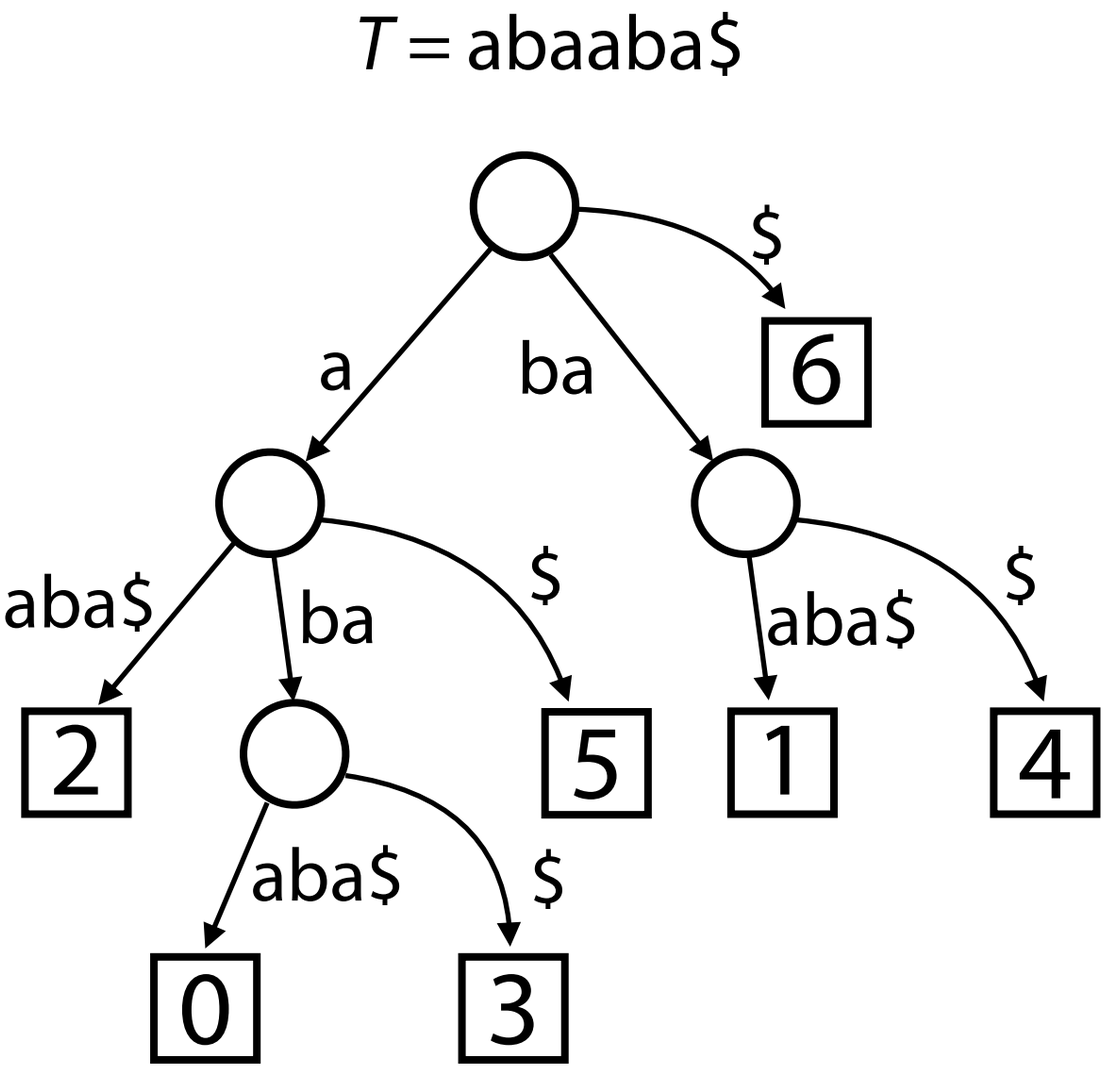
S = abab does not occur...

...but a chunk of it does

How do we discover this while traversing the tree?

We were deep in the tree when we fell off, telling us a prefix of the query matched

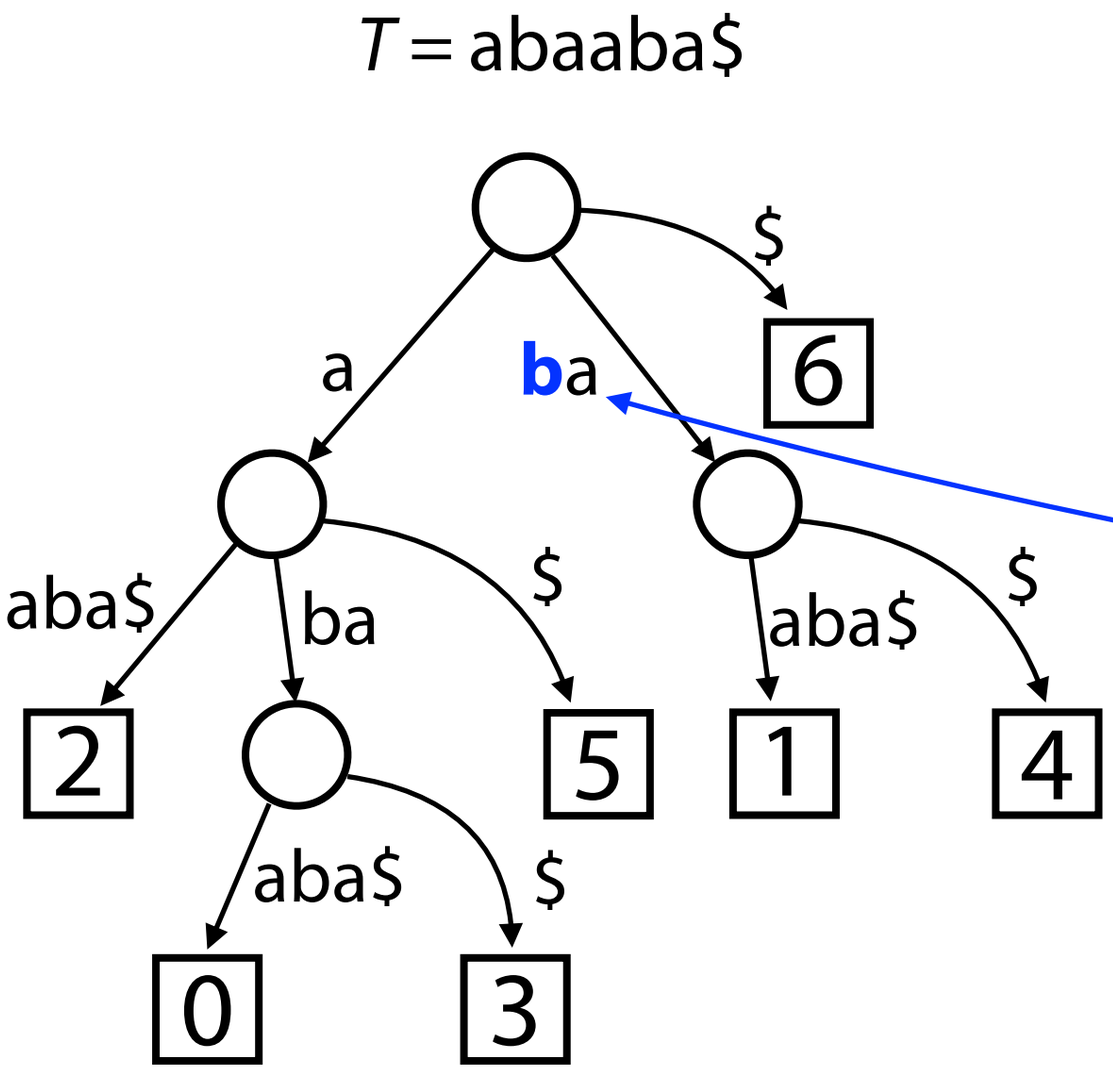
Suffix tree



Consider S = bbaa

Again it does not occur.
But again, a chunk of it does.

Suffix tree



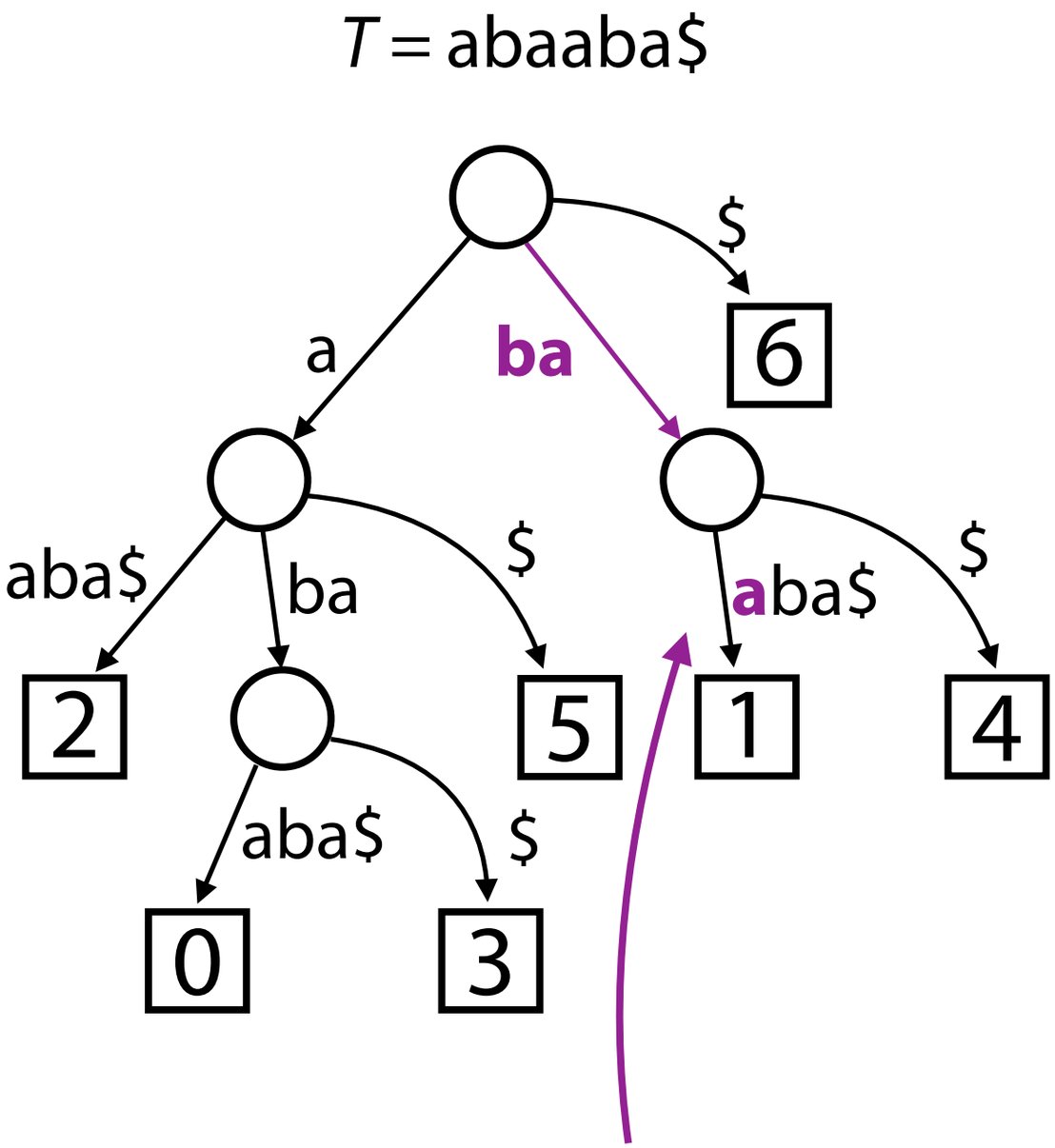
Consider S = bbaa

Again it does not occur.
But again, a chunk of it does.

Consider where we **fall off**

Instead of falling off, what if we **start over** at the next character...

Suffix tree



We would find this longer match. Starting over helped!

Consider S = bbaa

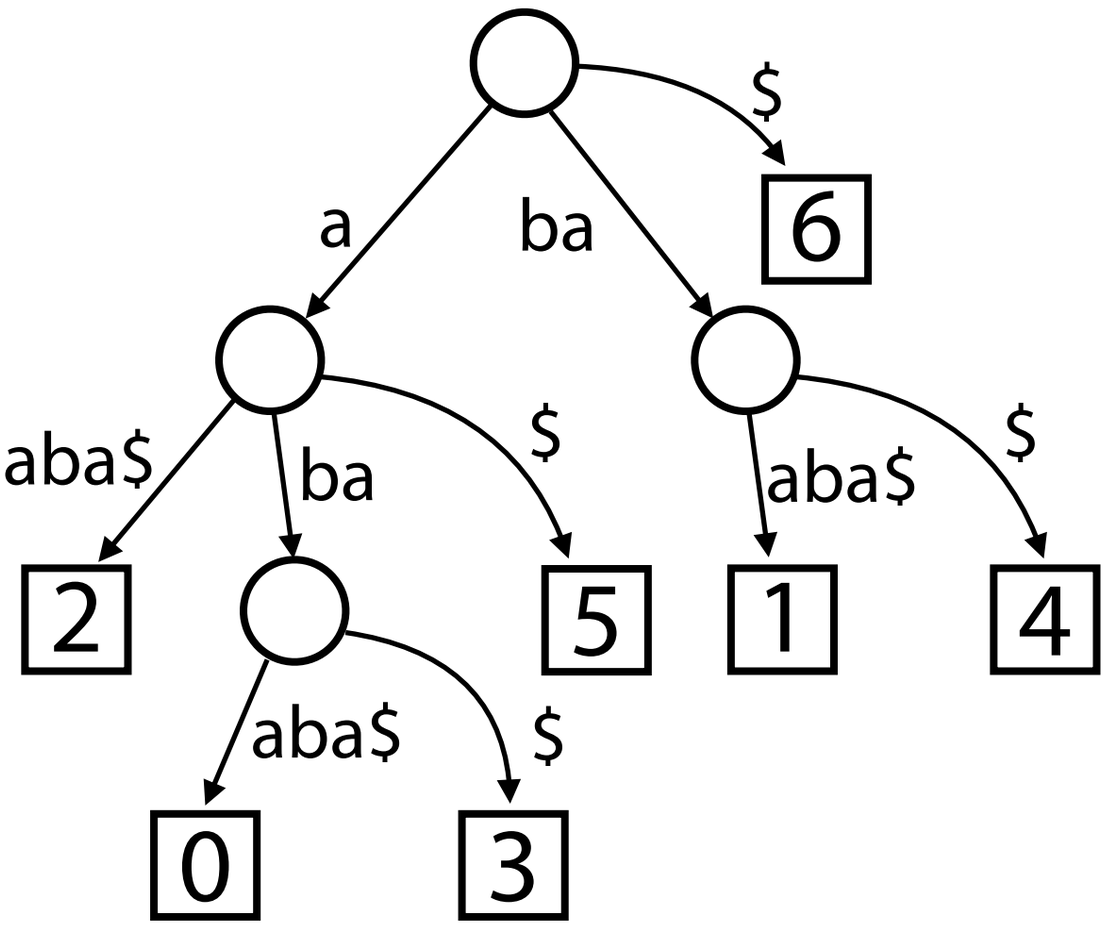
Again it does not occur.
But again, a chunk of it does.

Consider where we **fall off**

Instead of falling off, what if we **start over** at the next character...

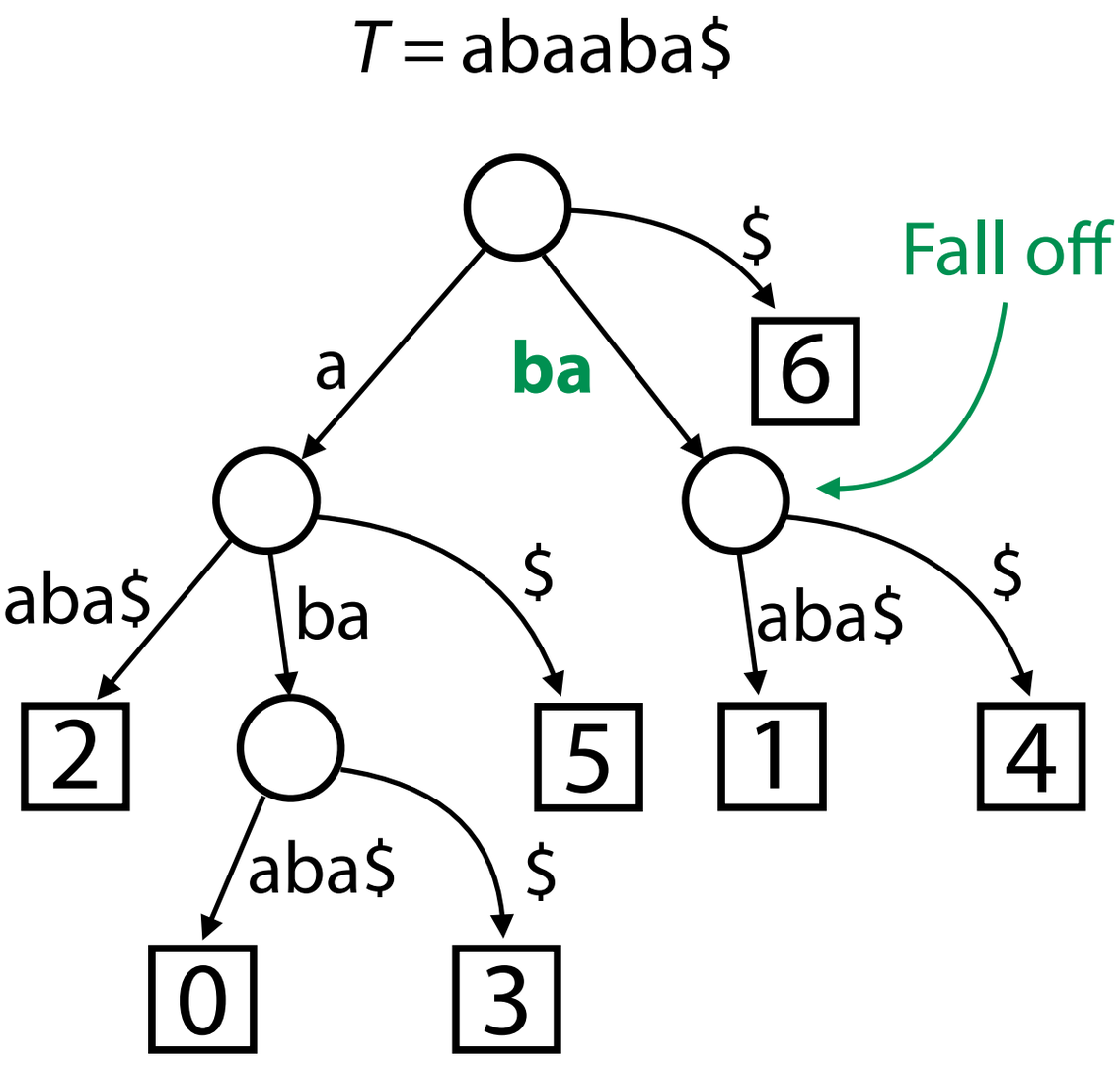
Suffix tree

T = abaaba$



Now take S = baba

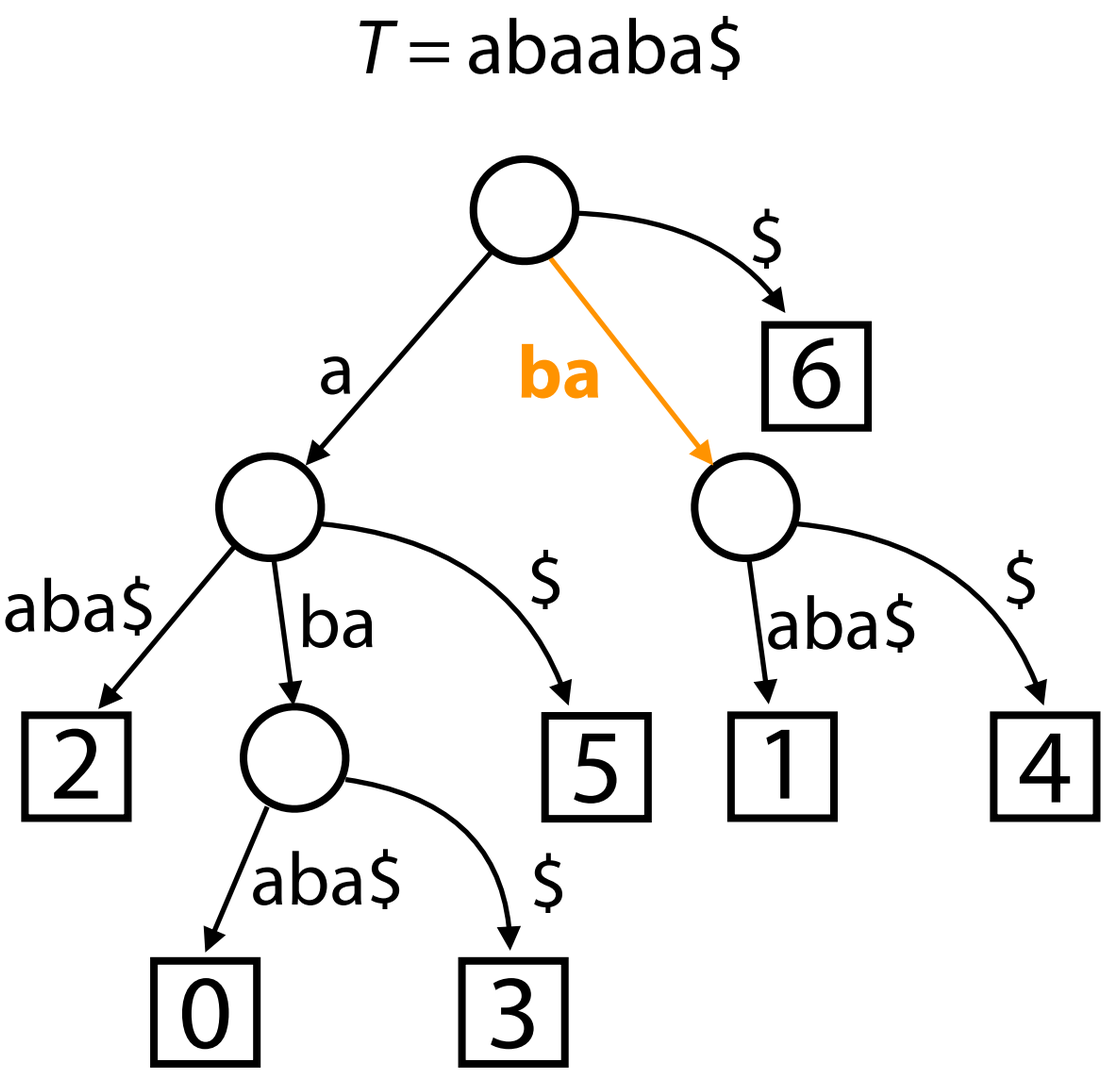
Suffix tree



Now take S = baba

If we reset and start matching again at the second b...

Suffix tree

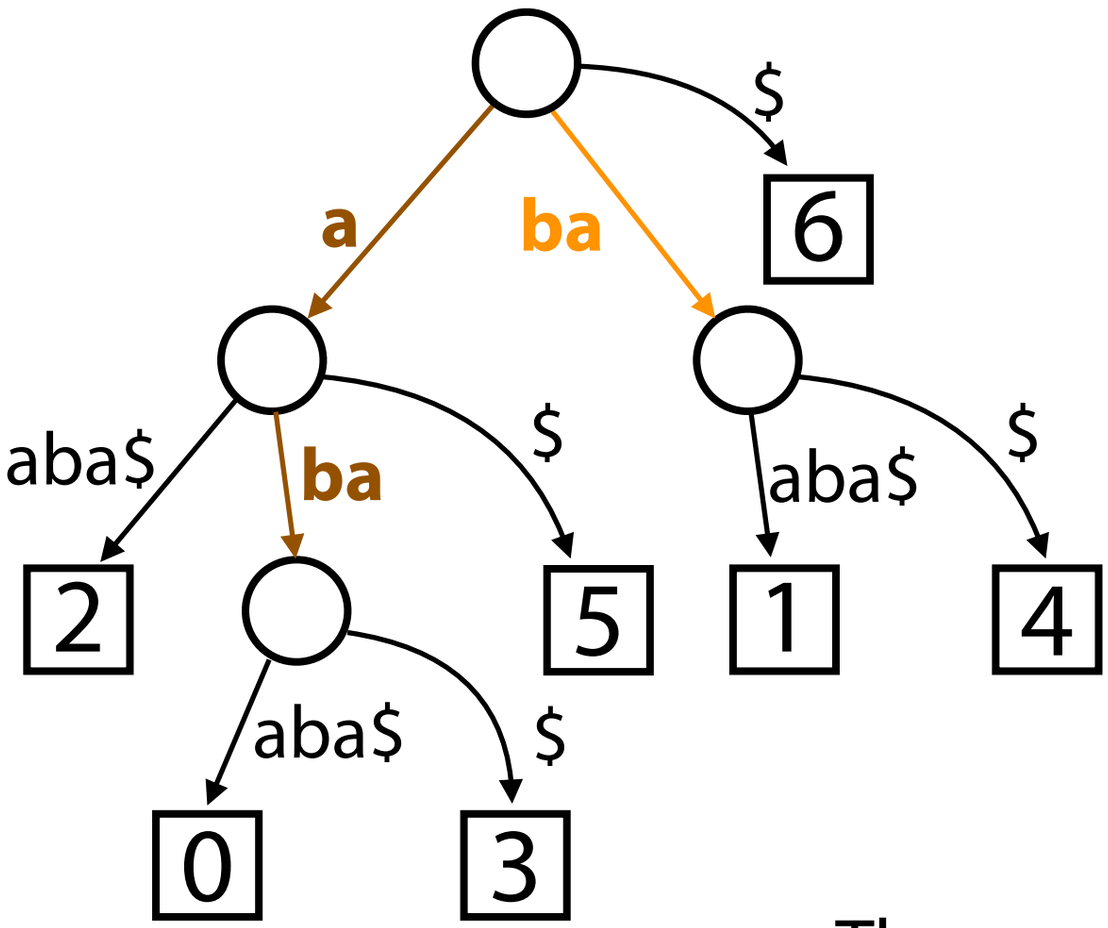


Now take S = baba

If we reset and start matching again at the second b...

We find the match **ba**

Suffix tree



Now take S = baba

If we reset and start matching again at the second b...

We find the match **ba**

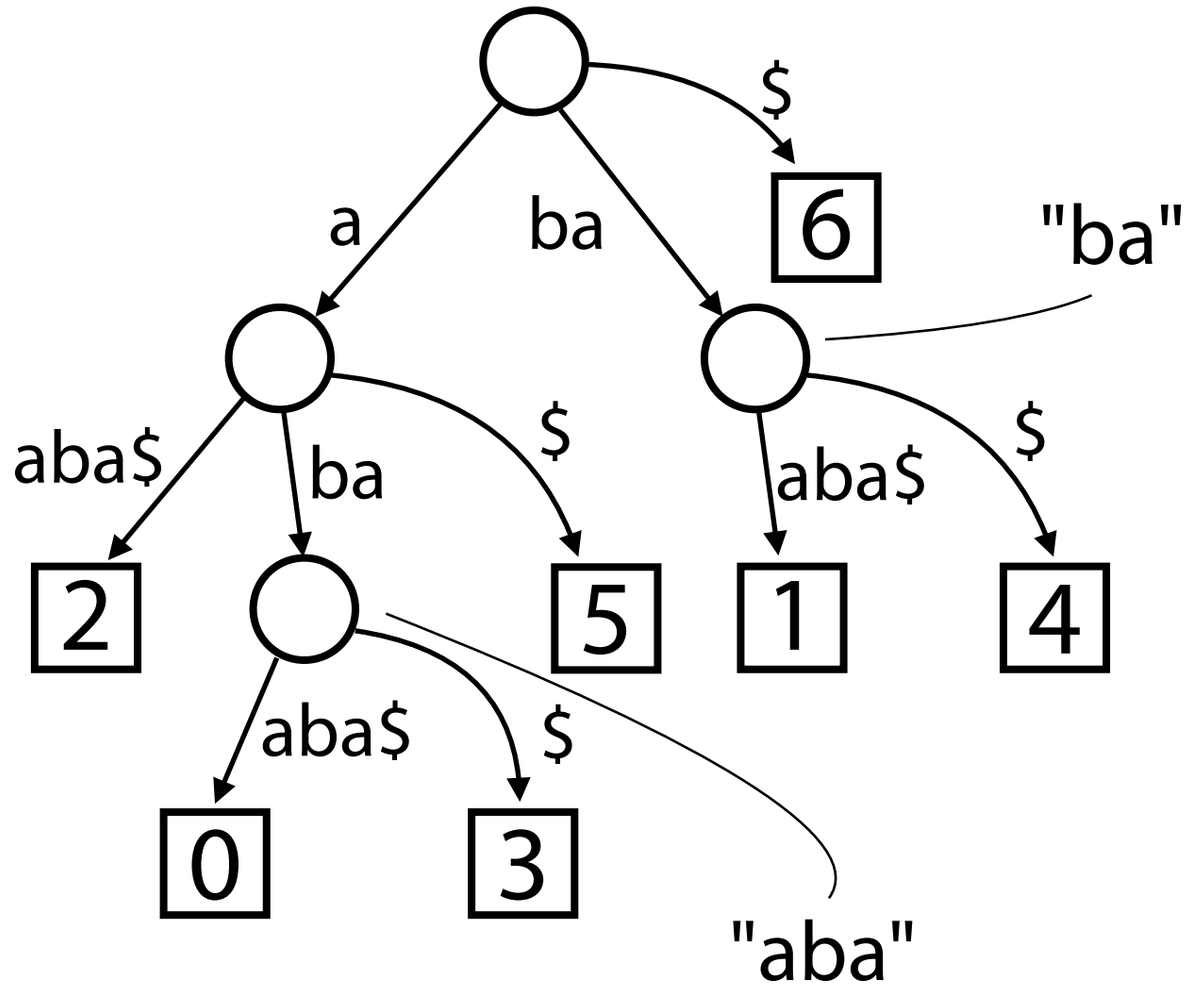
But we missed the longer match **aba**

The **reset** failed to "carry over" the **a** from the first match into the second match

Suffix links

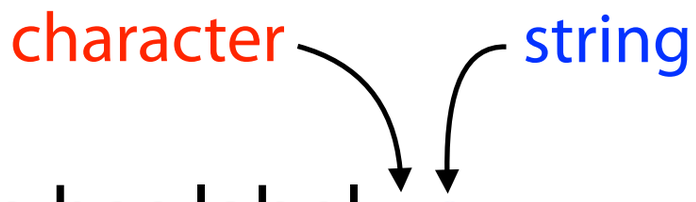
To "carry over" partial matches, we augment the suffix tree with **suffix links**

Recall: we consider that nodes have *labels*

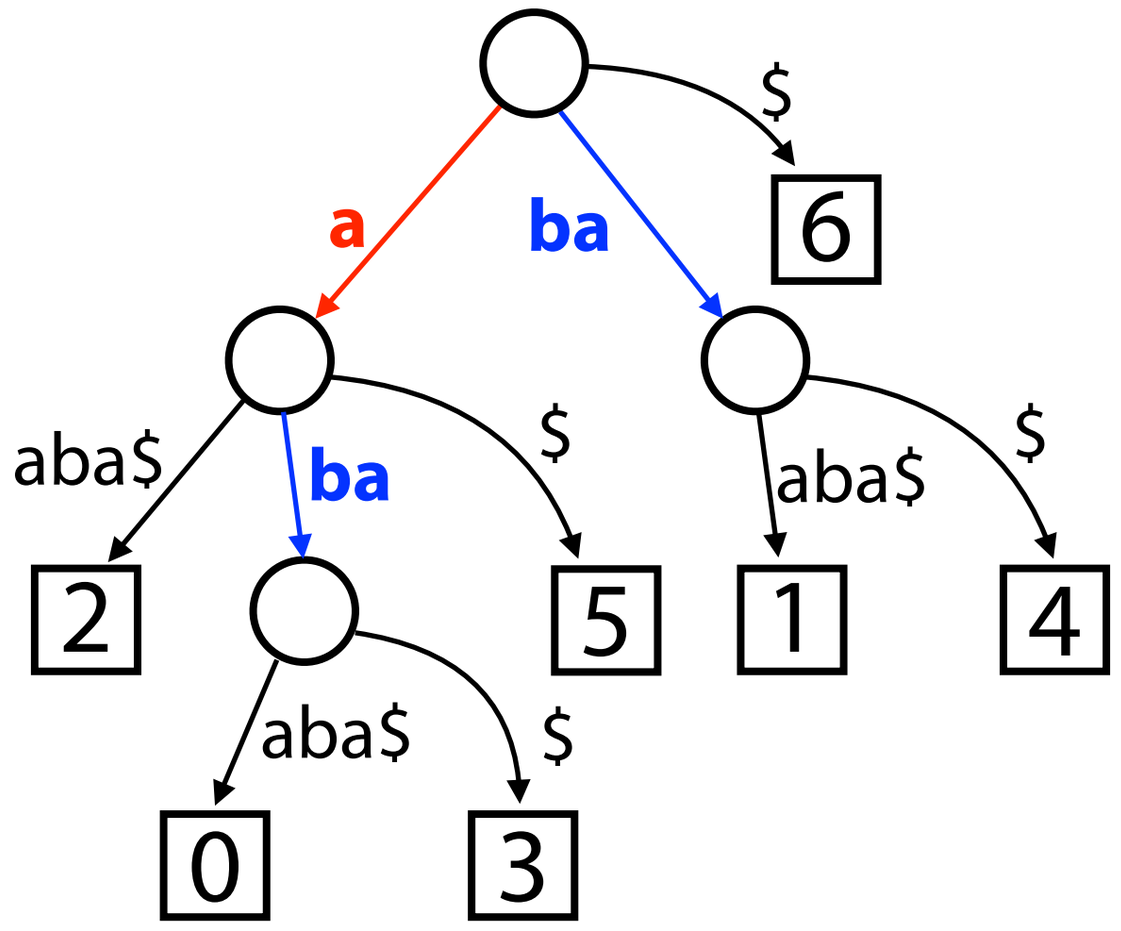


Suffix links

To "carry over" partial matches, we augment the suffix tree with **suffix links**

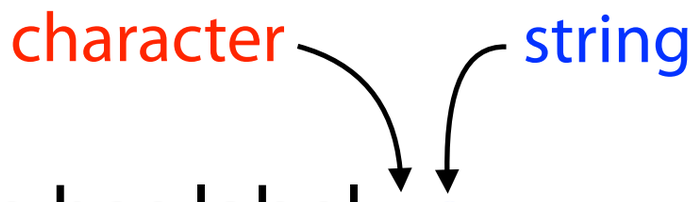


If one node has label $x\alpha$ and the other has label α, we draw a special edge ("suffix link") from $x\alpha$ to α

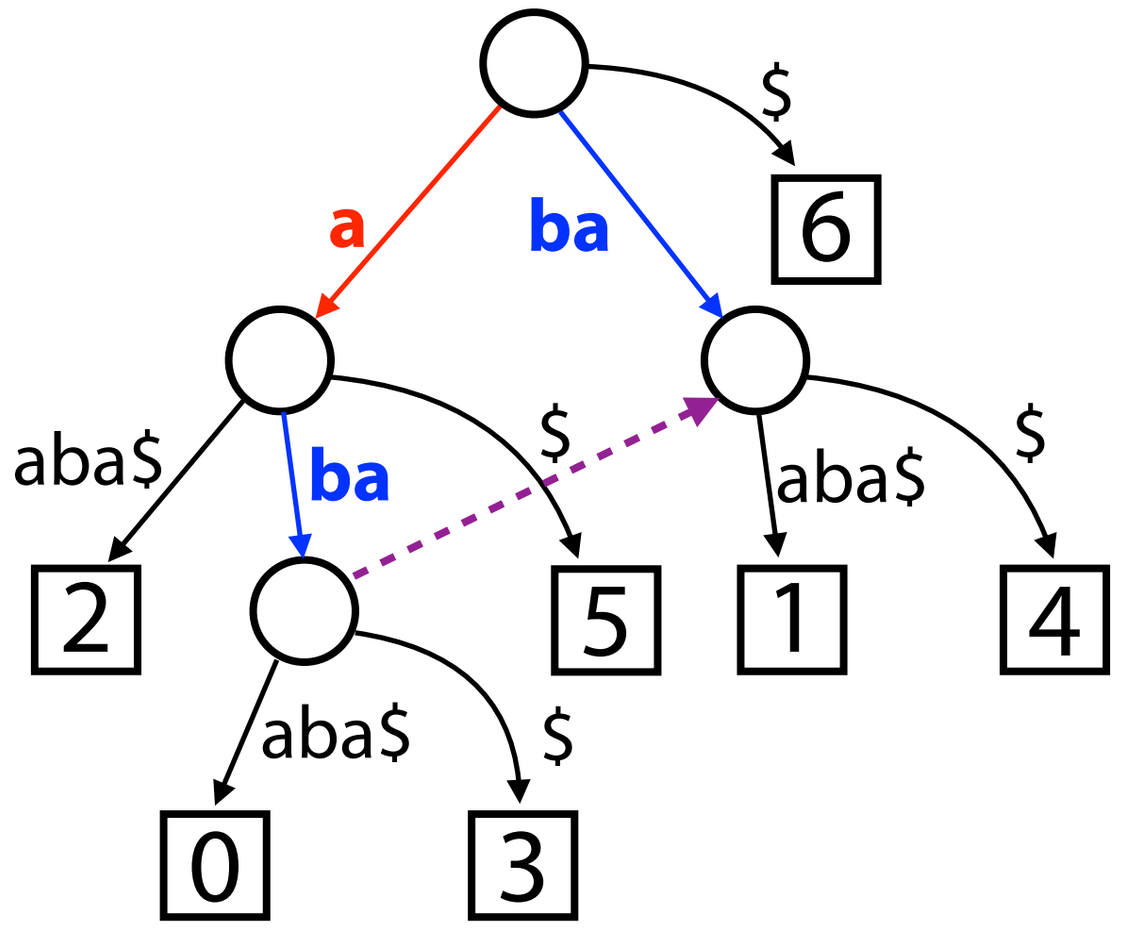


Suffix links

To "carry over" partial matches, we augment the suffix tree with **suffix links**



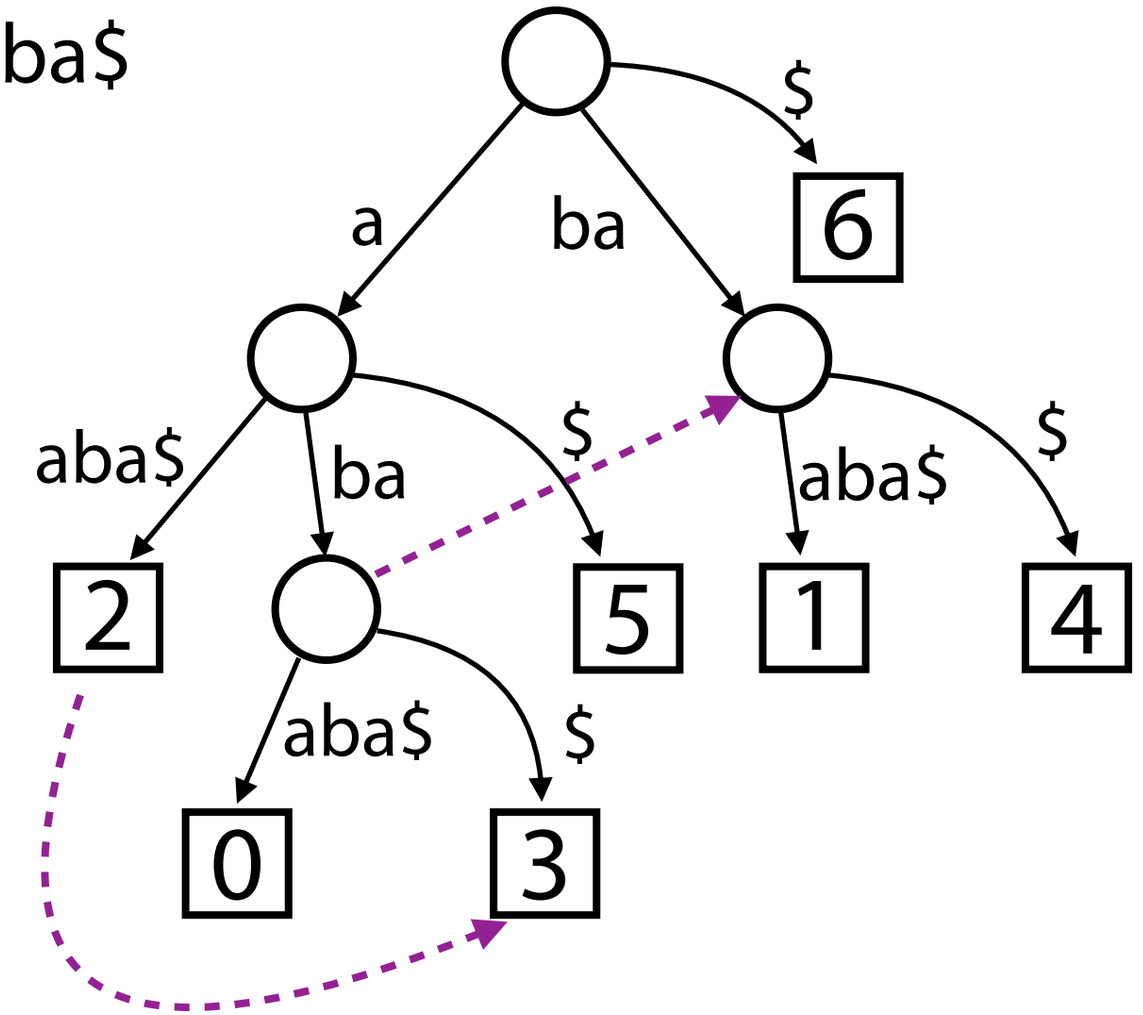
If one node has label $x\alpha$ and the other has label α, we draw a special edge ("suffix link") from $x\alpha$ to α



In this case α = ba and x = a

Suffix links

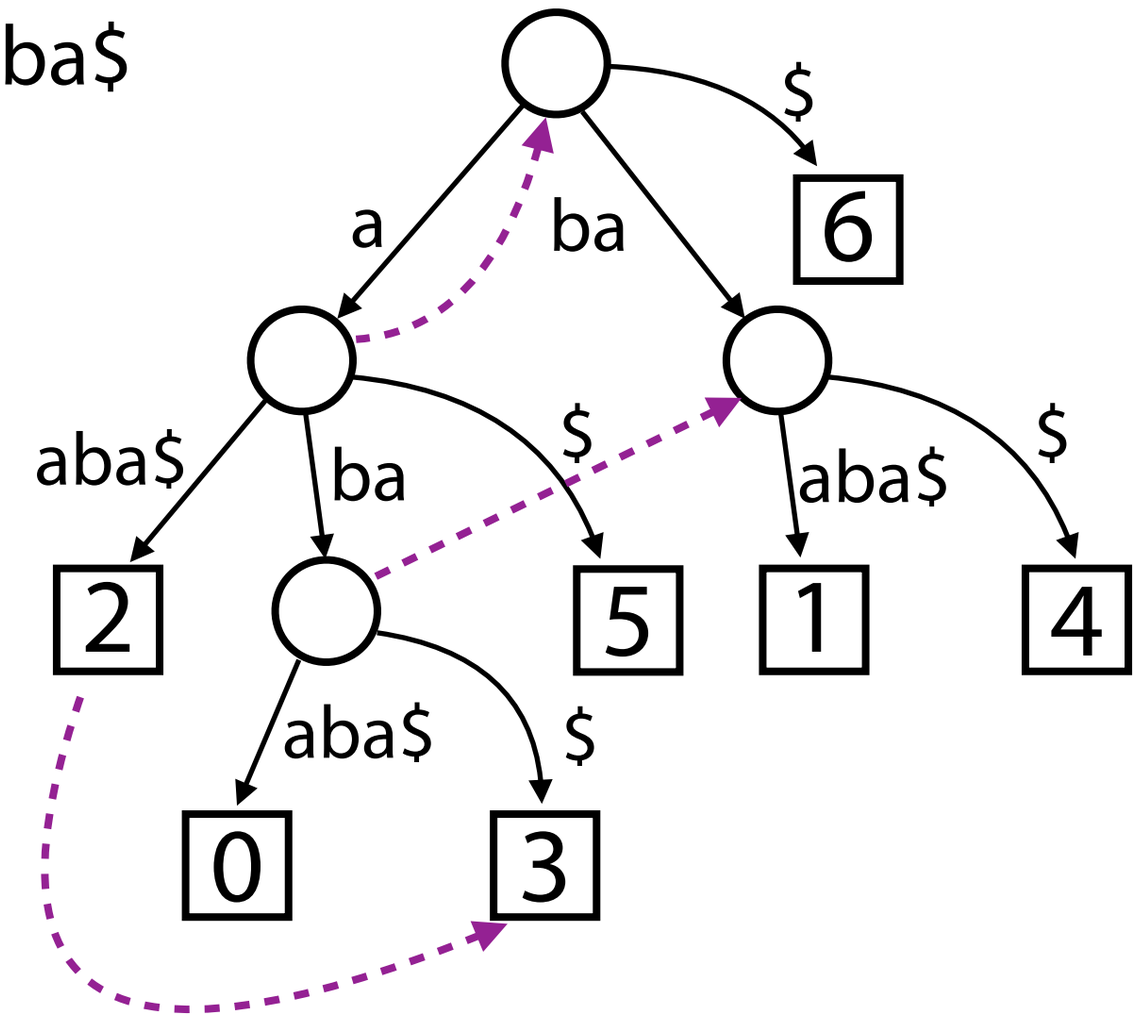
A suffix link can have a leaf as a source or sink, e.g. aaba$ → aba$



Suffix links

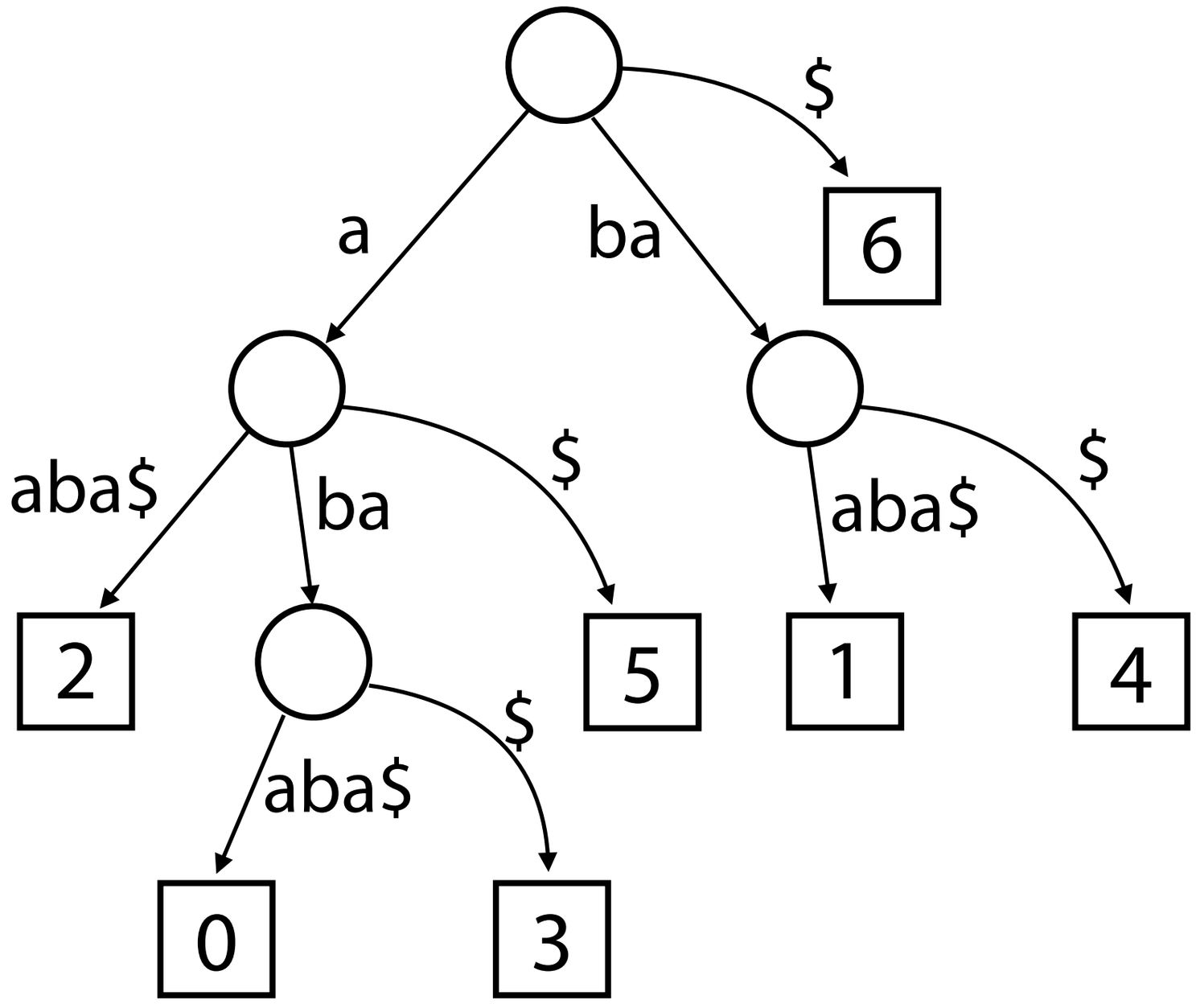
A suffix link can have a leaf as a source or sink, e.g. aaba$ → aba$

A suffix link can also have the root as a sink e.g. a → ϵ (empty string)



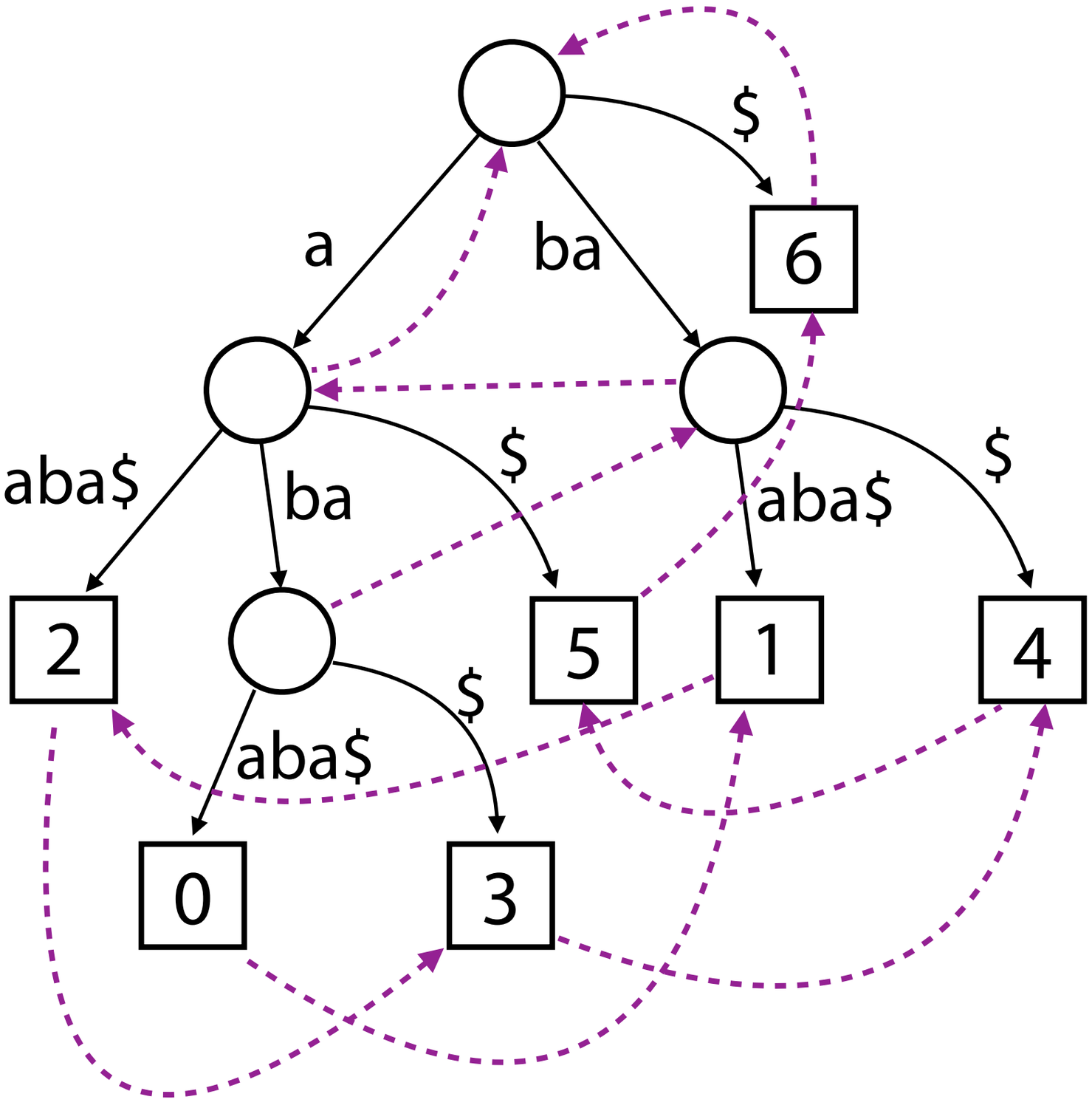
Suffix links

Let's put in all the suffix links!

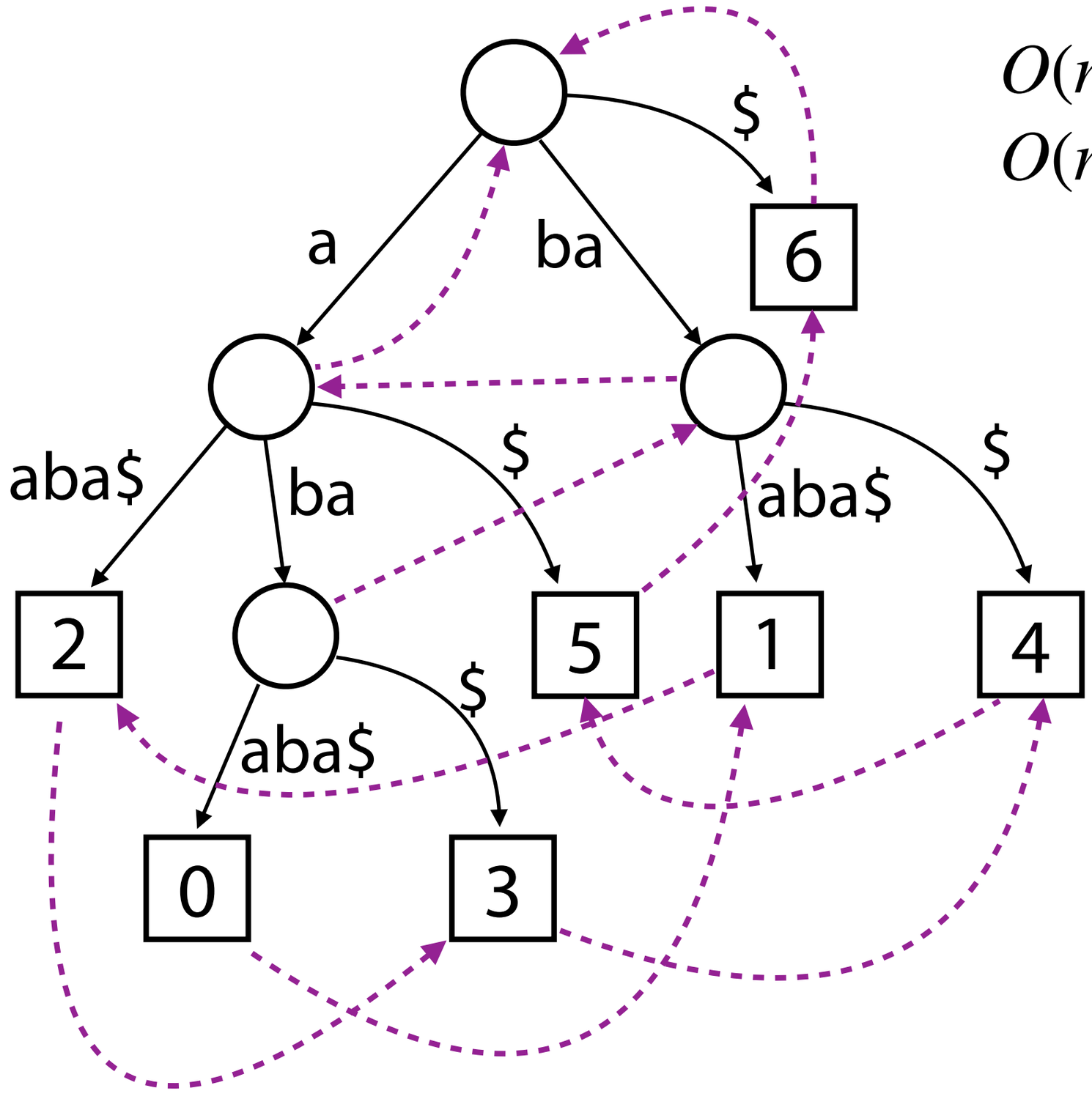


Suffix links

Let's put in all the suffix links!



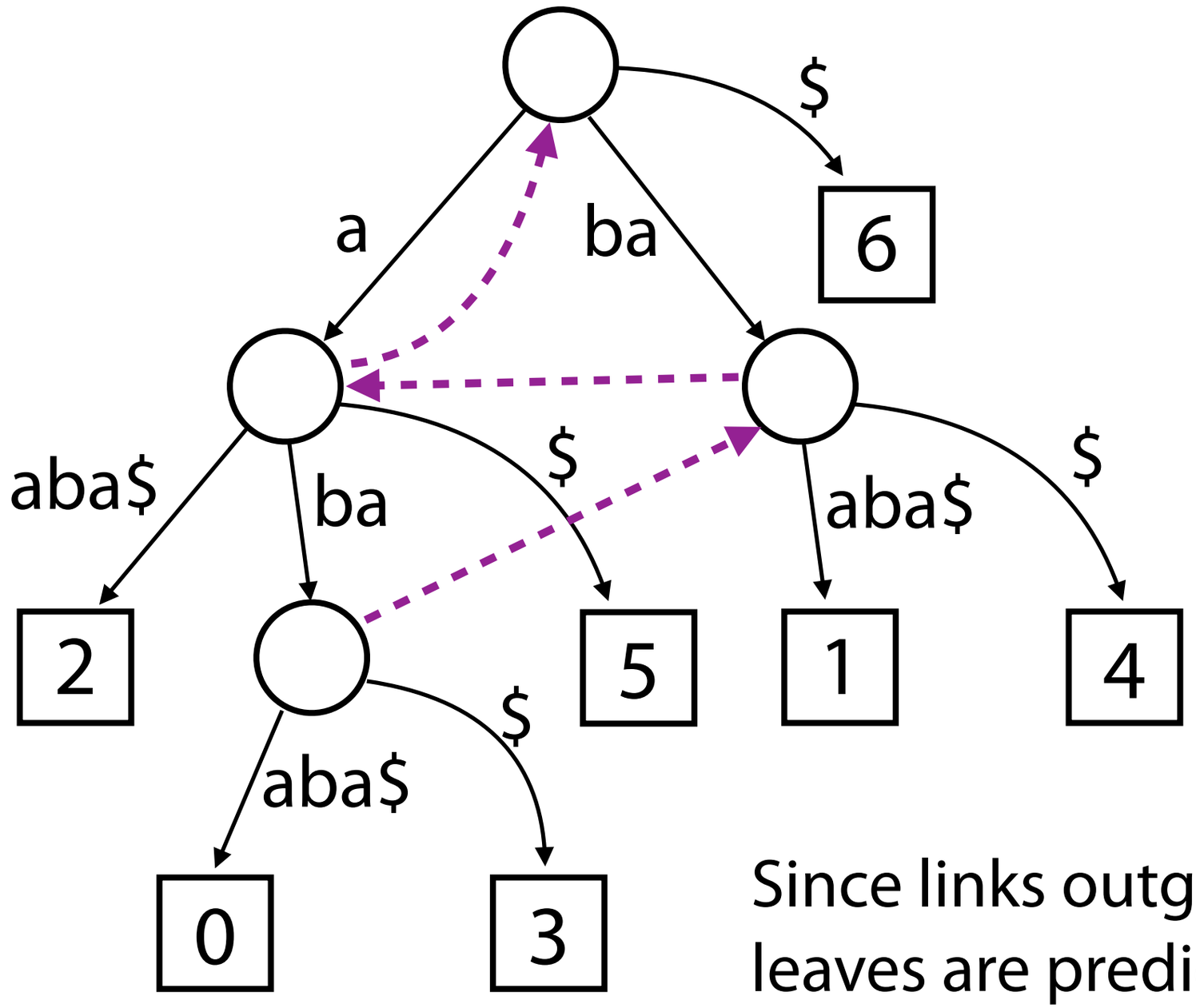
Suffix links



$O(m)$ links, so still

$O(m)$ overall

Suffix links



Since links outgoing from leaves are predictable, let's leave them out of the picture

Suffix links

Suffix links allow partial matches to "carry over"

Coming videos: using suffix links to find similar substrings between P and T, i.e. "matching statistics"