Algorithms theory

15 – Text search (1)

Prof. Dr. S. Albers

Text search

Various scenarios:

Static texts

- Literature databases
- Library systems
- Gene databases
- World Wide Web

Dynamic texts

- Text editors
- Symbol manipulators

Properties of suffix trees

Search index

for a text σ in order to search for several patterns α.

Properties:

1. **Substring searching** in time $O(|\alpha|)$.

2. **Queries to σ itself**, e.g.:
   Longest substring of σ that occurs at least twice.

3. **Prefix search:** all positions in σ with prefix α.

Properties of suffix trees

4. **Range search:** all locations (substrings) in σ belonging to an interval
$[\alpha, \beta]$ with $\alpha \leq_{\text{lex}} \beta$, e.g.

abrakadabra, acacia $\in$ [abc, acc],

abacus $\notin$ [abc, acc] .

5. **Linear complexity:**
Space requirement and construction time in $O(|\sigma|)$.

Tries

Trie: A tree representing a set of keys.

Alphabet Σ, set S of keys, $S \subset \Sigma^*$

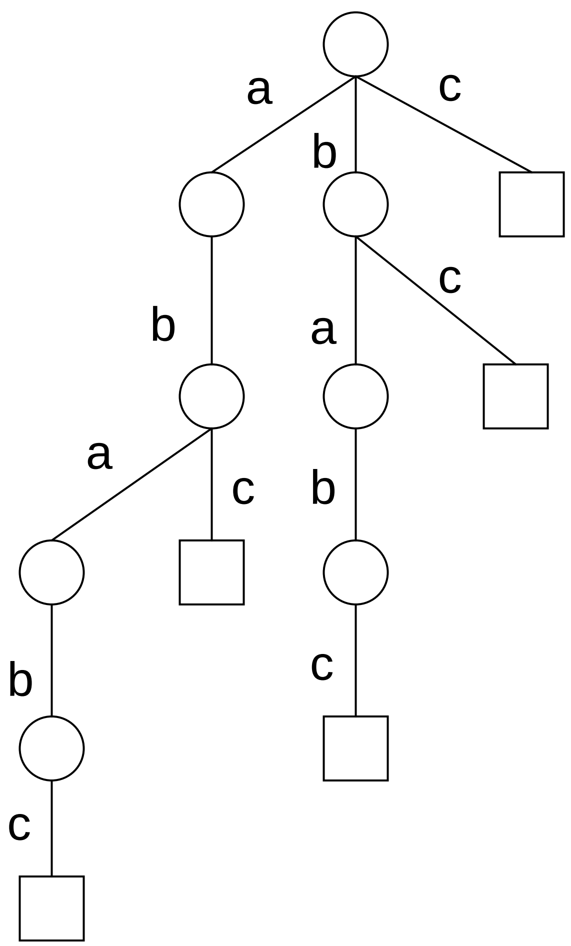
Key: string in Σ^*

Edge of a trie T: labeled with a single character of Σ

Neighboring edges (edges that lead to different children of a node):
labeled with different characters

Tries

Example:



Tries

A **leaf** represents a key:

The corresponding key is the string consisting of the edge labels along the path from the root to the leaf.

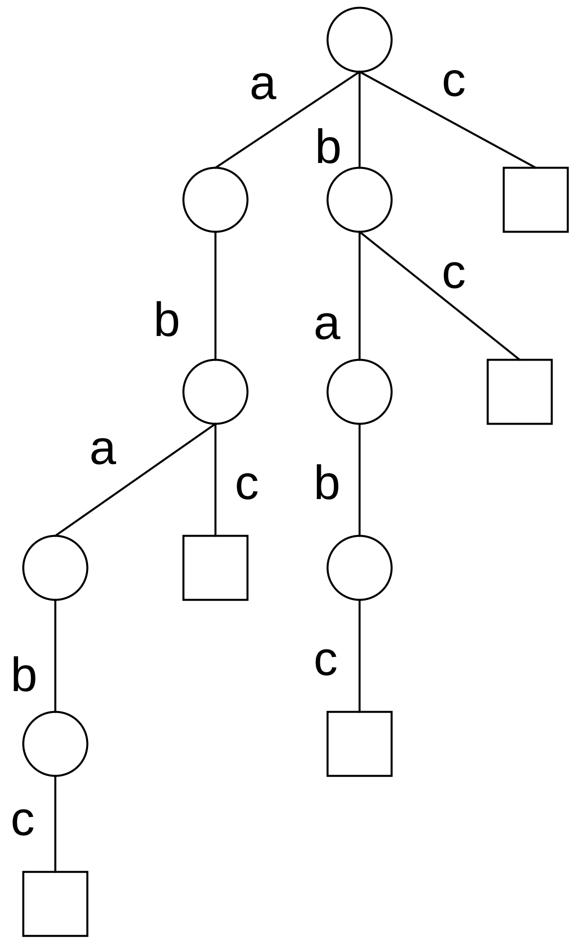
Keys are not stored in nodes!

Suffix tries

Trie representing all suffixes of a string

Example: σ = ababc

suffixes: ababc = suf_1
babc = suf_2
abc = suf_3
bc = suf_4
c = suf_5



Suffix tries

Internal nodes of a suffix trie $\triangleq$ substrings of σ

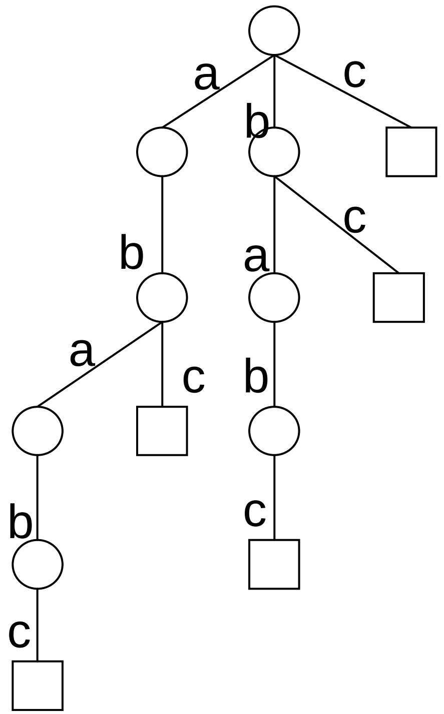
Each proper substring of σ is represented by an internal node.

Let $\sigma = a^n b^n$. Then, there are $n^2 + 2n + 1$ different substrings (or internal nodes).

$\Rightarrow$ space requirement in $O(n^2)$

Suffix tries

A suffix trie T satisfies some of the desired properties:



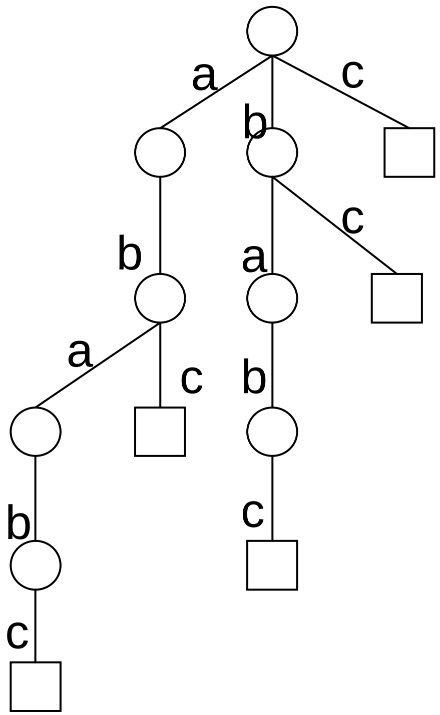
1. String matching for α: Following the path with edge labels α takes $O(|\alpha|)$ time.
   leaves of the subtree $\triangleq$ occurrences of α

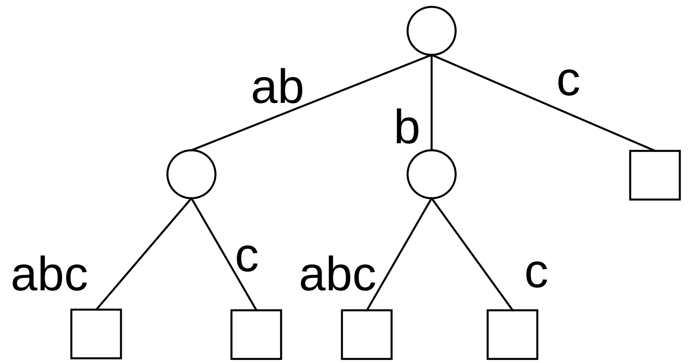
2. Longest substring occurring at least twice: internal node with maximum depth having at least two chilren

3. Prefix search: All occurrences of strings with prefix α are represented by the nodes of the subtree rooted at the internal node corres-ponding to α.

Suffix trees

A suffix tree is obtained from a suffix trie by contracting unary nodes:





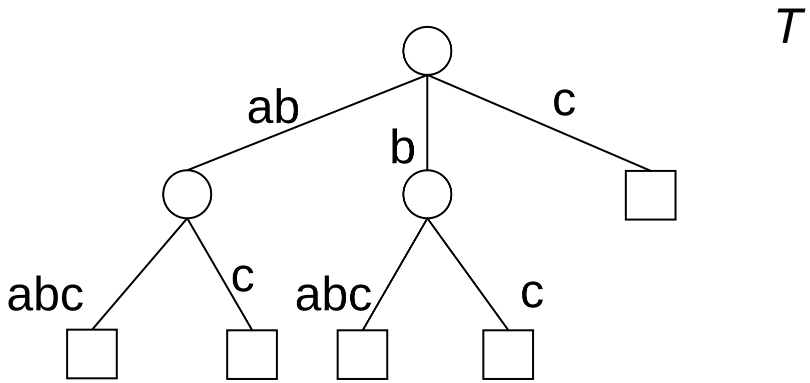
suffix tree = contracted suffix trie

Internal representation of suffix trees

Child-sibling representation

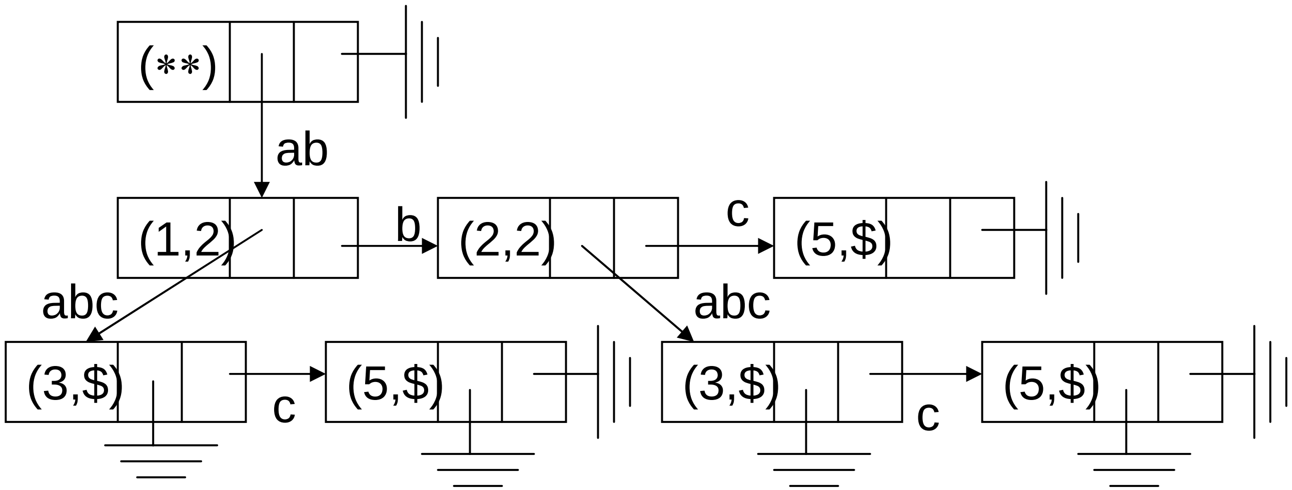
substring: pair of numbers (i,j)

Example: σ = ababc



Internal representation of suffix trees

Example: σ = ababc



node $v = (v.l, v.u, v.c, v.s)$

Further pointers (suffix links) are added later.

Properties of suffix trees

(S1) No suffix is prefix of another suffix.
This holds if the last character of σ is $\$ \notin \Sigma$.

Search:

(T1) edge $\hat{=}$ non-empty substring of σ.

(T2) neighboring edges :
corresponding substrings start with different characters

Properties of suffix trees

Size

(T3) each internal node ($\neq$ root) has at least two children

(T4) leaf $\triangleq$ (non-empty) suffix of σ.

Let $n = |\sigma| \neq 1$.

$\overset{(T4)}{\Longrightarrow}$ number of leaves $= n$

$\overset{(T3)}{\Longrightarrow}$ number of internal nodes $\leq n-1$

$\Rightarrow$ space requirement in $\mathrm{O}(n)$

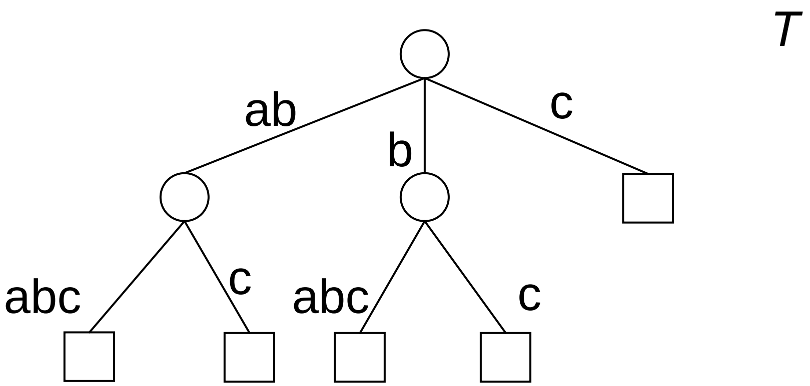
Construction of suffix trees

Definitions:

Partial path: Path from the root to a node in *T*.

Path: A partial path ending at a leaf.

Location of a string α : Node where the partial path corresponding to α ends (if it exists).

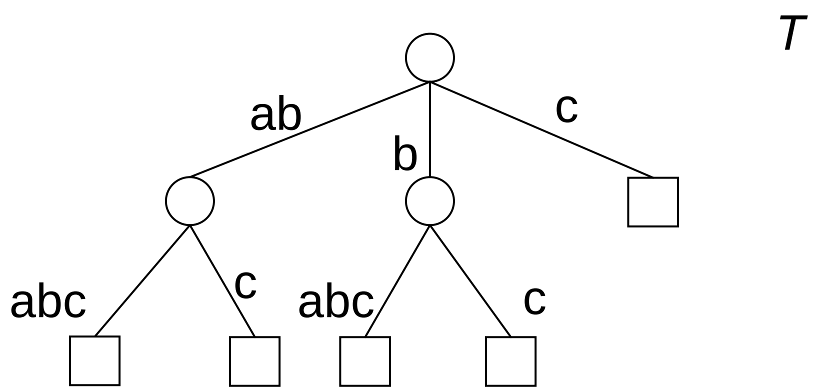


Construction of suffix trees

Extension of a string α : string with prefix α

Extended location of a string α : location of the shortest extension of α whose location is defined

Contracted location of a string α : location of the longest prefix of α whose location is defined



Construction of suffix trees

Definitions:

suf_i : suffix of σ beginning at position i, e.g. $suf_1 = \sigma$, $suf_n = \$$.

$head_i$: longest prefix of suf_i which is also a prefix of suf_j for some $j < i$.

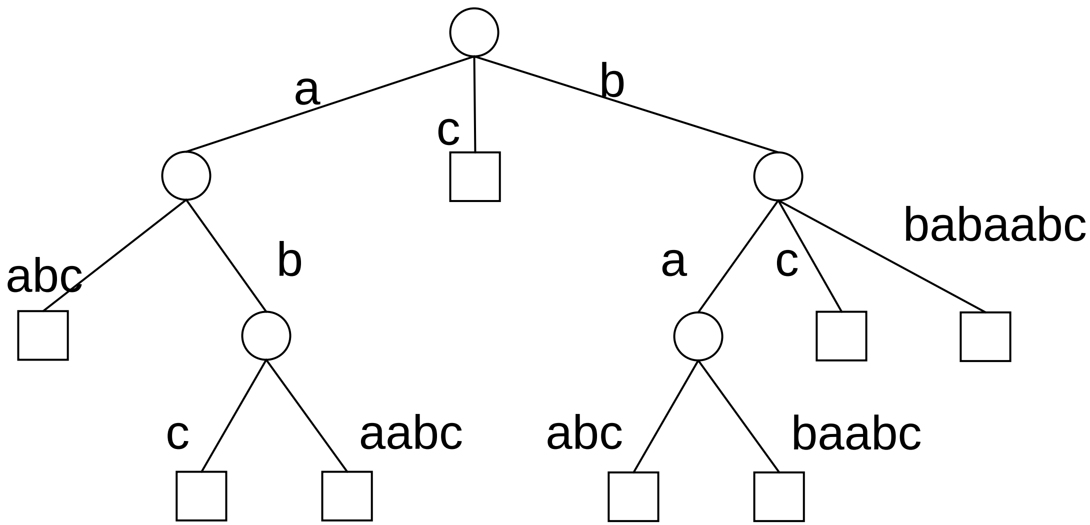
Example: σ = bbabaabc α = baa (has no location)

suf_4 = baabc

$head_4$ = ba

Construction of suffix trees

σ = bbabaabc



Naive suffix tree construction

Start with the empty tree T_0 .

The tree T_{i+1} is constructed from T_i by inserting the suffix suf_{i+1}.

Algorithm *suffix-tree*

Input: string σ

Output: suffix tree T for σ

$n := |\sigma|$; $T_0 := \emptyset$;
for $i := 0$ **to** $n - 1$ **do**
    insert suf_{i+1} into T_i, store the result in T_{i+1} ;
end for

Naive suffix tree construction

All suffixes suf_j with $j \leq i$ have a location in T_i .

$\rightarrow$ $head_{i+1}$ = longest prefix of suf_{i+1} whose extended location exists in T_i

Definition:

$tail_{i+1} := suf_{i+1} - head_{i+1}$ i.e. $suf_{i+1} = head_{i+1}\, tail_{i+1}$.

$$\overset{(S1)}{\Longrightarrow} \quad tail_{i+1} \neq \varepsilon.$$

Naive suffix tree construction

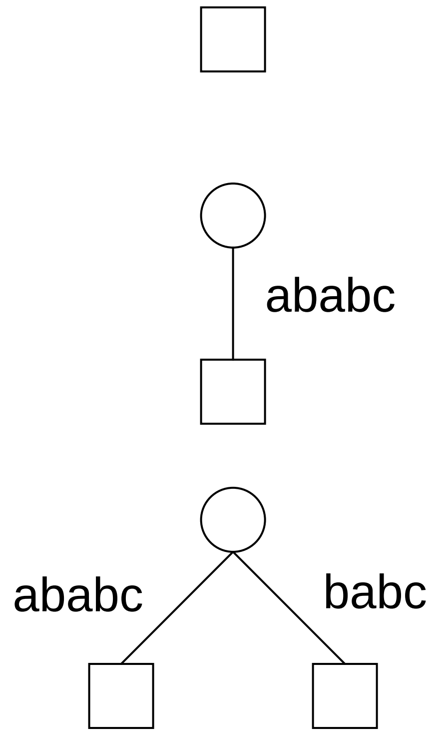
Example: σ = ababc

suf_3 = abc
$head_3$ = ab
$tail_3$ = c

T_0 =

T_1 =

T_2 =

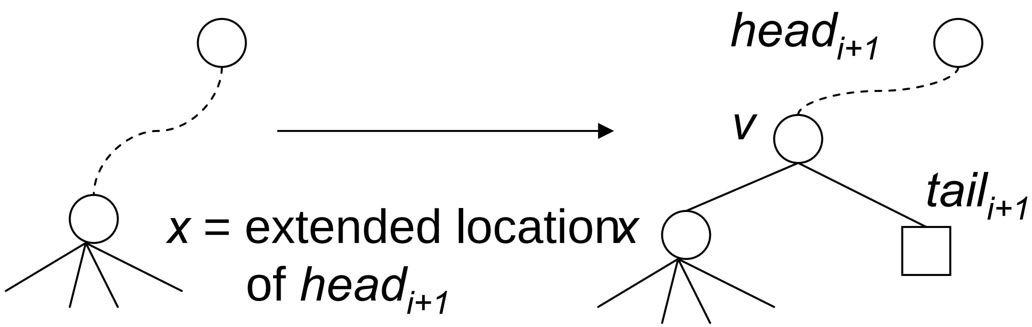


Naive suffix tree construction

T_{i+1} can be constructed from T_i as follows:

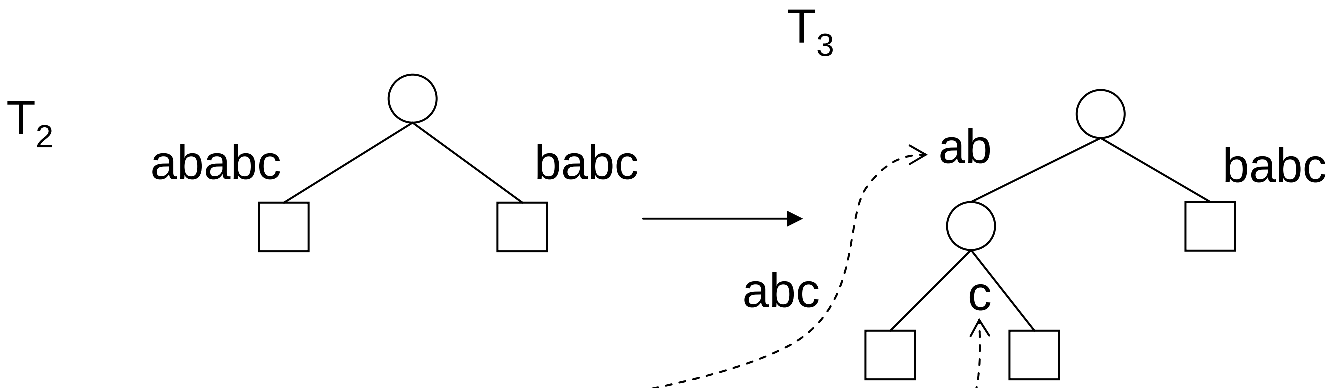
1. Determine the extended location of $head_{i+1}$ in T_i and split the last edge leading to this location into two new edges by inserting a new node.

2. Insert a new leaf as location for suf_{i+1} .



Naive suffix tree construction

Example: σ = ababc



$head_3$ = ab

$tail_3$ = c

Naive suffix tree construction

Algorithm *suffix-insertion*

Input: tree T_i and suffix suf_{i+1}

Output: tree T_{i+1}

v := root of T_i

$j := i$

repeat

    find child w of v with $\sigma_{w.l} = \sigma_{j+1}$

    $k := w.l - 1;$

    while $k < w.u$ **and** $\sigma_{k+1} = \sigma_{j+1}$ **do**

        $k := k + 1$; j $:= j + 1$

    end while

Naive suffix tree construction

        if $k = w.u$ **then** $v := w$

until k <$w.u$ or w = nil

/* v is the contracted location of $head_{i+1}$ */

insert the location of $head_{i+1}$ and $tail_{i+1}$ below v into T_i

Running time of *suffix-insertion* : O(   )

Total time required for the naive construction: O(   )

The algorithm *M*

(Mc Creight, 1976)

Idea: Extended location of $head_{i+1}$ in T_i is determined in constant amortized time. (Additional information required!)

When the extended location of $head_{i+1}$ in T_i has been found:
Creating a new node and splitting an edge takes O(1) time.

Theorem 1

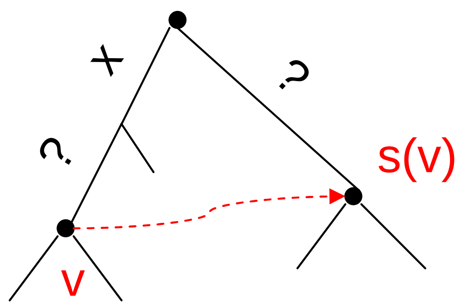
Algorithm *M* constructs a suffix tree for σ with $|\sigma|$ leaves and at most $|\sigma|$ - 1 internal nodes in time $O(|\sigma|)$.

Suffix links

Definition:

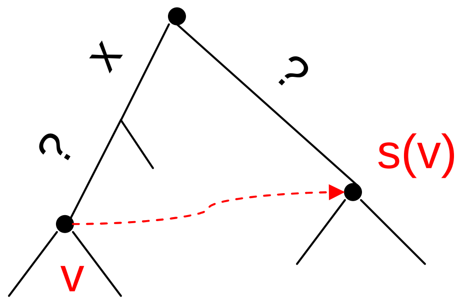
Let *x*? be an arbitrary string where *x* is a single character and ? some (possibly empty) substring.
For an internal node *v* with edge labels *x*? the following holds:

> If there exists a node *s*(*v*) with edge label ?, then there is a pointer from *v* to *s*(*v*) which is called a suffix link.



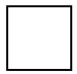
Suffix links

The idea is the following:
By following the suffix links, we do not have to start each search for a splitting point at the root node. Instead, we can use the suffix links in order to determine these nodes more efficiently, i.e. in constant amortized time.

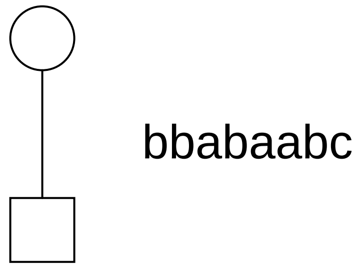


Suffix tree: example

T_0 =



T_1 =

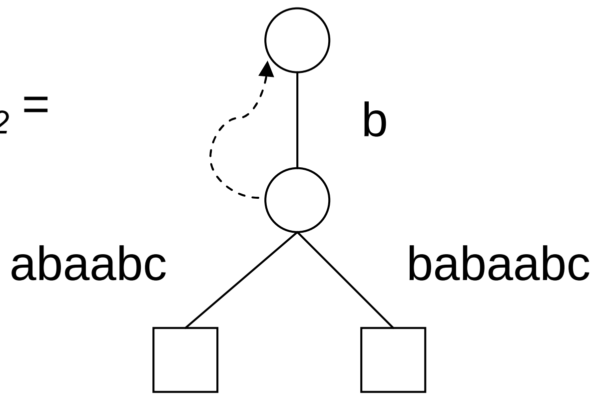


suf_1 = bbabaabc

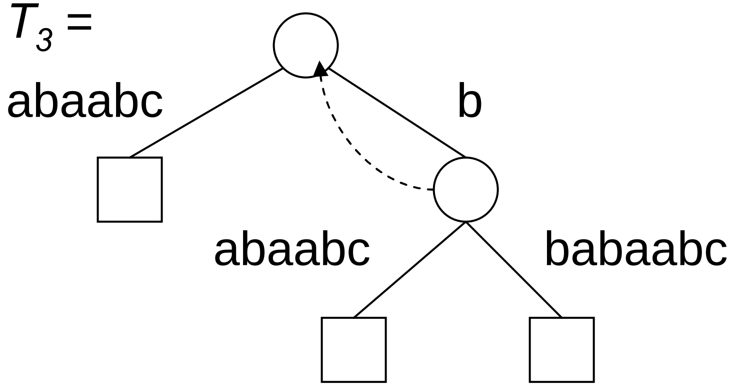
suf_2 = babaabc
$head_2$ = b

Suffix tree: example

T_2 =



T_3 =

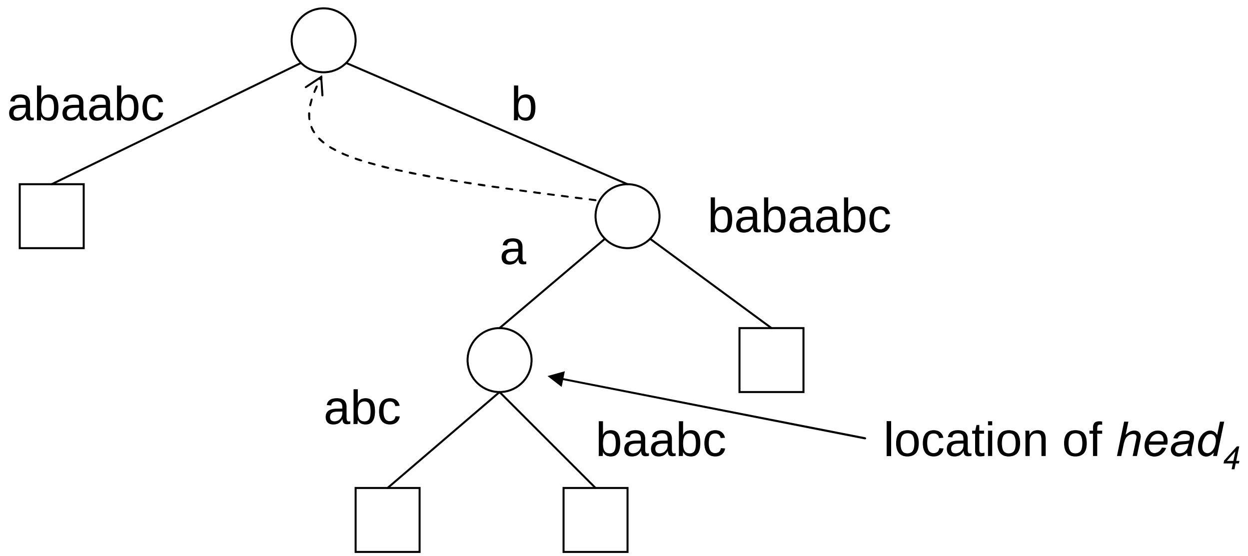


suf_3 = abaabc
$head_3 = \varepsilon$

suf_4 = baabc
$head_4$ = ba

Suffix tree: example

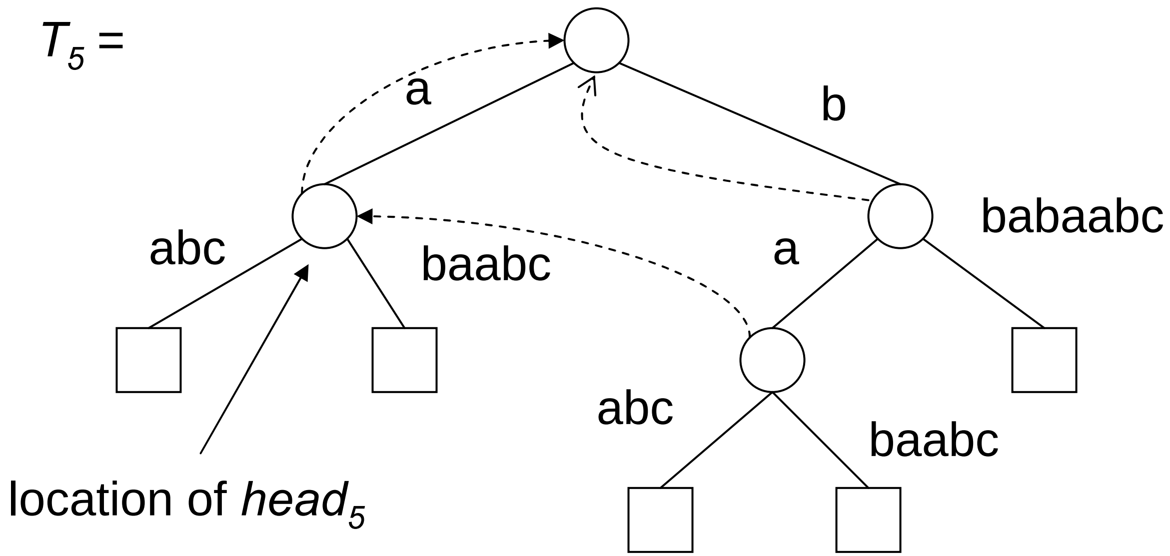
T_4 =



suf_5 = aabc

$head_5$ = a

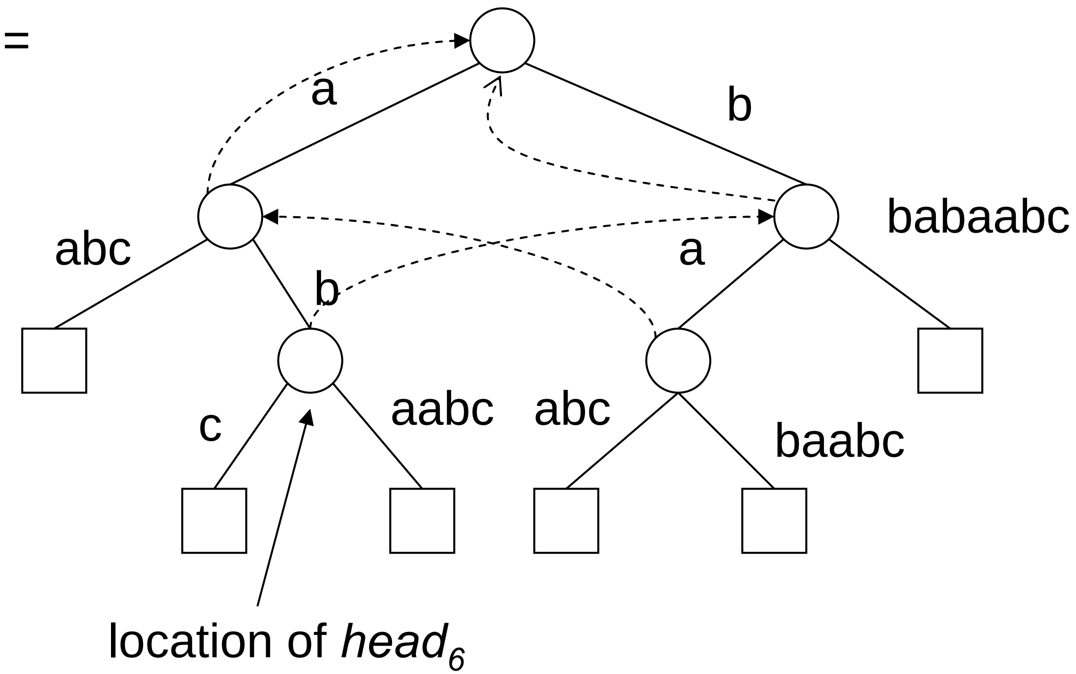
Suffix tree: example



suf_6 = abc
$head_6$ = ab

Suffix tree: example

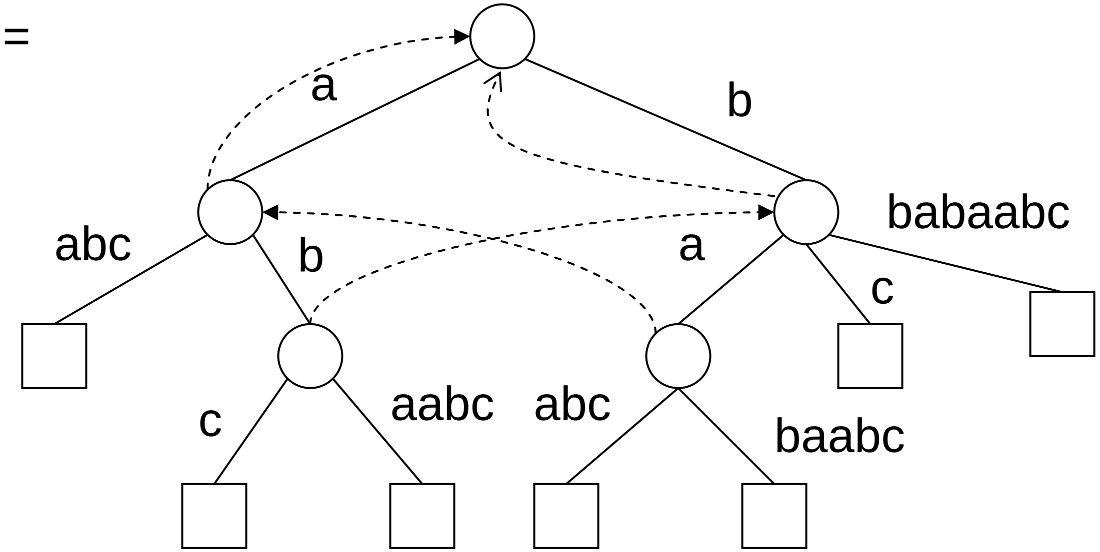
$T_6 =$



suf_7 = bc
$head_7$ = b

Suffix tree: example

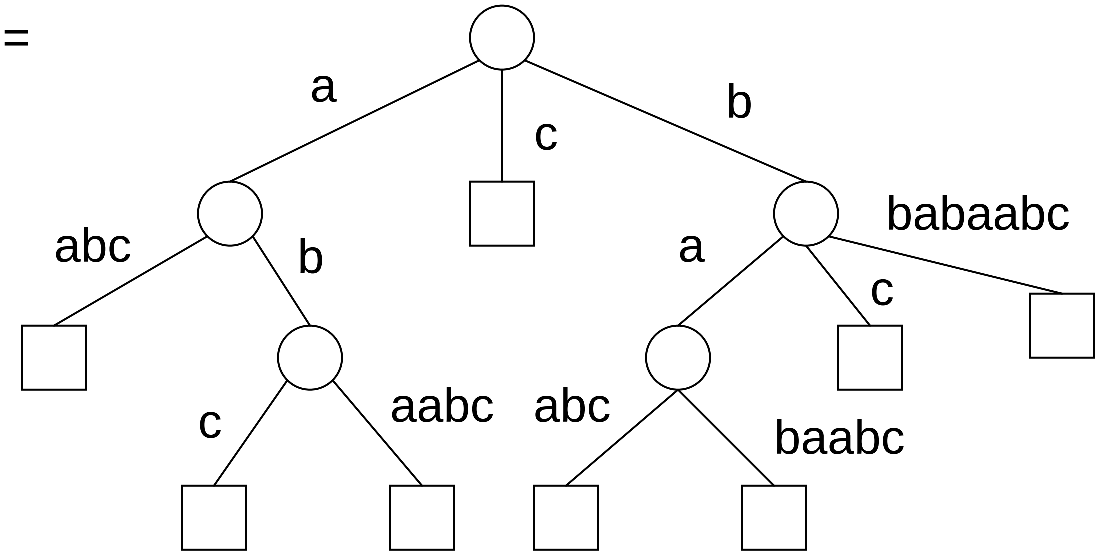
T_7 =



suf_8 = c

Suffix tree: example

$T_8 =$



Suffix tree: application

Usage of a suffix tree *T*:

1 Search for a string α:
Follow the path with edge labels α (takes $O(|\alpha|)$ time).
leaves of the subtree $\hat{=}$ occurrences of α

2 Search for the longest substring occurring at least twice:
Find the location of a substring with maximum weighted depth that is an internal node.

3 Prefix search:
All occurrences of strings with prefix α are represented by the nodes of the subtree rooted the location of α in *T*.

Suffix tree: application

4 Range search for [α, β] :