Recuperació de la informació

algorithmics genetics group

- Modern Information Retrieval (1999)
  Ricardo-Baeza Yates and Berthier Ribeiro-Neto

- Flexible Pattern Matching in Strings (2002)
  Gonzalo Navarro and Mathieu Raffinot

- Algorithms on strings (2001)
  M. Crochemore, C. Hancart and T. Lecroq

- http://www-igm.univ-mlv.fr/~lecroq/string/index.html

String Matching

String matching: definition of the problem (text,pattern)

- Exact matching: depends on what we have: text or patterns
  - The patterns ---> Data structures for the patterns
    - 1 pattern ---> The algorithm depends on |p| and |Σ|
    - k patterns ---> The algorithm depends on k, |p| and |Σ|
  - Extensions
  - Regular Expressions
  - The text ----> Data structure for the text (suffix tree, ...)
- Approximate matching:
  - Dynamic programming
  - Sequence alignment (pairwise and multiple)
  - Sequence assembly: hash algorithm
- Probabilistic search: Hidden Markov Models

Index

algorithmics genetics group

1a. Part: Suffix trees

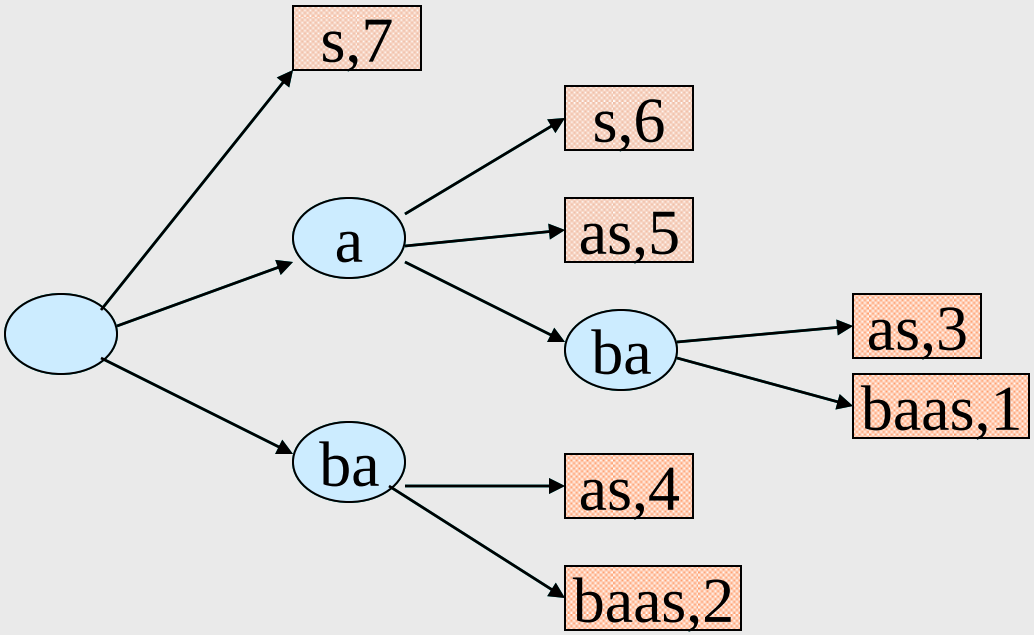
Algorithms on strings, trees and sequences,
Dan Gusfield
Cambridge University Press

2a. Part: Suffix arrays

Suffix-arrays: a new method for on-line string searches,
G. Myers, U. Manber

Suffix trees

Given string abababaas:



Suffixes:

7: s

6: as

5: aas

3: abaas

1: ababaas

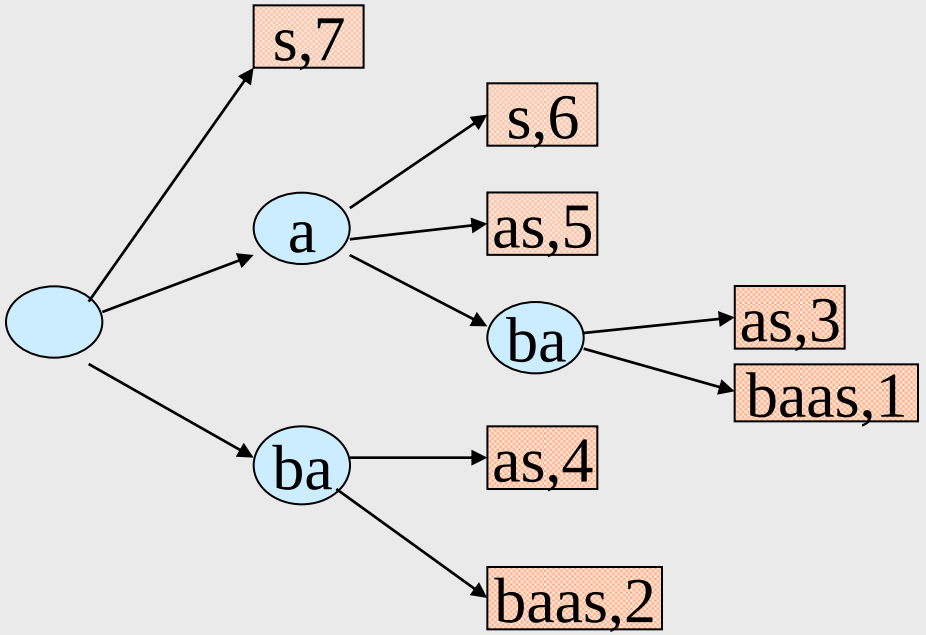
4: baas

2: babaas

What kind of queries?

Applications of Suffix trees

1. Exact string matching
   - Does the sequence ababaas contain any ocurrence of patterns abab, aab, and ab?



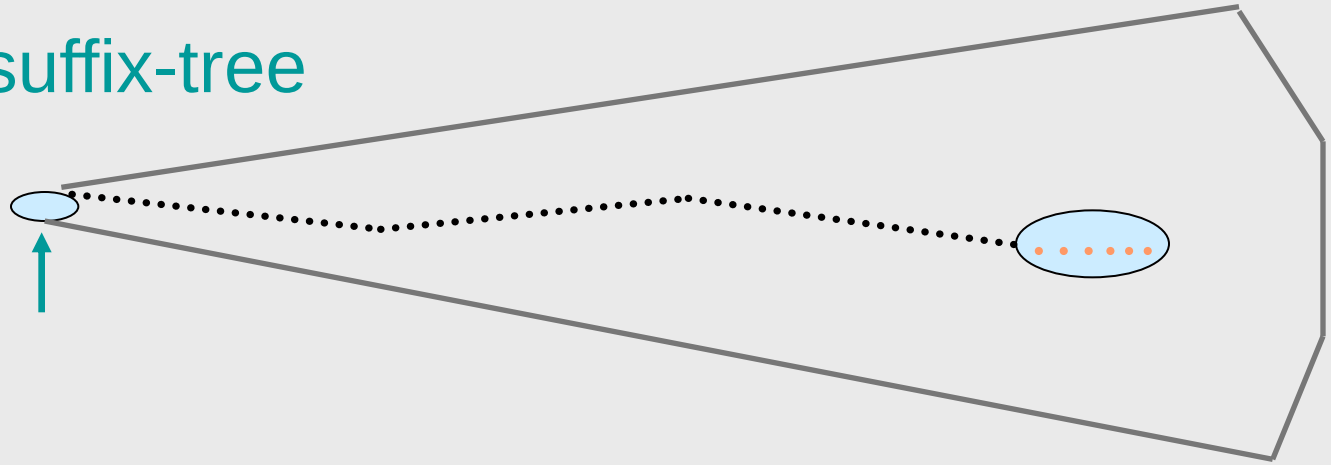
Quadratic insertion algorithm

algorithmics genetics group

Invariant Properties:

Given the string

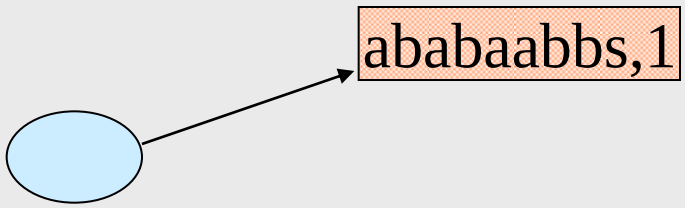
and the suffix-tree



P1: the leaves of suffixes from α have been inserted

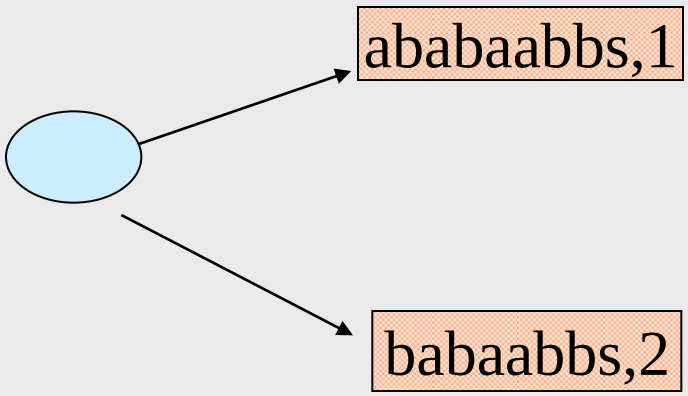
Quadratic insertion algorithm

Given the string ababaabbs



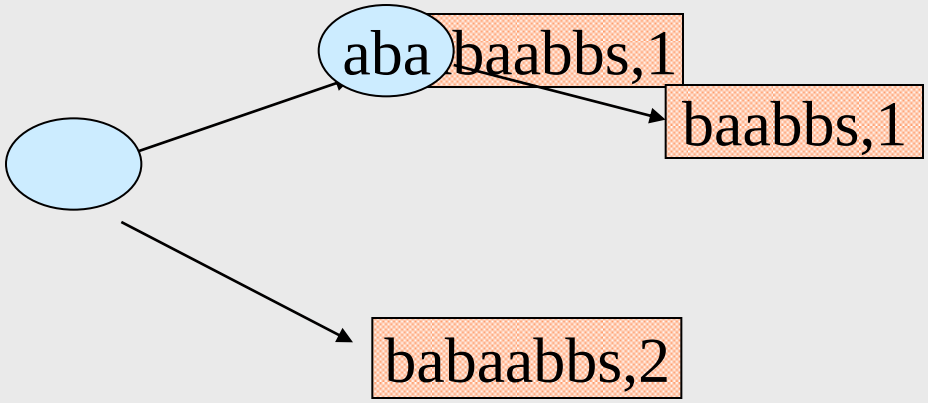
Quadratic insertion algorithm

Given the string ababaabbs



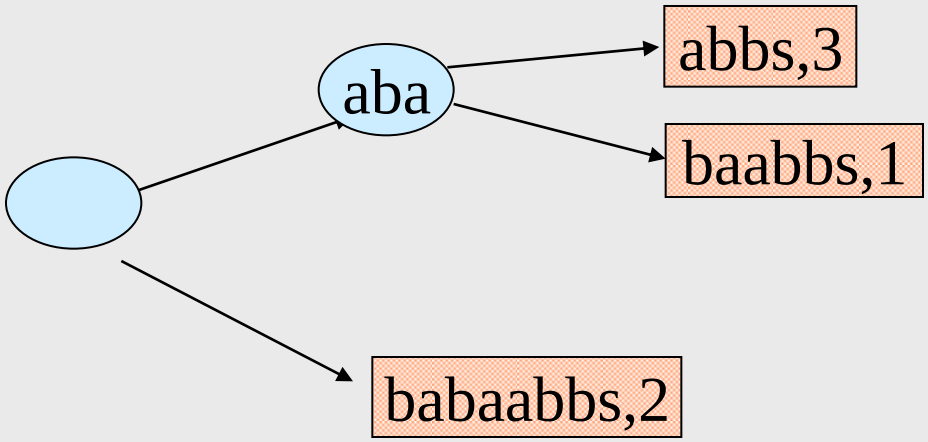
Quadratic insertion algorithm

Given the string ababaabbs



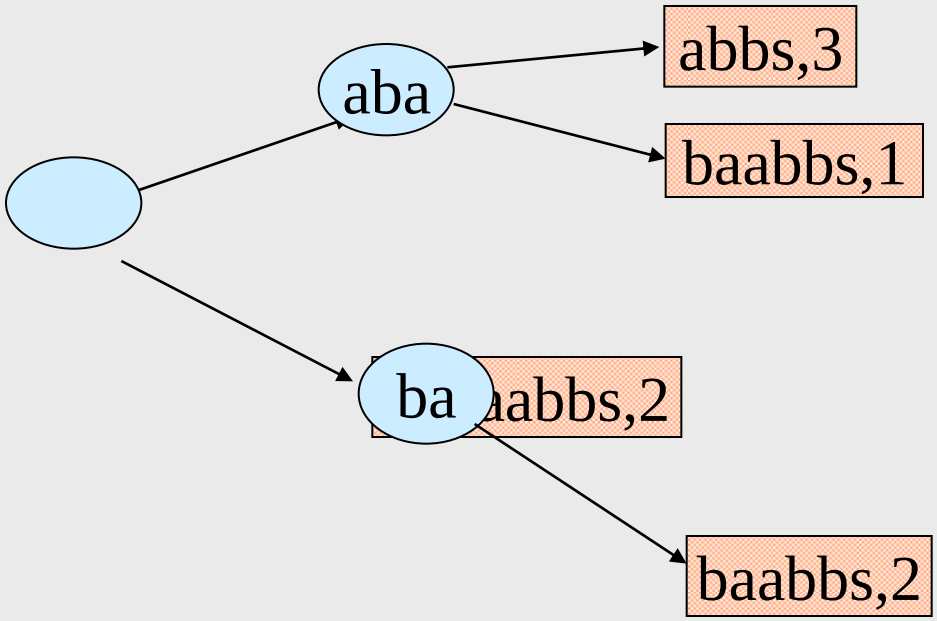
Quadratic insertion algorithm

Given the string ababaabbs



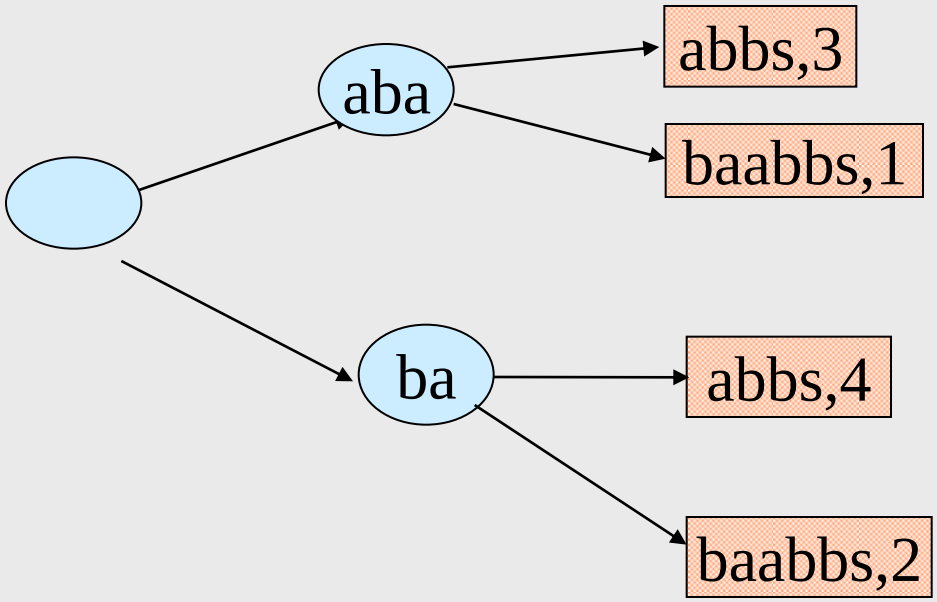
Quadratic insertion algorithm

Given the string ababaabbs



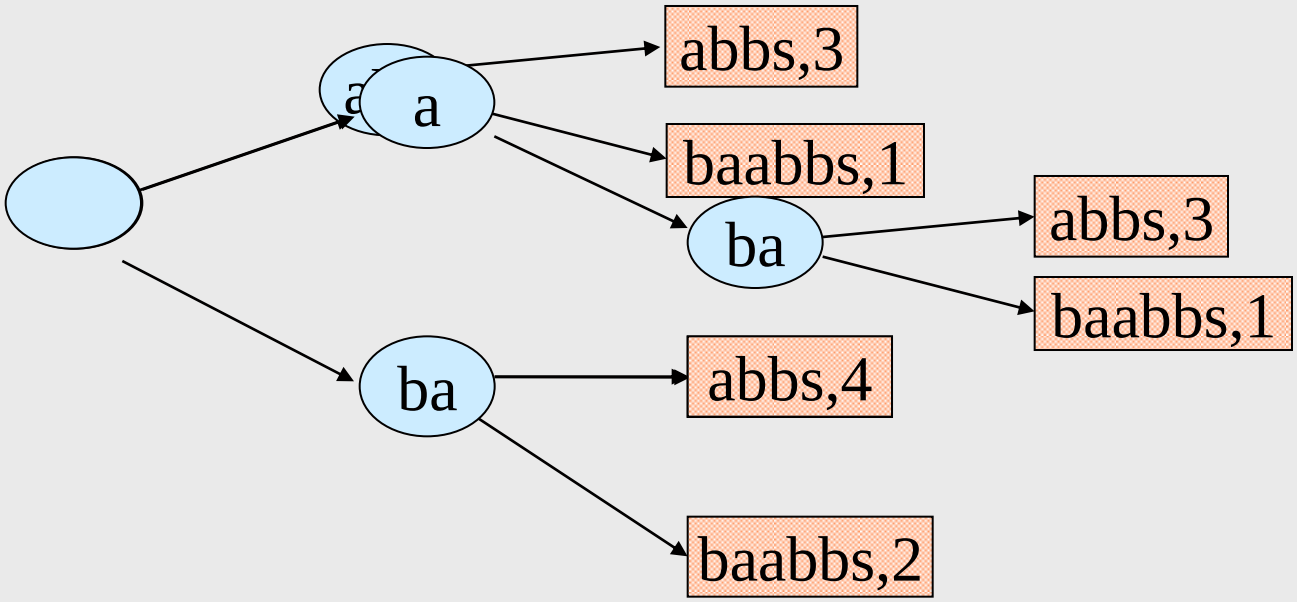
Quadratic insertion algorithm

Given the string ababaabbs



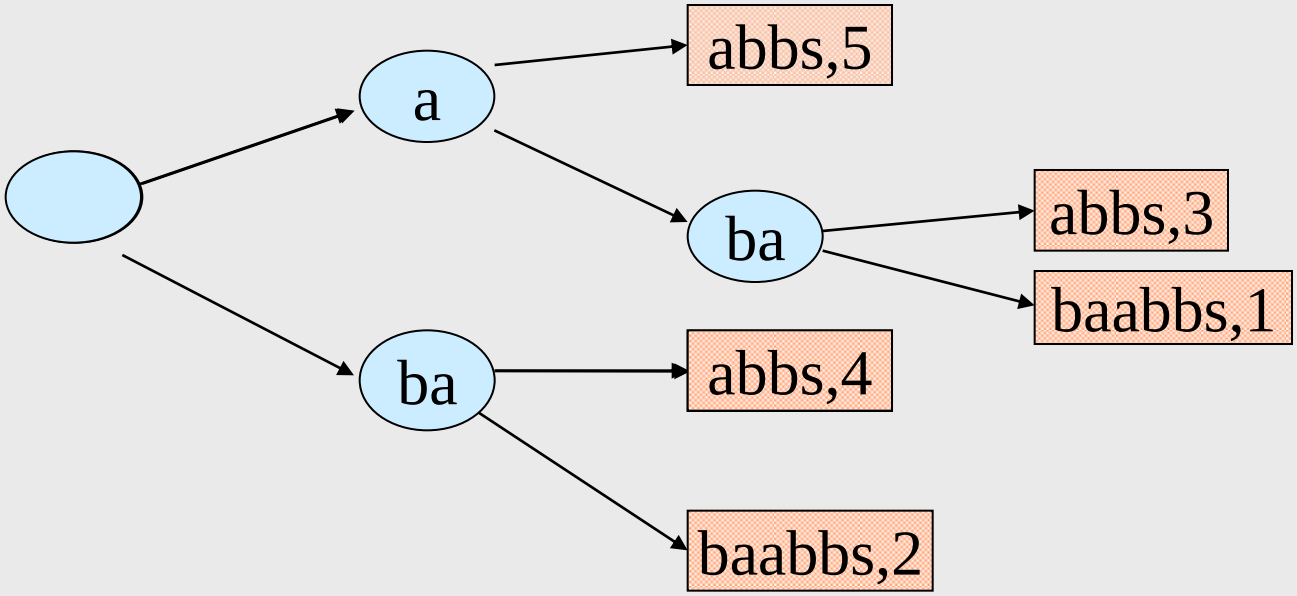
Quadratic insertion algorithm

Given the string ababaabbs



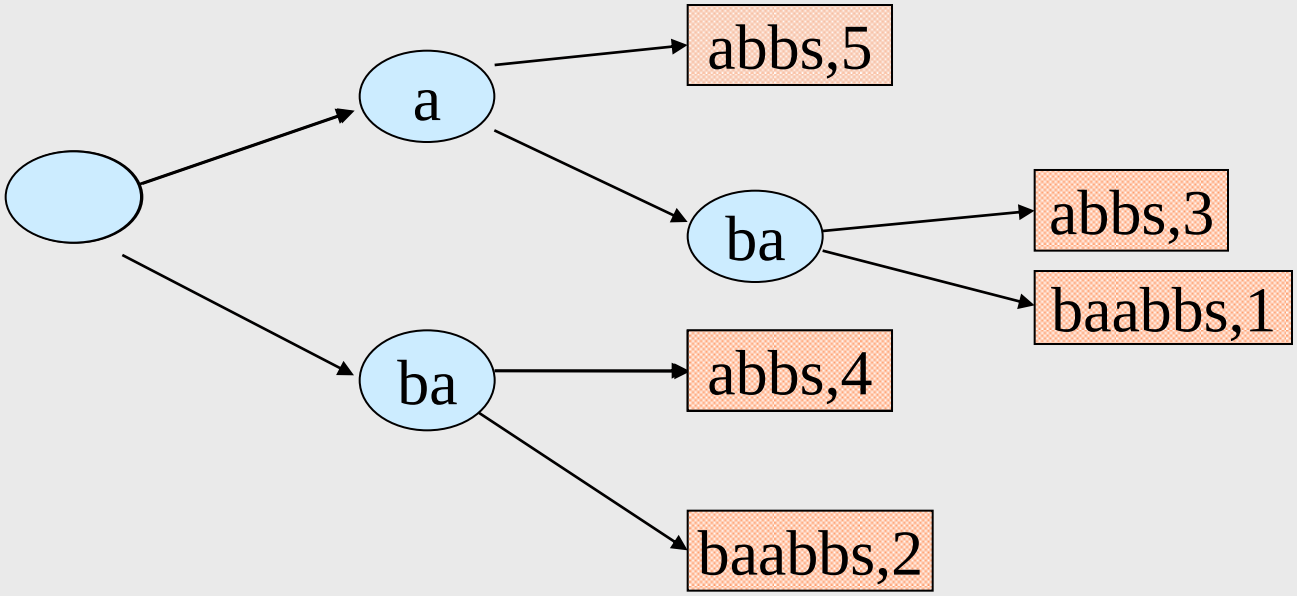
Quadratic insertion algorithm

Given the string ababaabbs



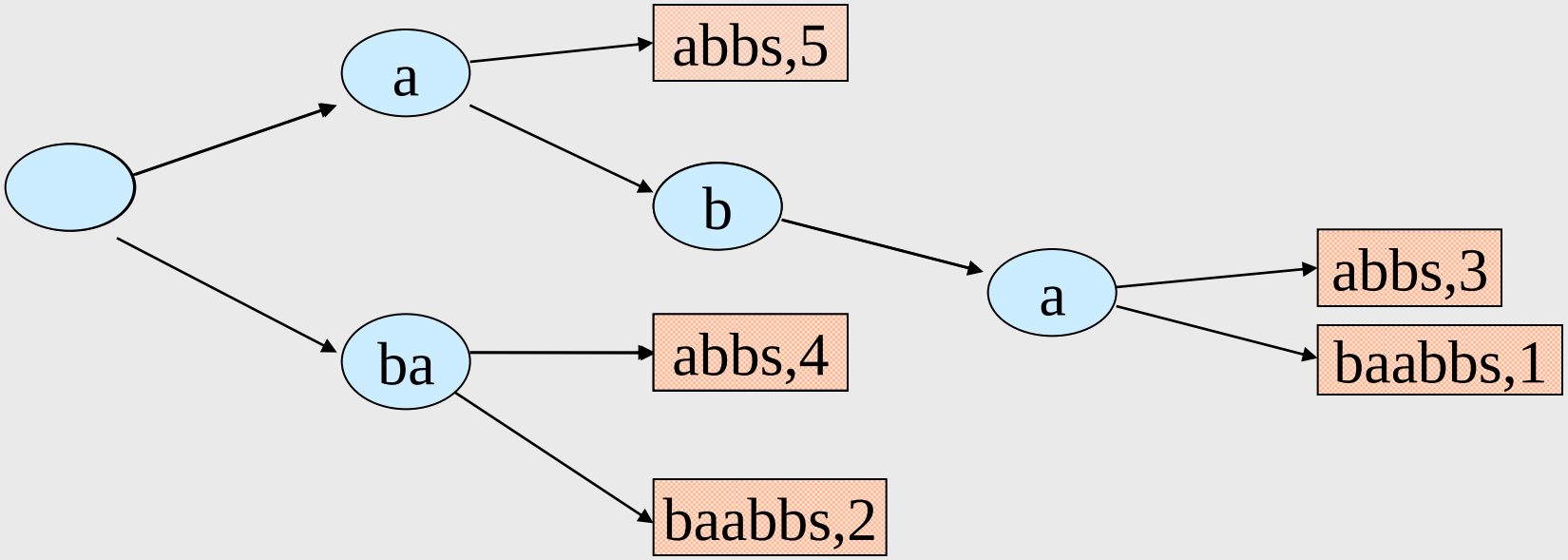
Quadratic insertion algorithm

Given the string ababaabbs



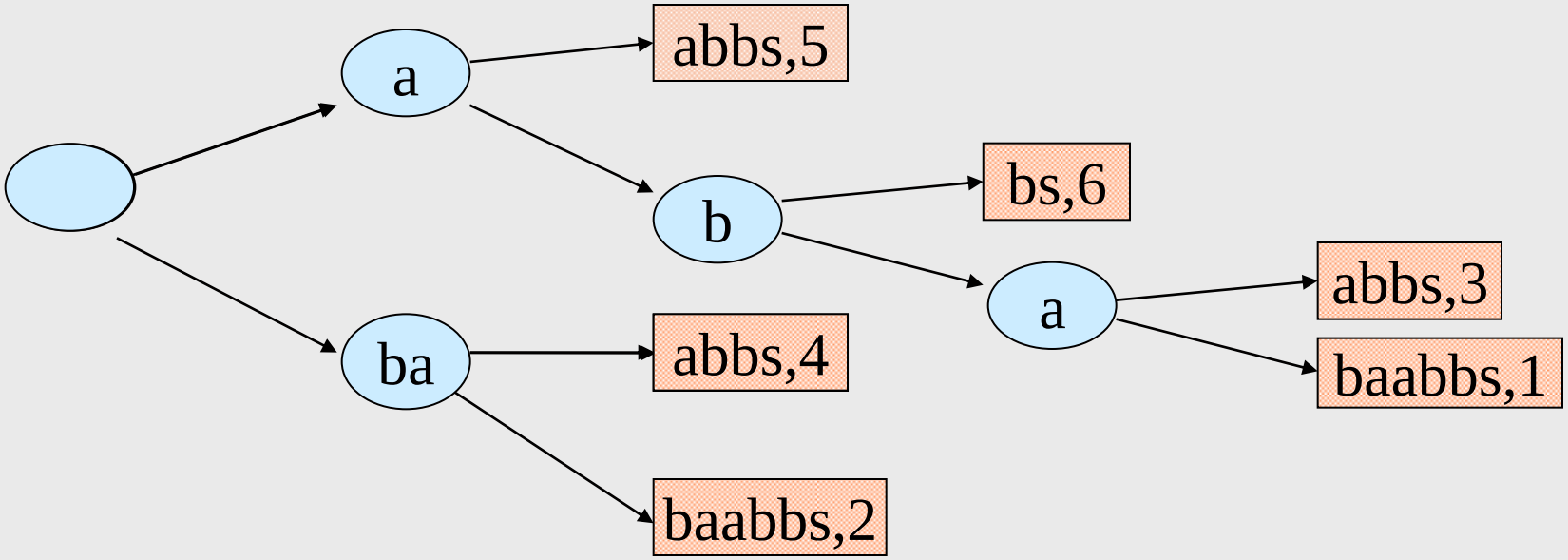
Quadratic insertion algorithm

Given the string ababaabbs



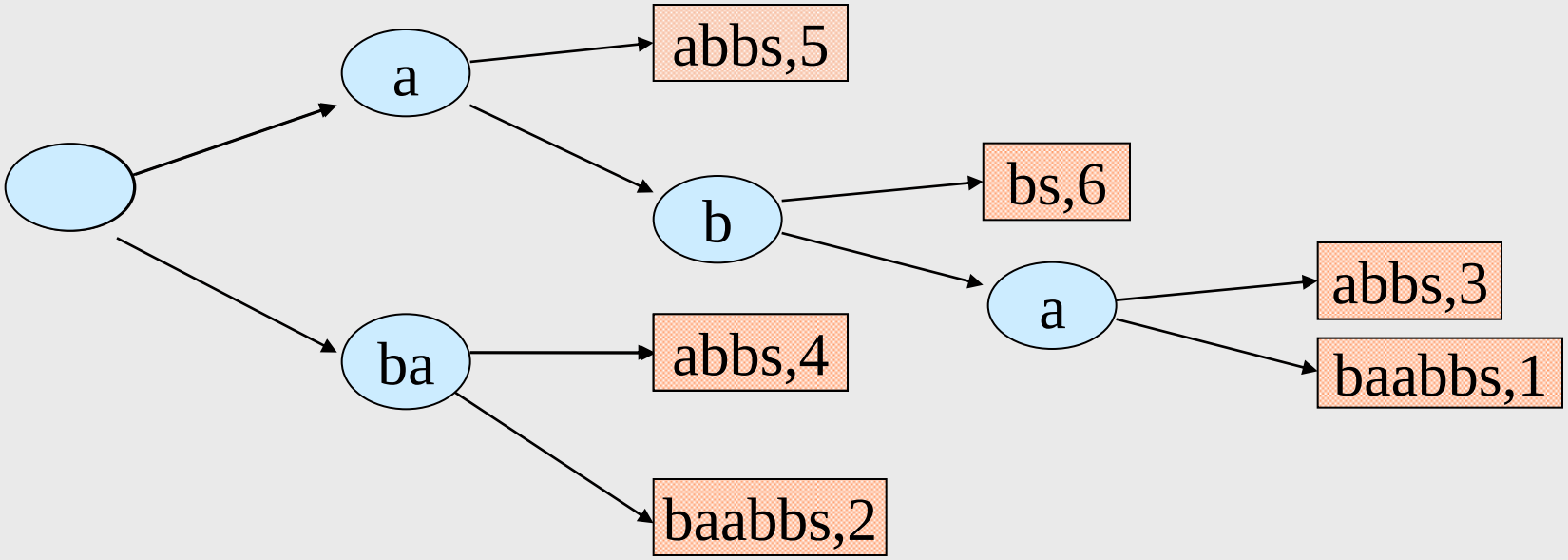
Quadratic insertion algorithm

Given the string ababaabbs



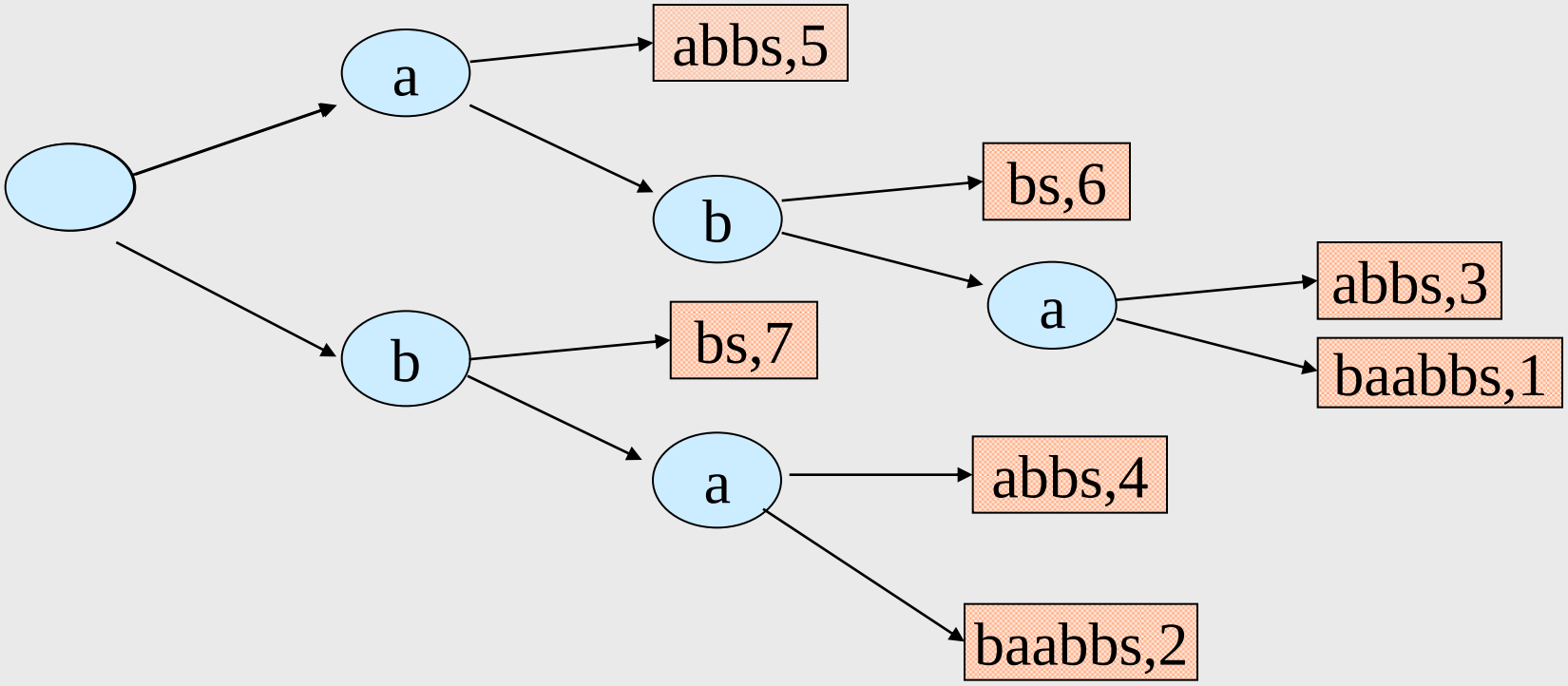
Quadratic insertion algorithm

Given the string ababaabbs



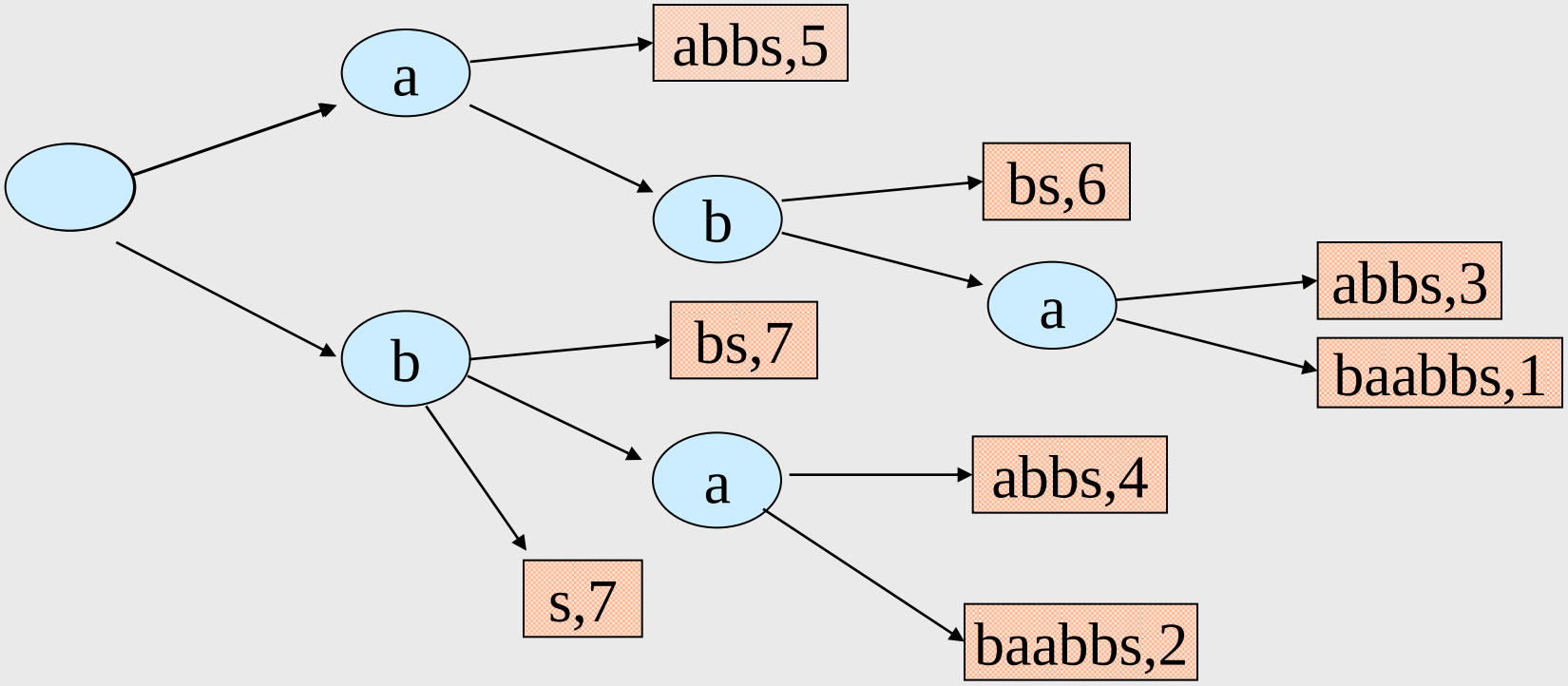
Quadratic insertion algorithm

Given the string ababaabbs



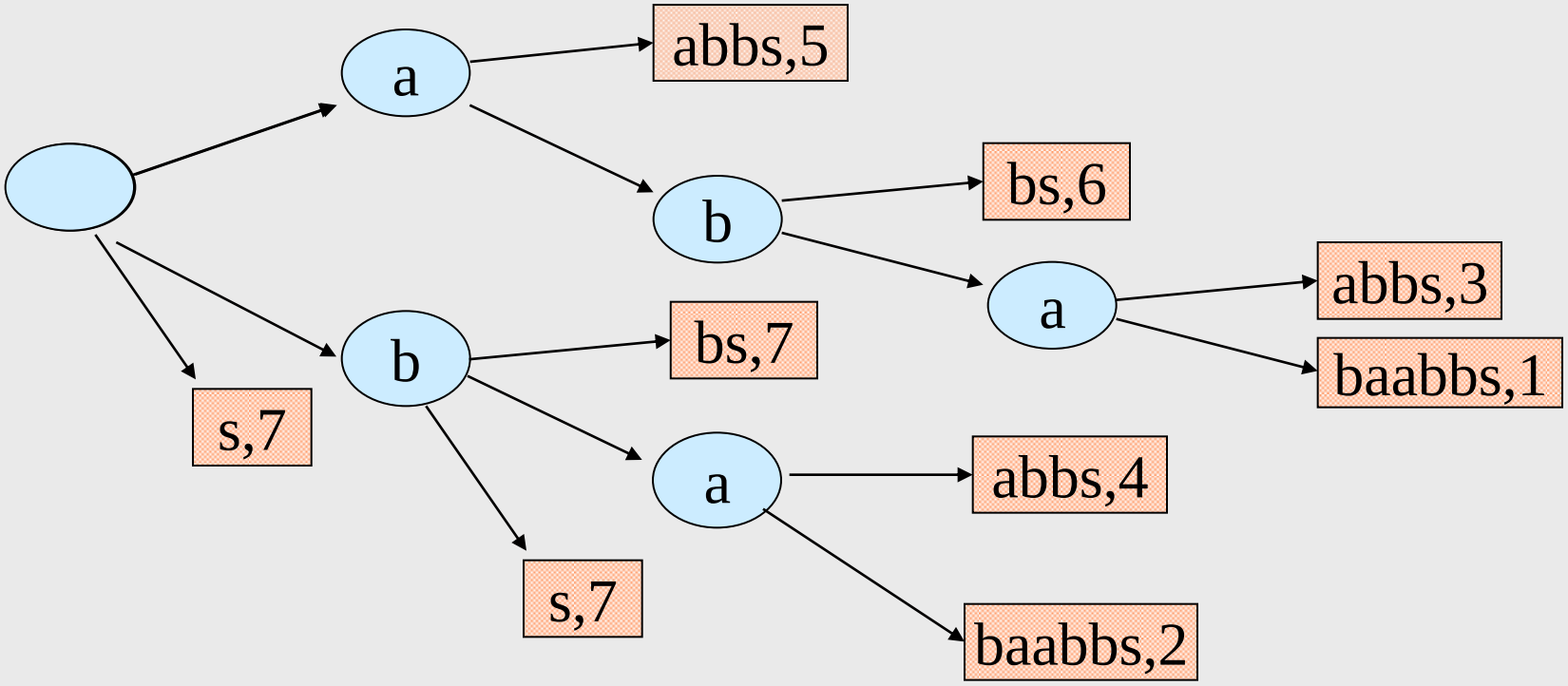
Quadratic insertion algorithm

Given the string ababaabbs



Quadratic insertion algorithm

Given the string ababaabbs



Generalizad suffix tree

algorithmics genetics group .

The suffix tree of many strings …

is called the generalized suffix tree …

and it is the suffix tree of the concatenation of strings.

For instance,

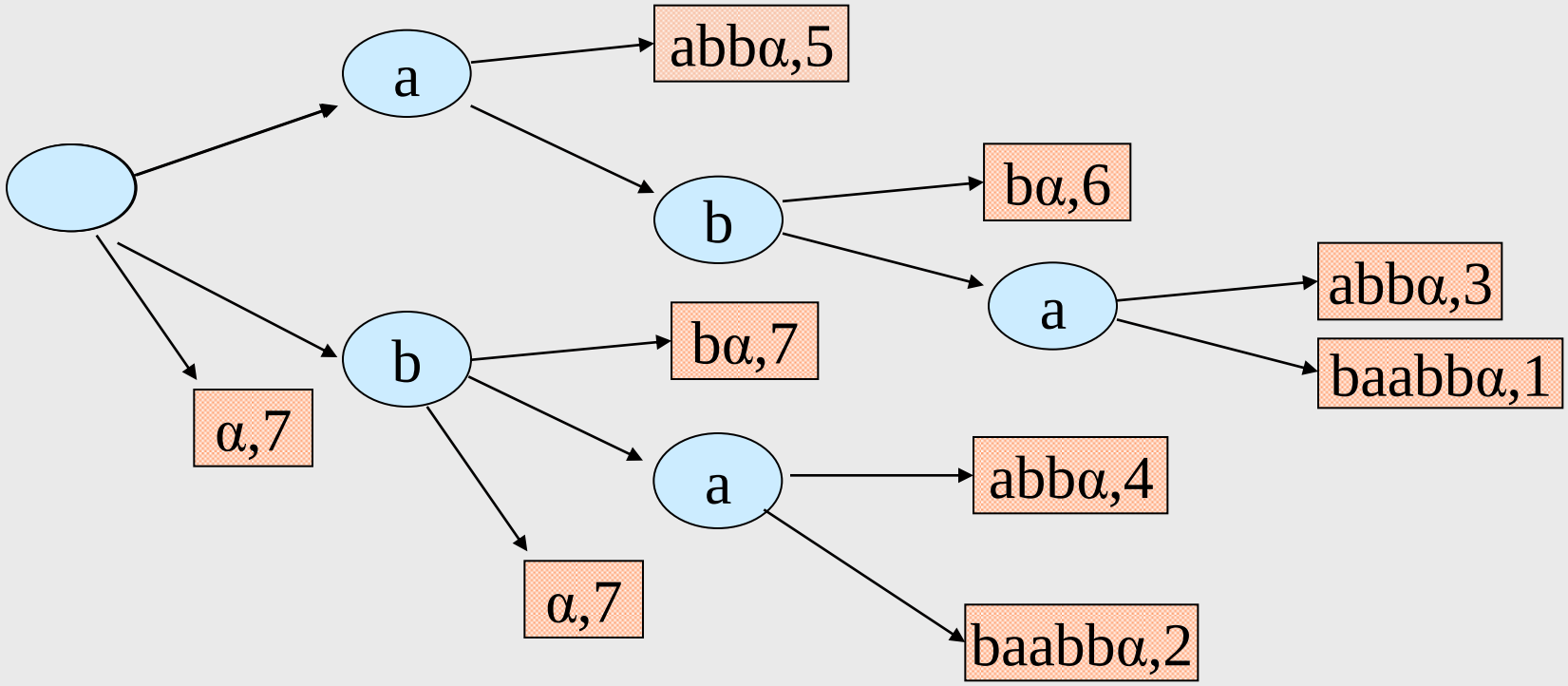
the generalized suffix tree of abababb and aabaat …

is the suffix tree of abababbαaabaatβ, :

Generalizad suffix tree

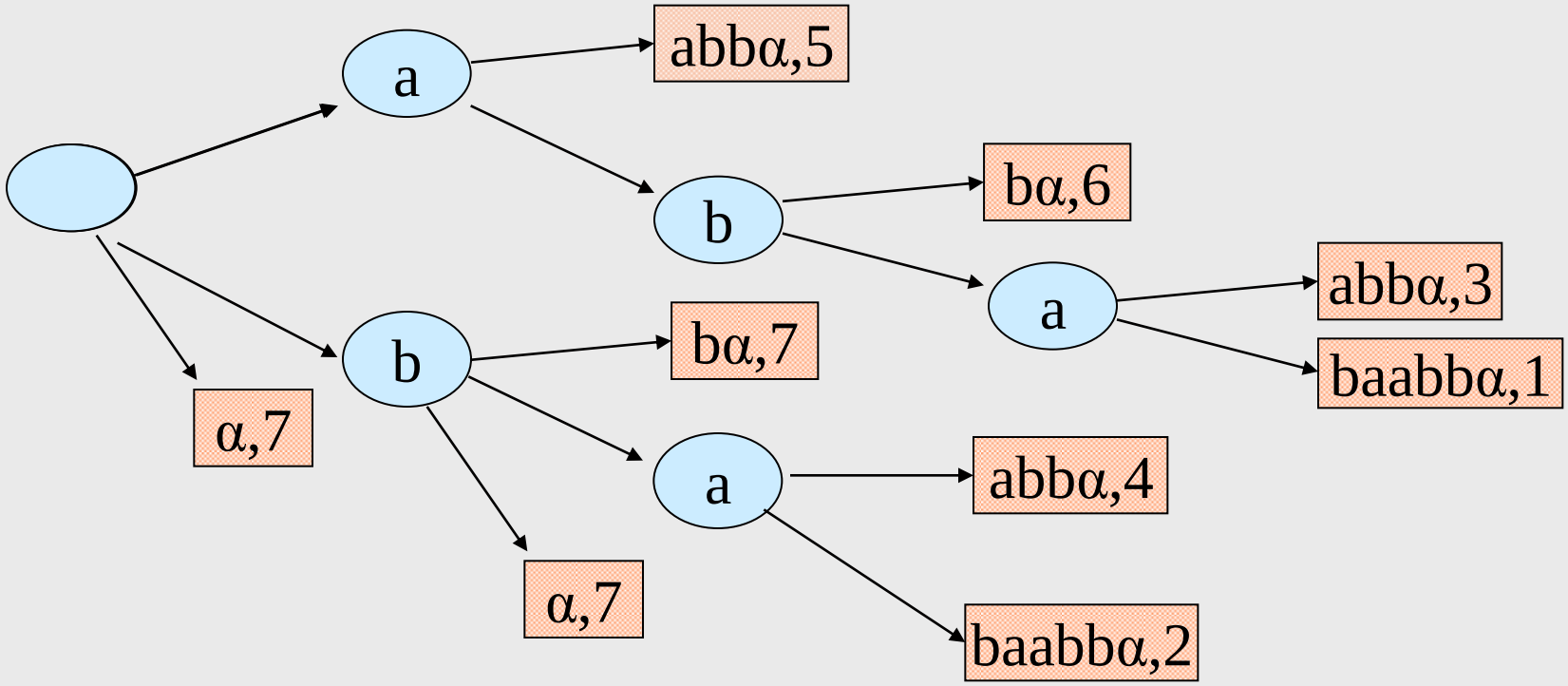
Construction of the suffix tree of abababaabbαaabaaβ :

Given the suffix tree of abababaabα :



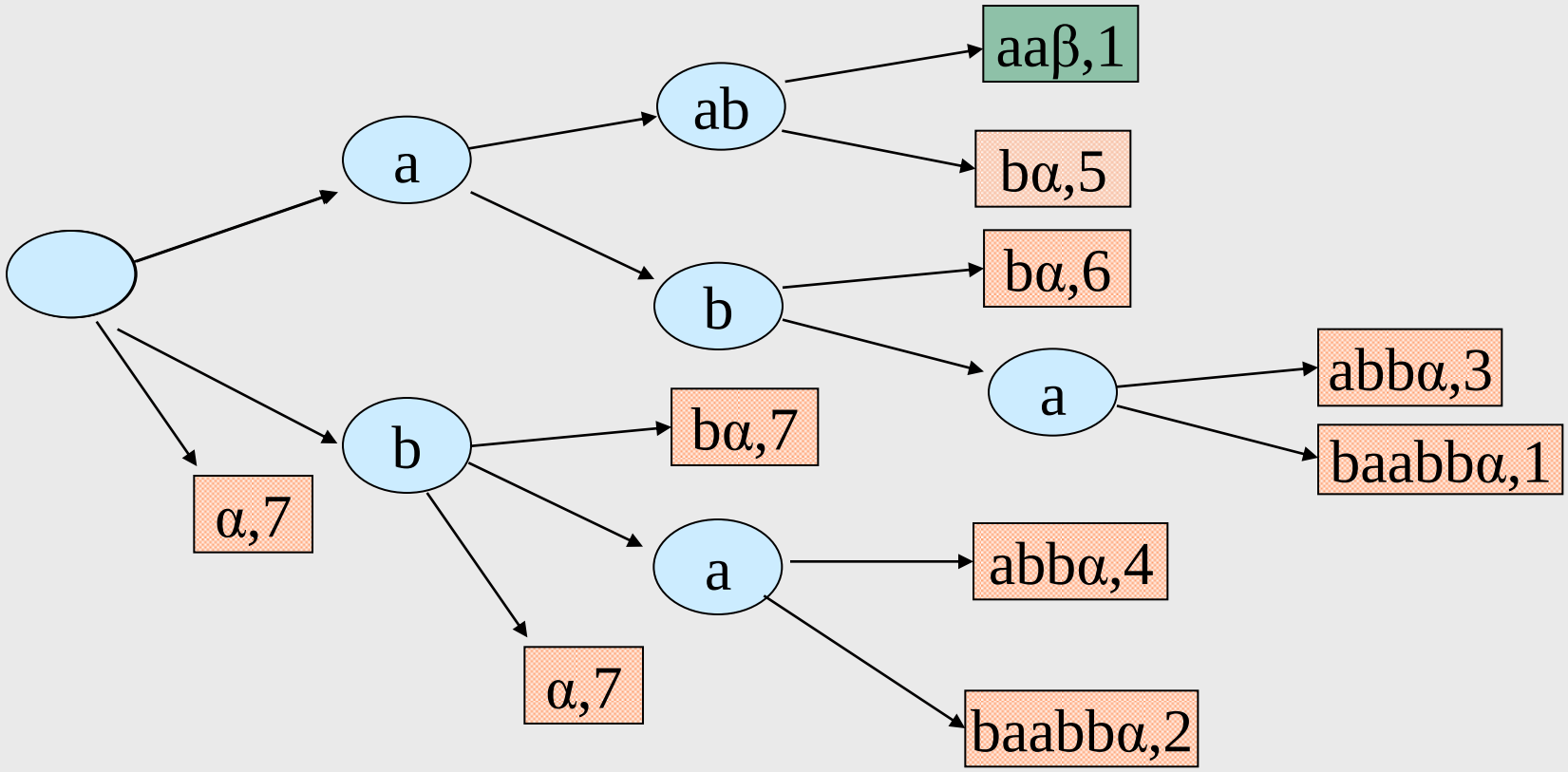
Generalizad suffix tree

Construction of the suffix tree of abababbαaabaaβ :



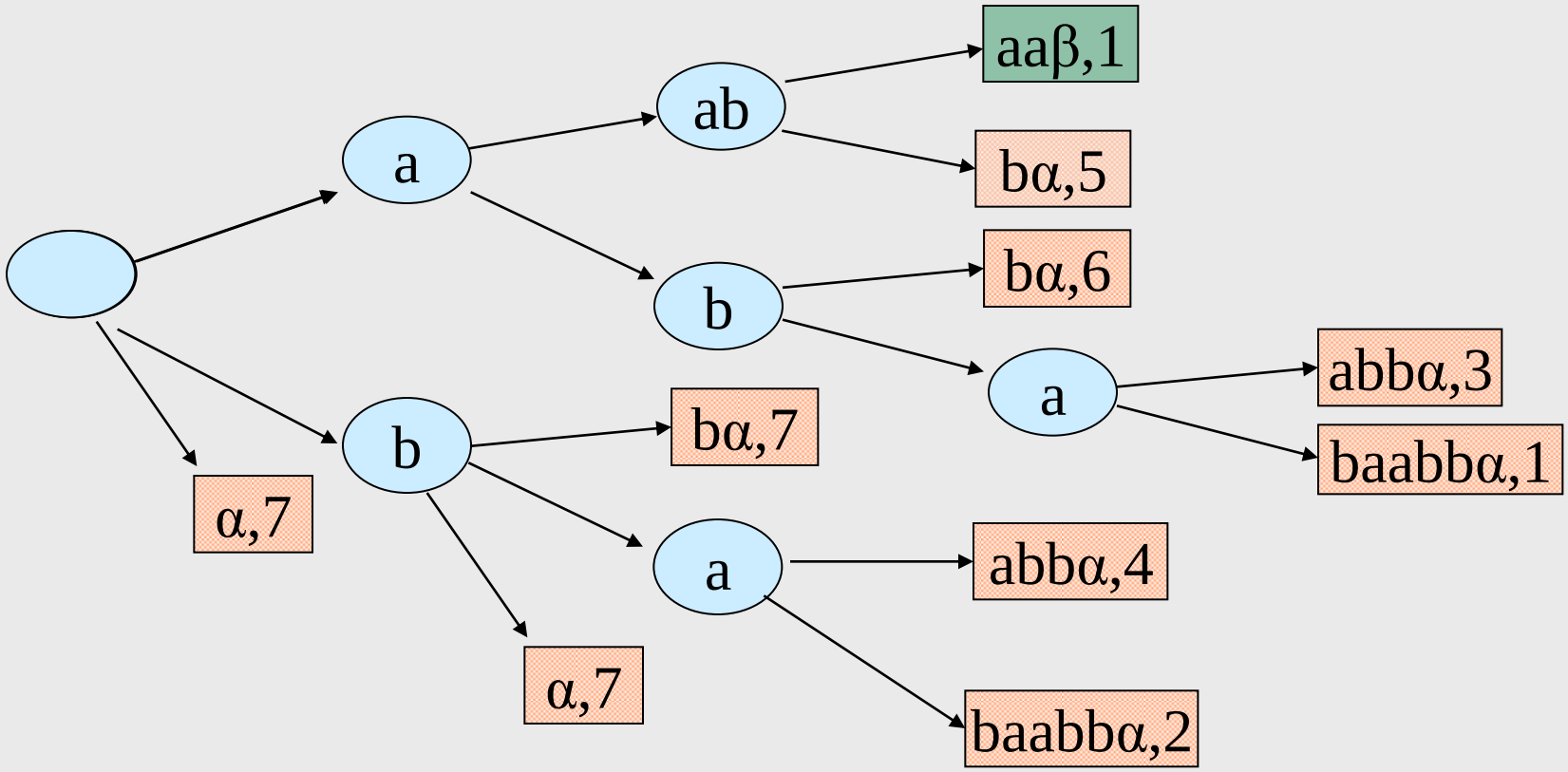
Generalizad suffix tree

Construction of the suffix tree of abababaabbαaabaaβ :



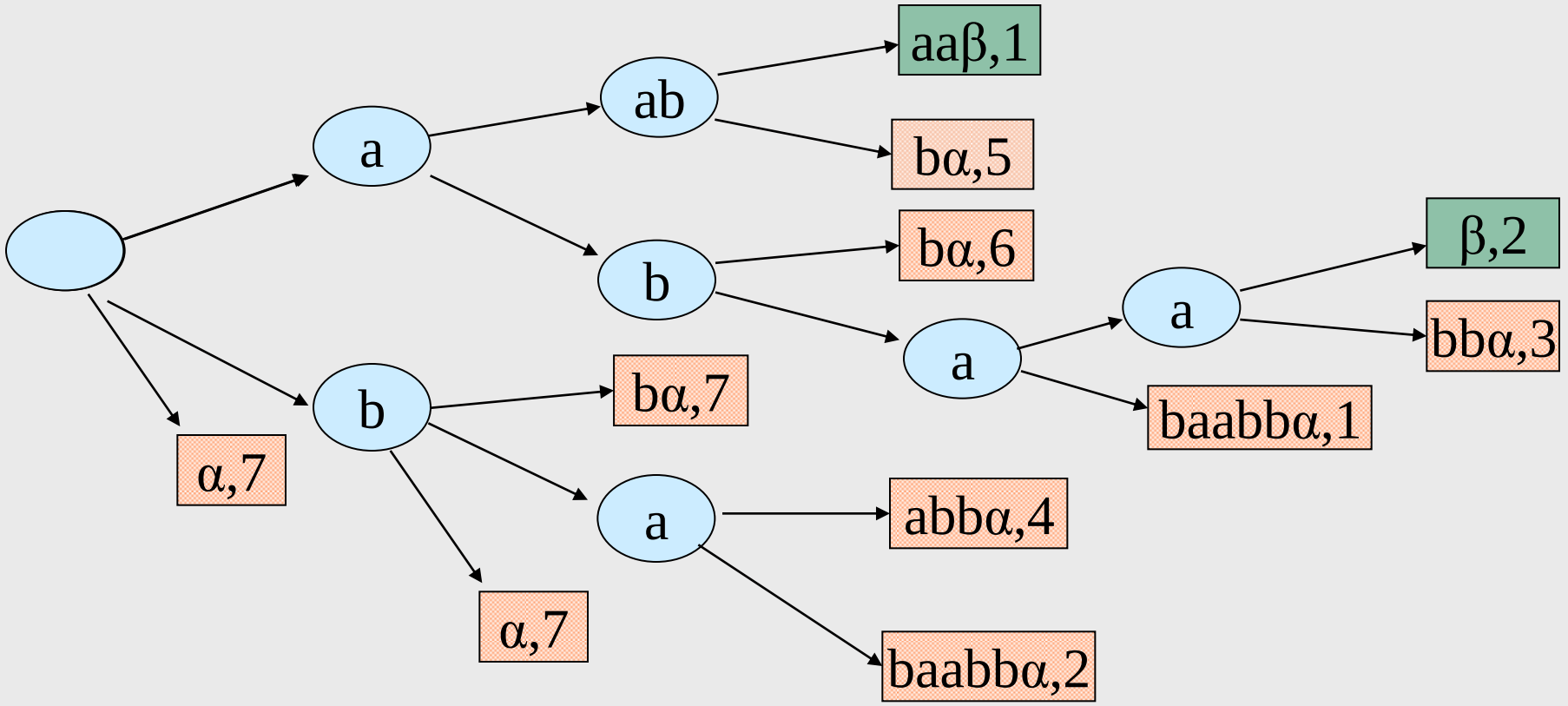
Generalizad suffix tree

Construction of the suffix tree of abababaabbαaabaaβ :



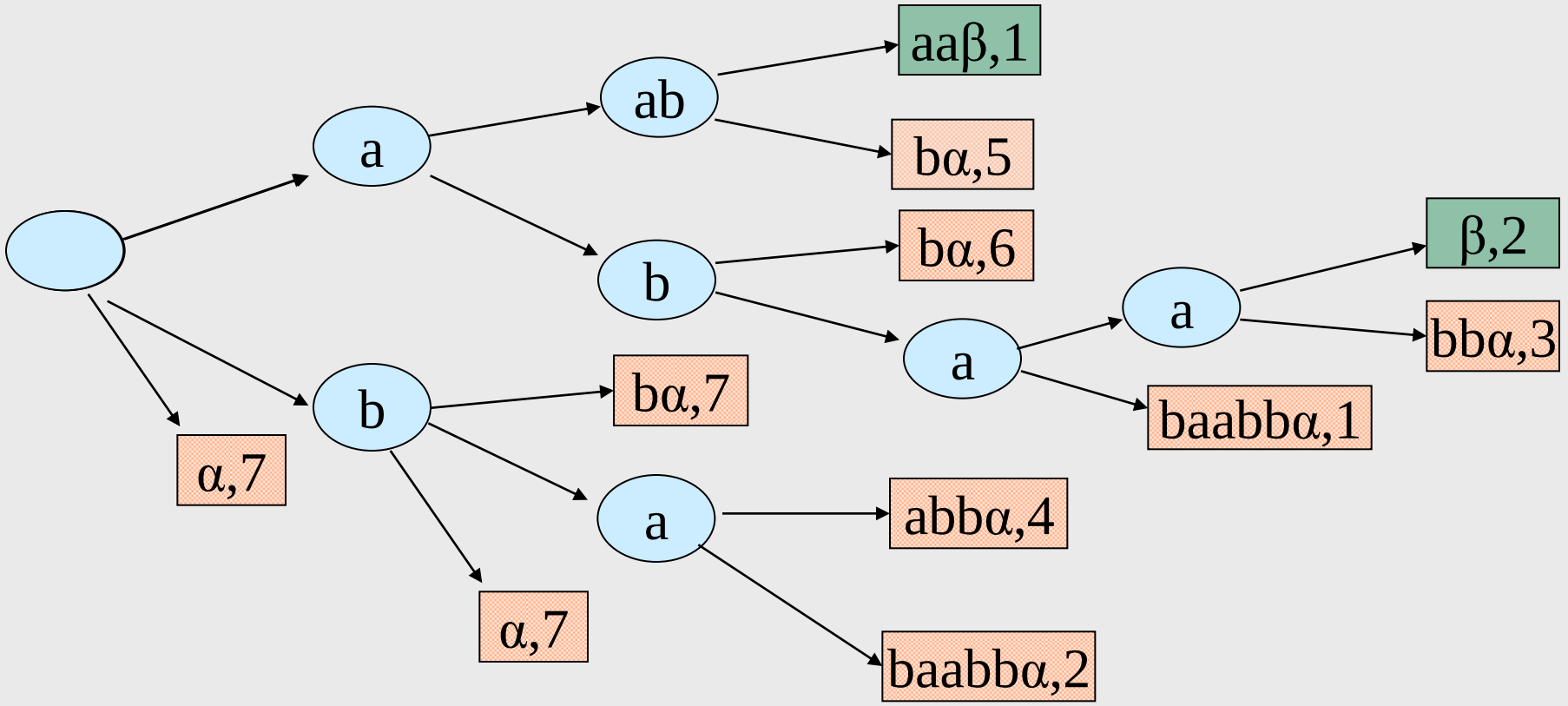
Generalizad suffix tree

Construction of the suffix tree of abababaabbαaabaaβ :



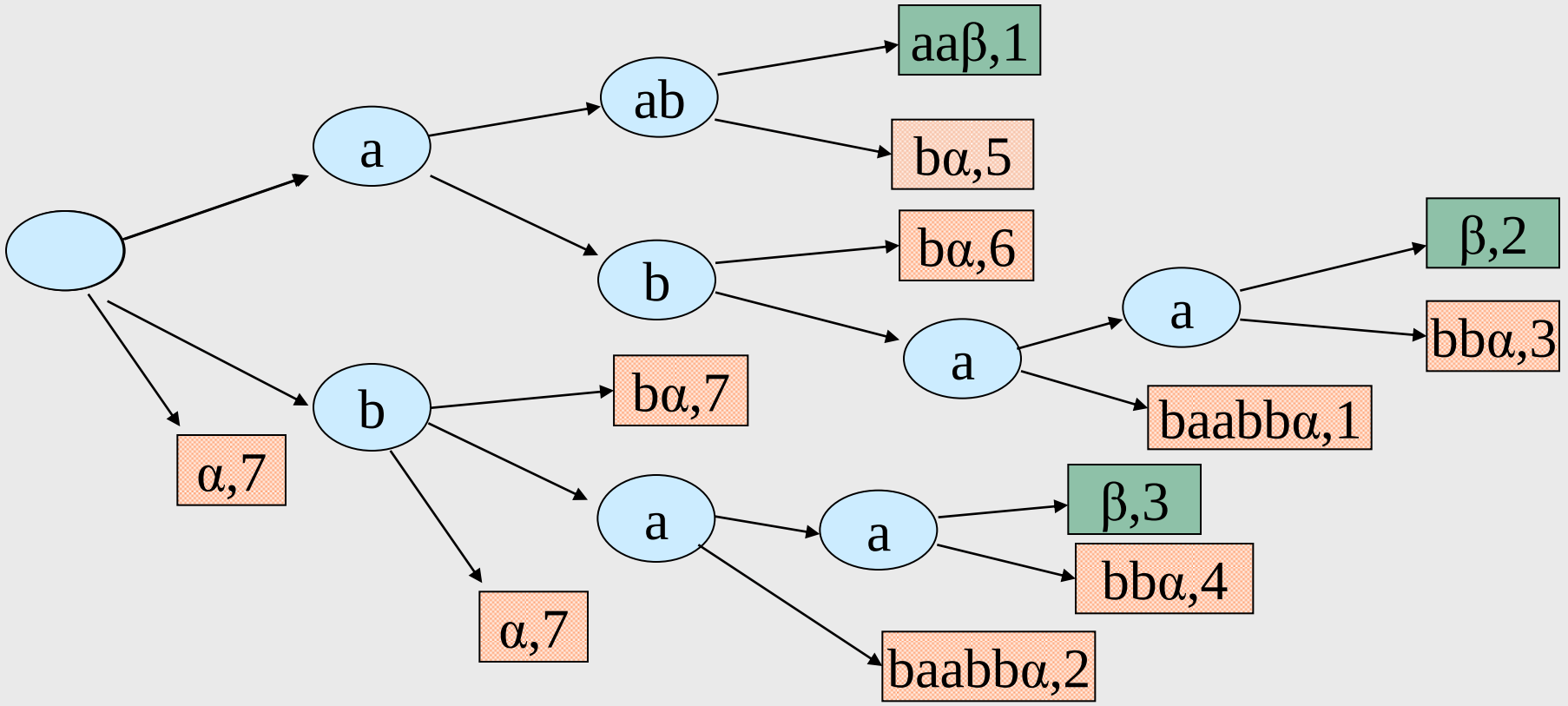
Generalizad suffix tree

Construction of the suffix tree of abababaabbαaabaaβ :



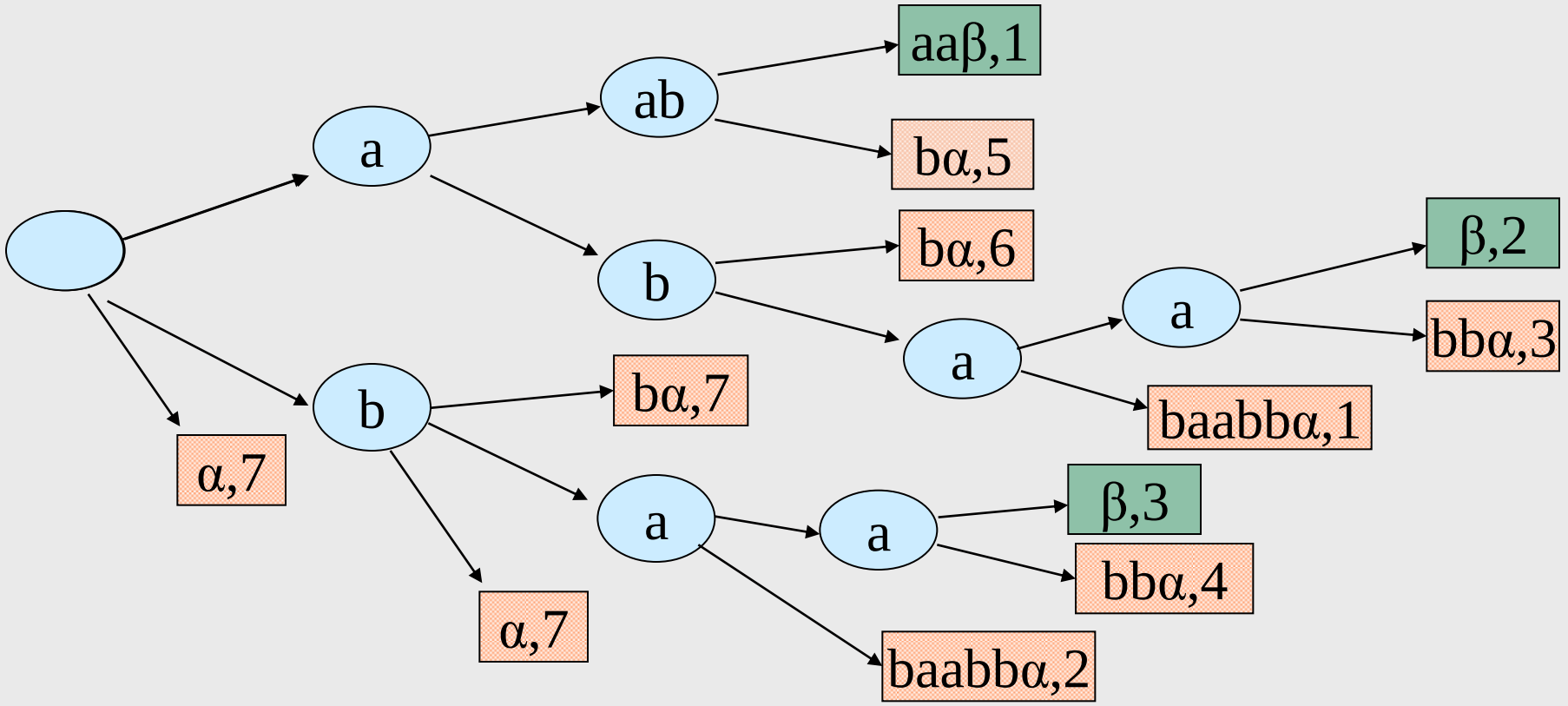
Generalizad suffix tree

Construction of the suffix tree of abababaabbαaabaaβ :



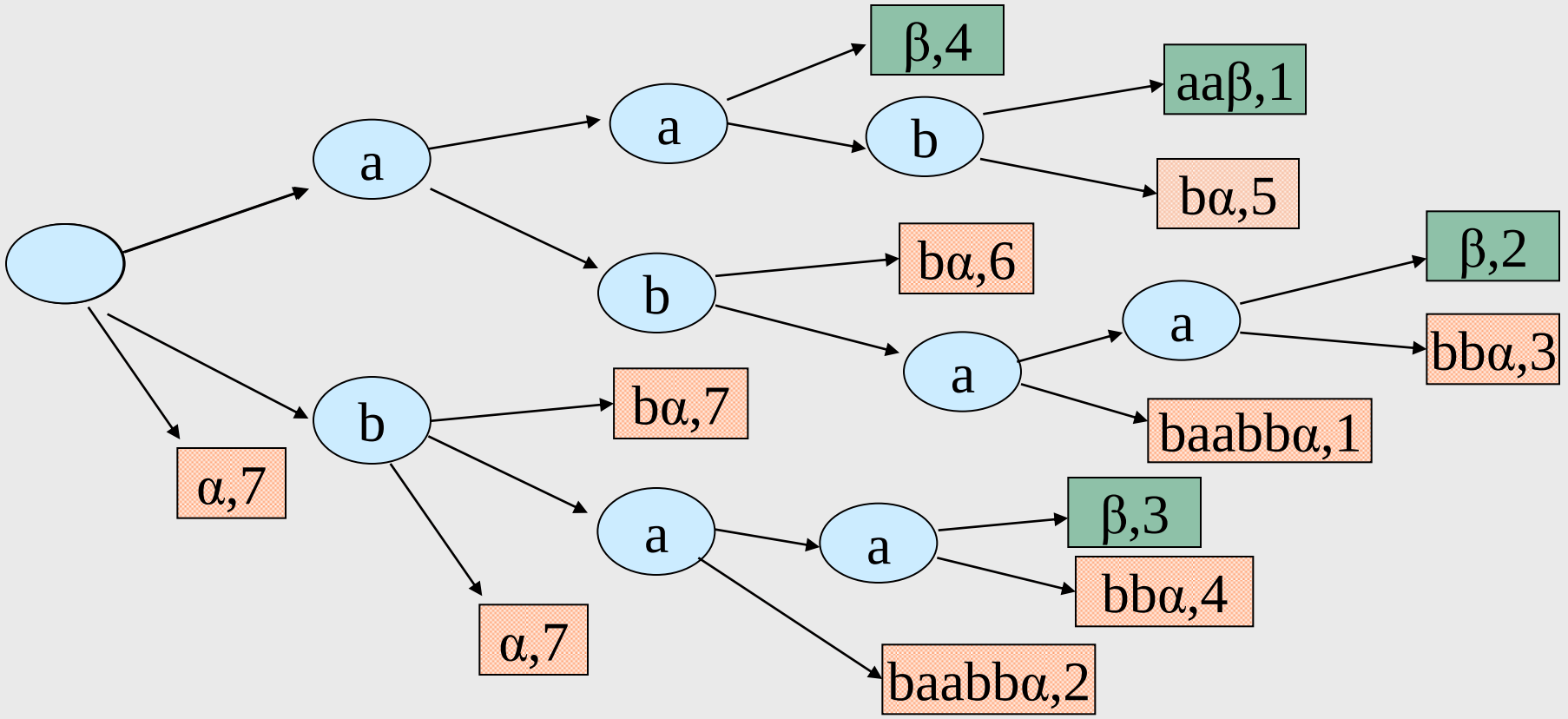
Generalizad suffix tree

Construction of the suffix tree of abababaabbαaabaaβ :



Generalizad suffix tree

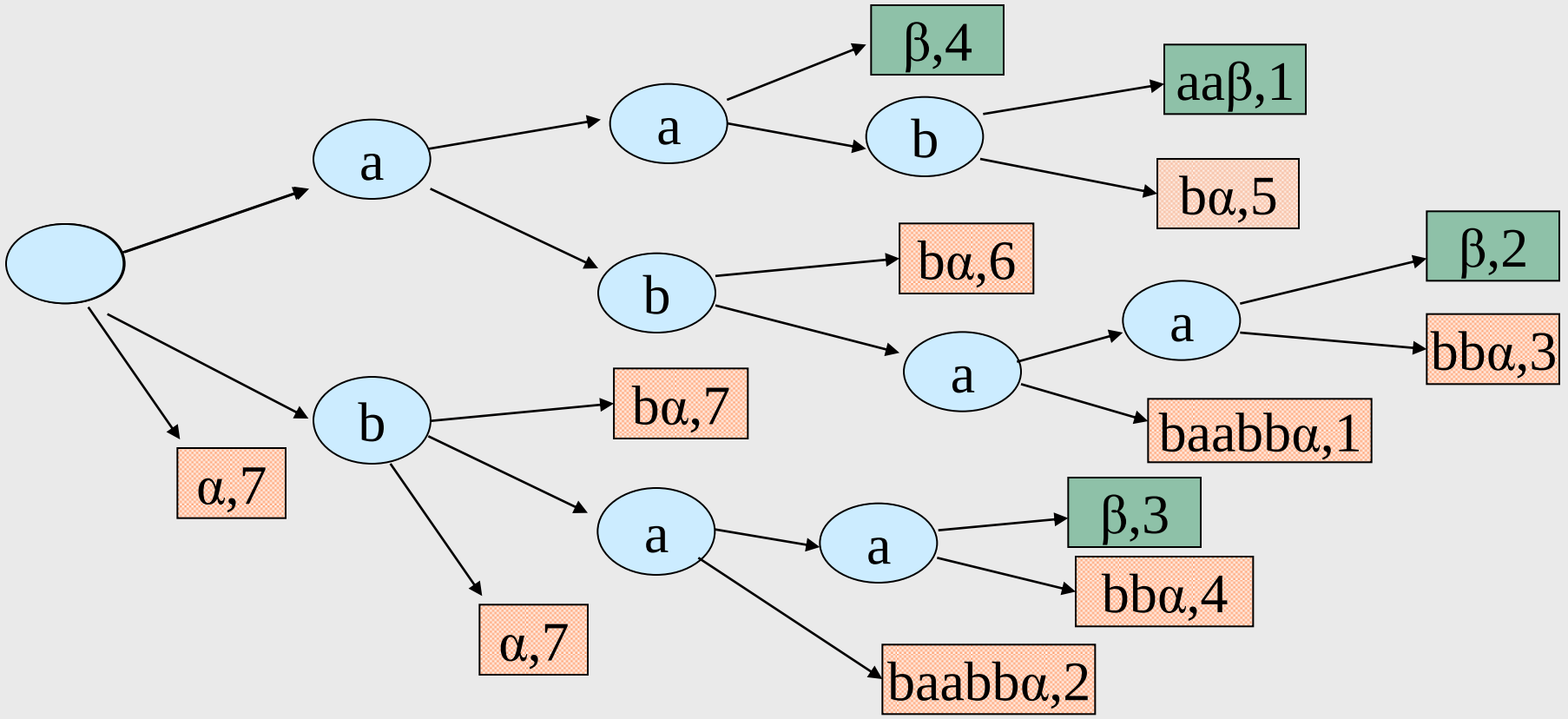
Construction of the suffix tree of abababaabbαaabaaβ :



Generalizad suffix tree

algorithmics genetics group .

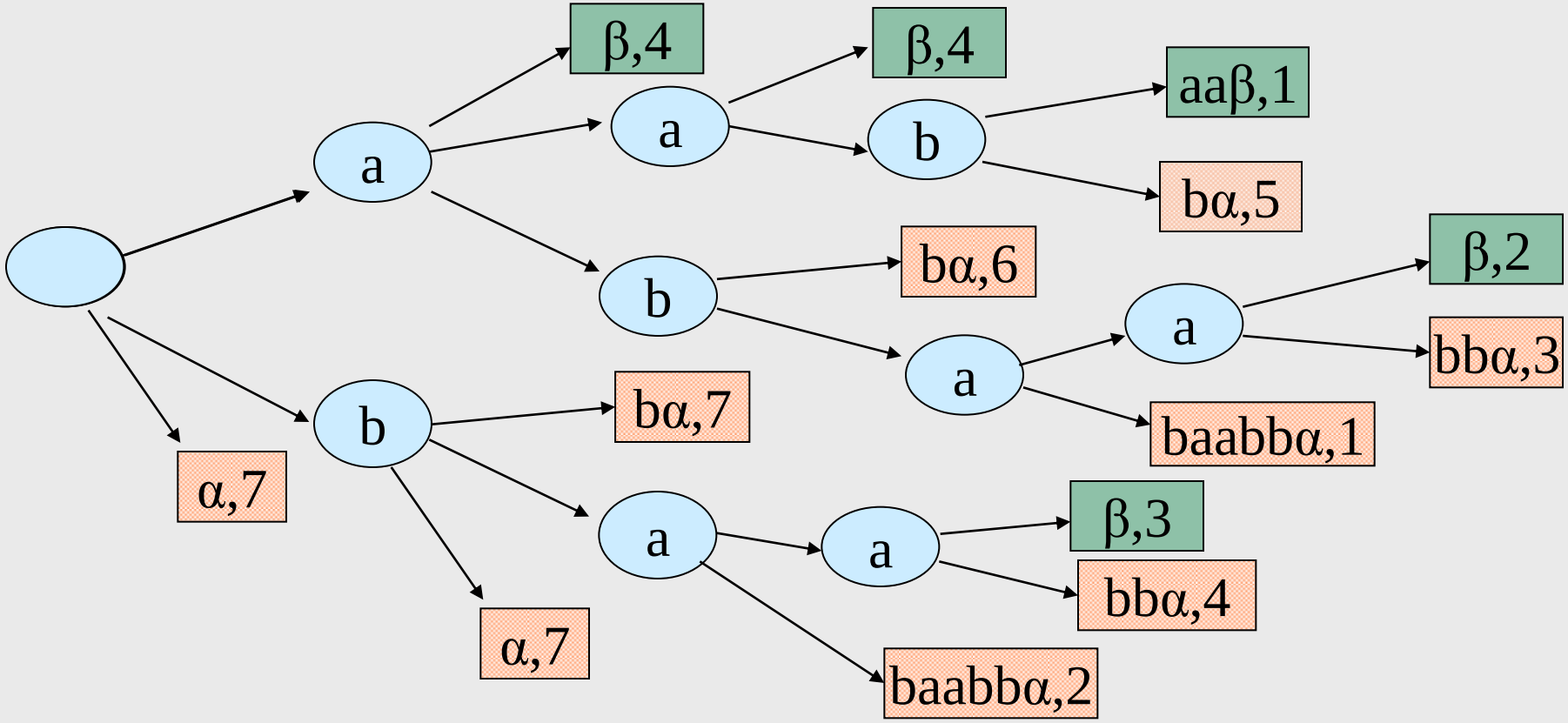
Construction of the suffix tree of abababaabbαaabaaβ :



Generalizad suffix tree

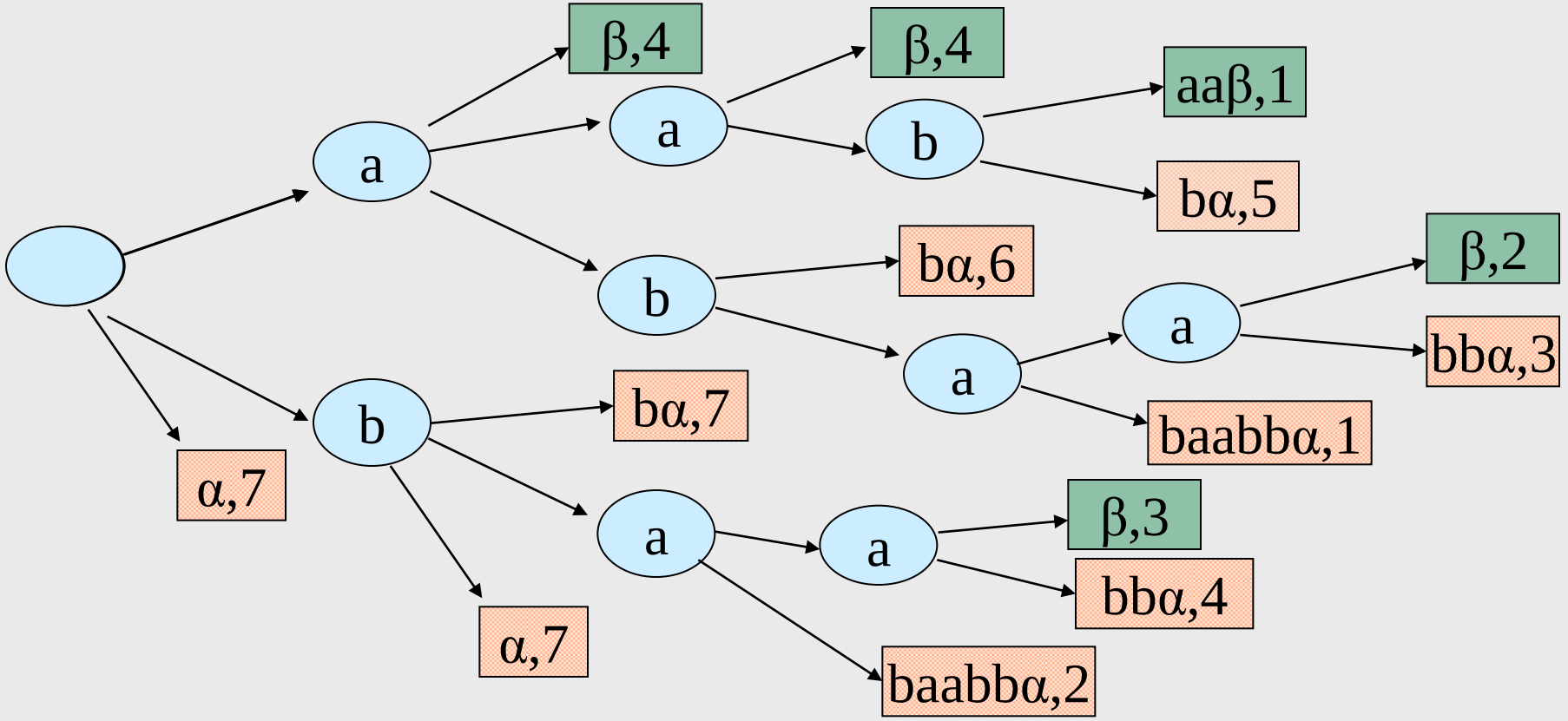
algorithmics genetics group .

Construction of the suffix tree of abababaabbαaabaaβ :



Generalizad suffix tree

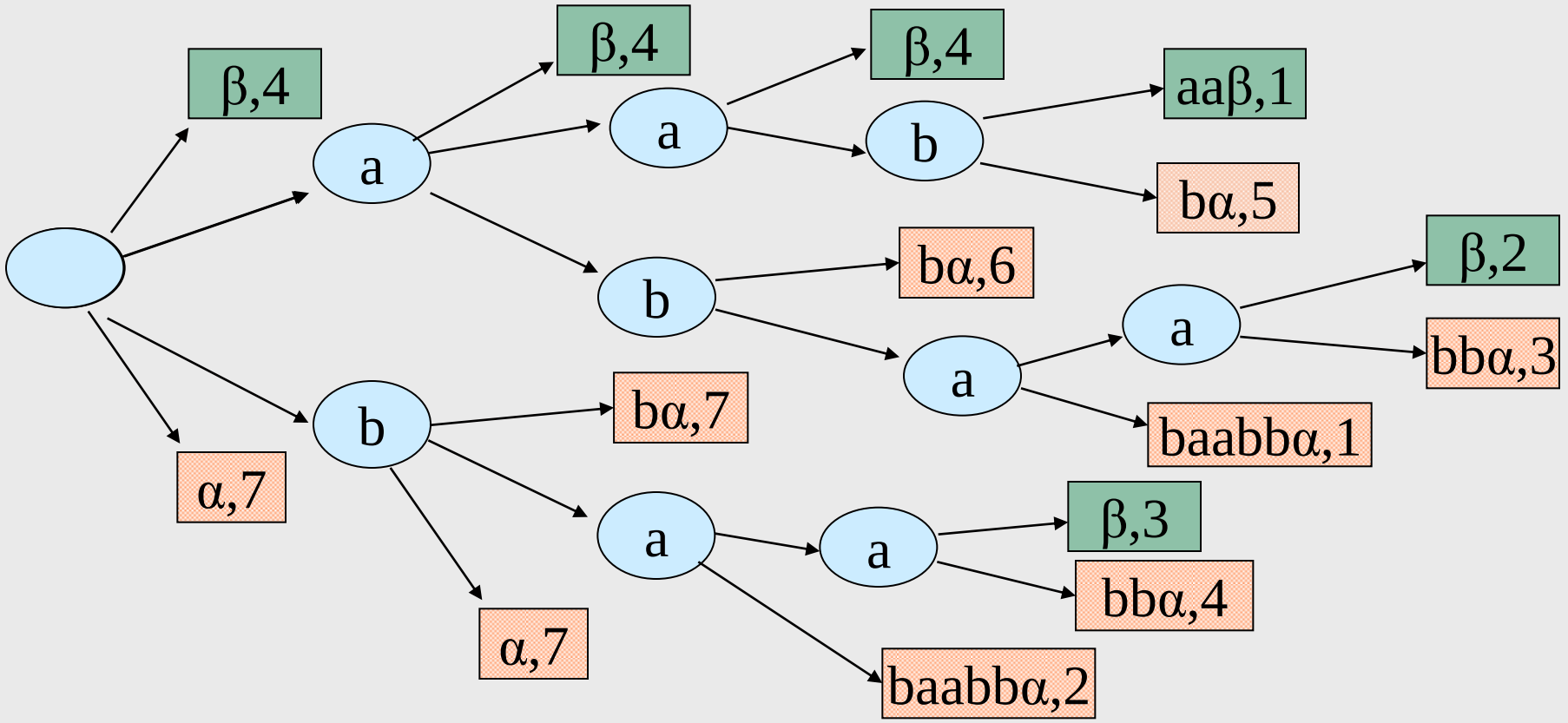
Construction of the suffix tree of abababaabbαaabaaβ :



Generalizad suffix tree

algorithmics genetics group .

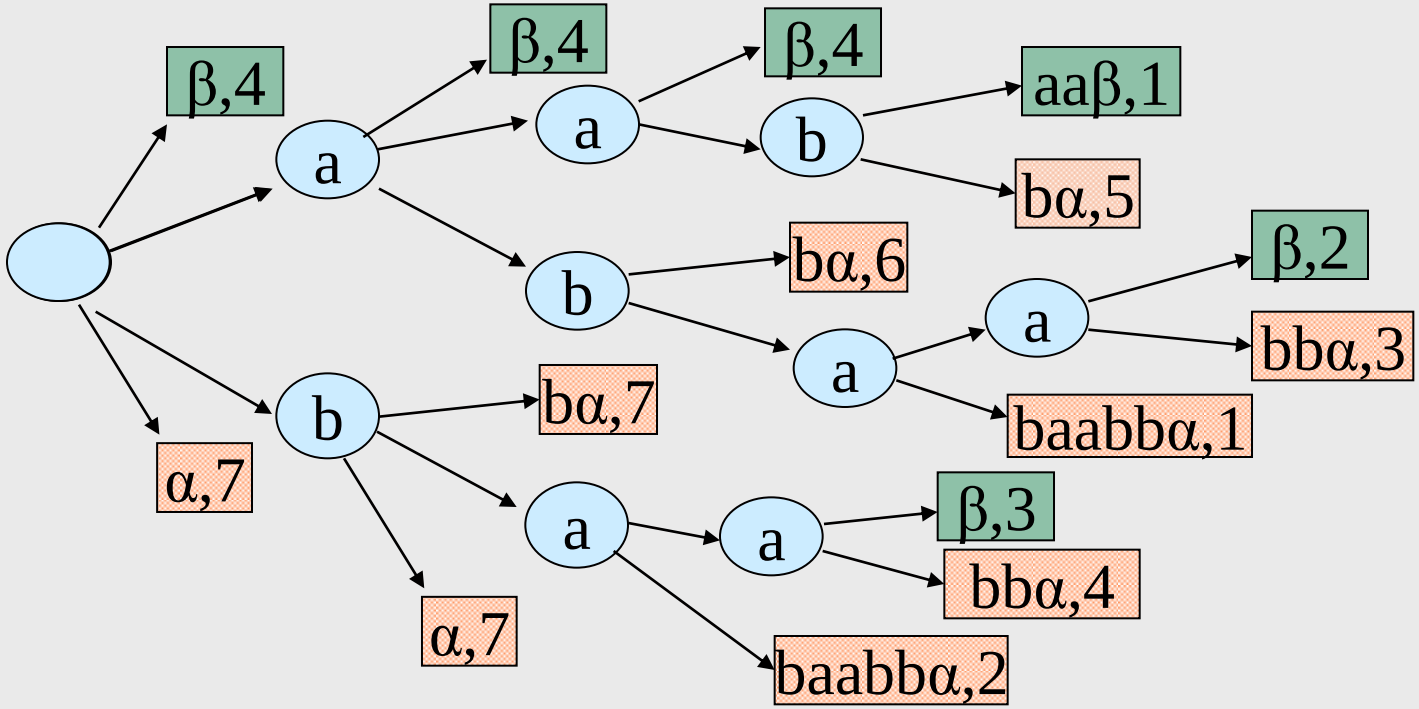
Construction of the suffix tree of abababaabbαaabaaβ :



Generalizad suffix tree

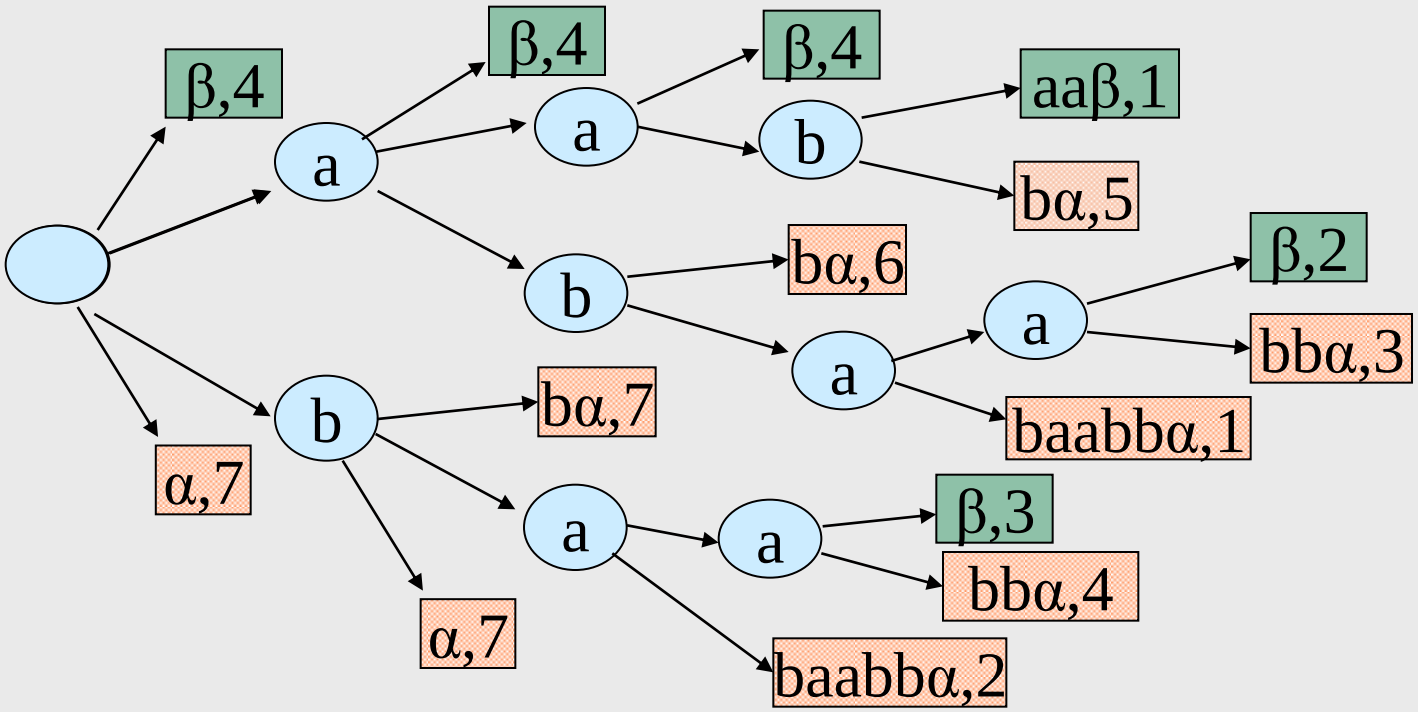
algorithmics genetics group .

Generalized suffix tree of abababaabbαaabaaβ :



Applications of Generalized Suffix trees

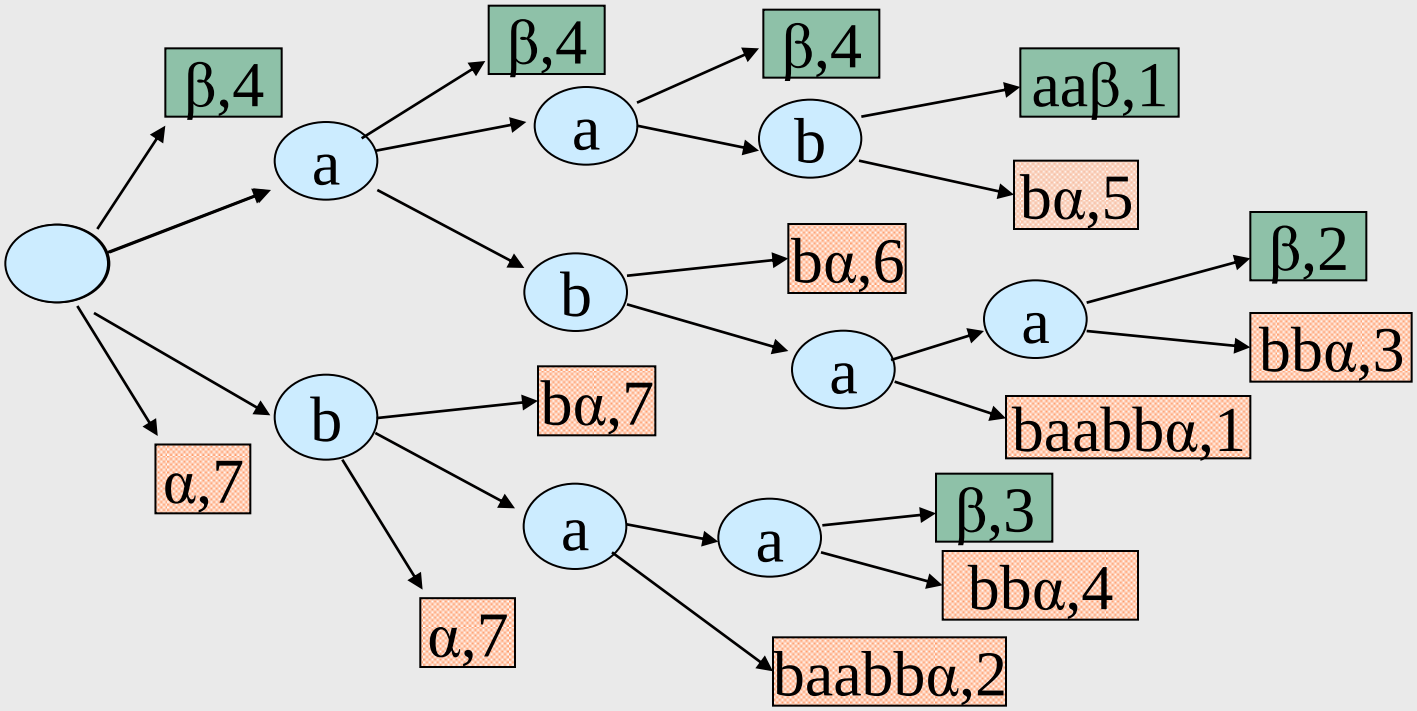
1. The substring problem for a database of strings DB
   - Does the DB contain any ocurrence of patterns abab, aab, and ab?



Applications of Generalized Suffix trees

algorithmics genetics group .

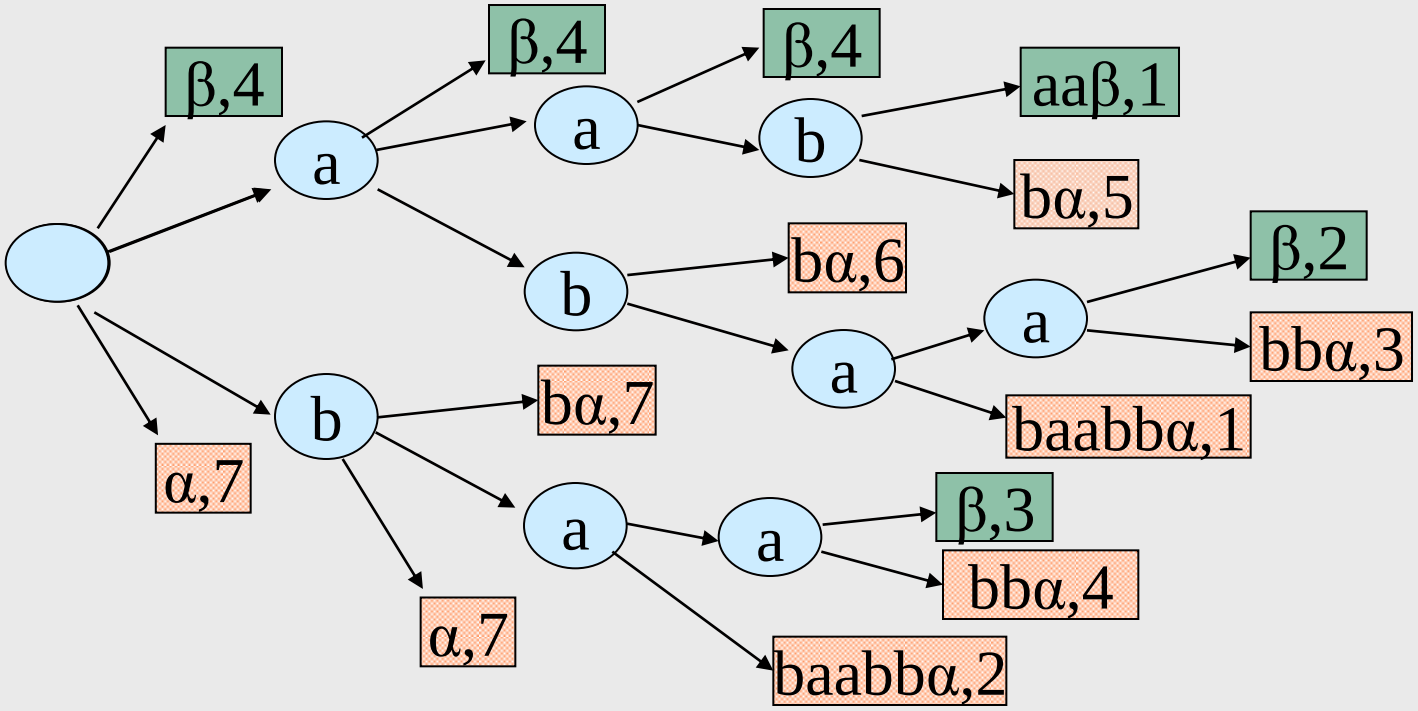
2. The longest common substring of two strings



Applications of Generalized Suffix trees

algorithmics genetics group

3. Finding MUMs.

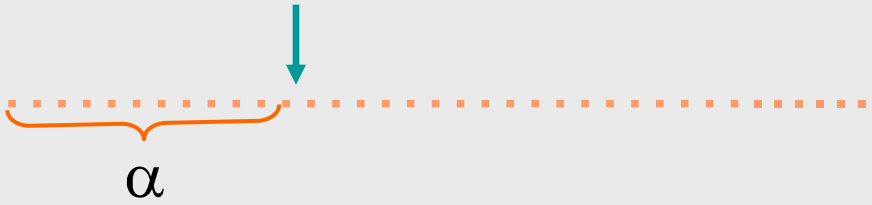


Quadratic insertion algorithm

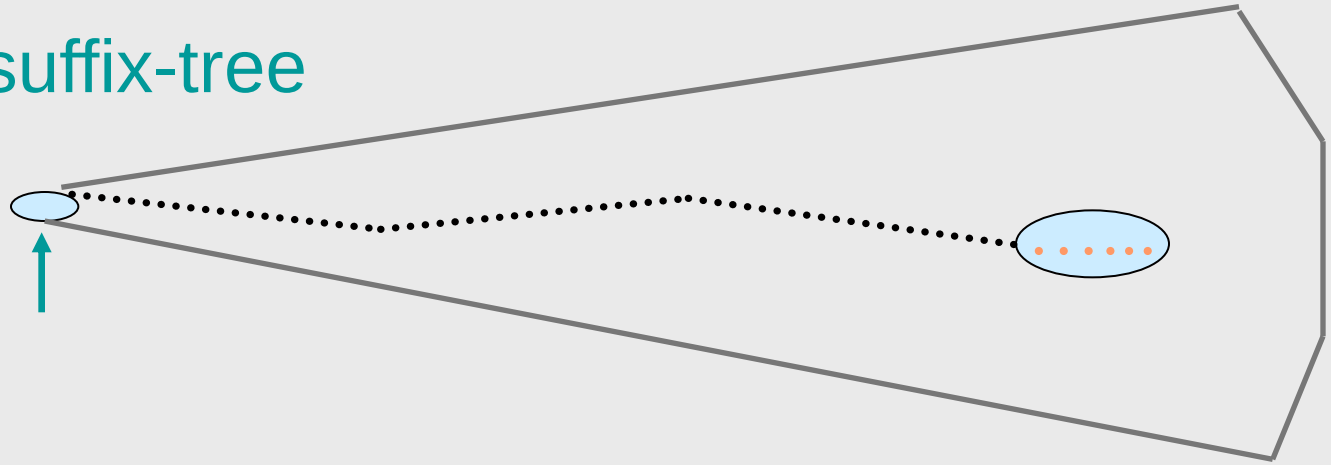
algorithmics genetics group

Invariant Properties:

Given the string



and the suffix-tree



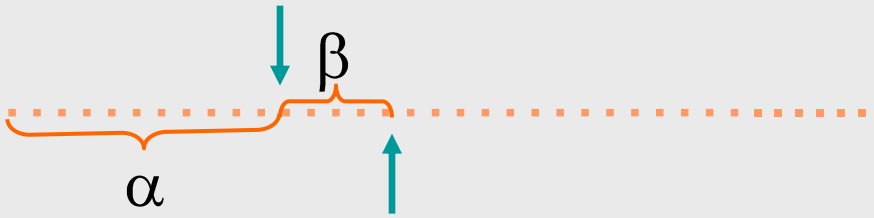
P1: the leaves of suffixes from α have been inserted

Linear insertion algorithm

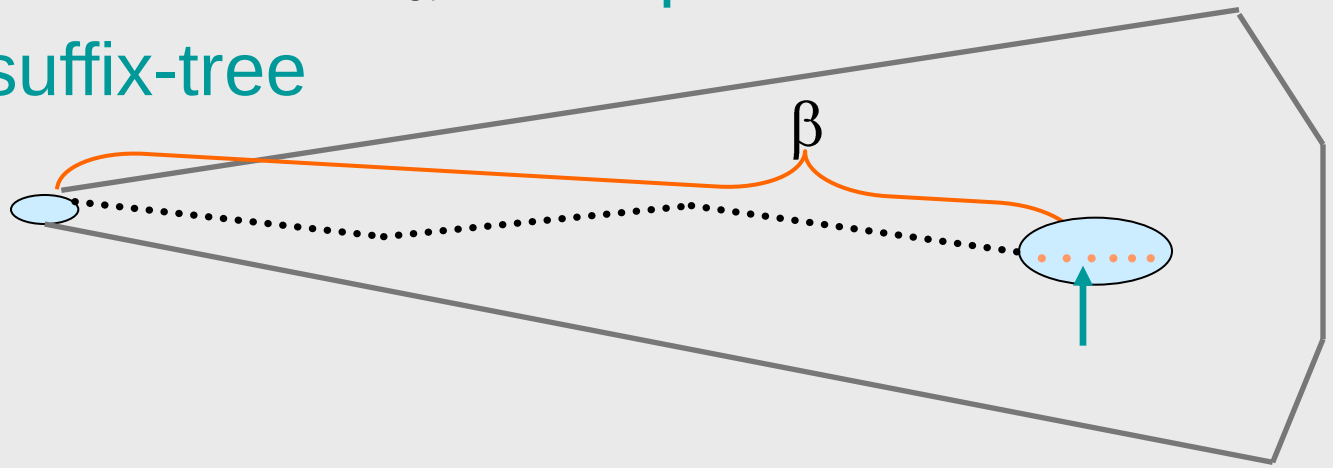
algorithmics genetics group .

Invariant Properties:

Given the string



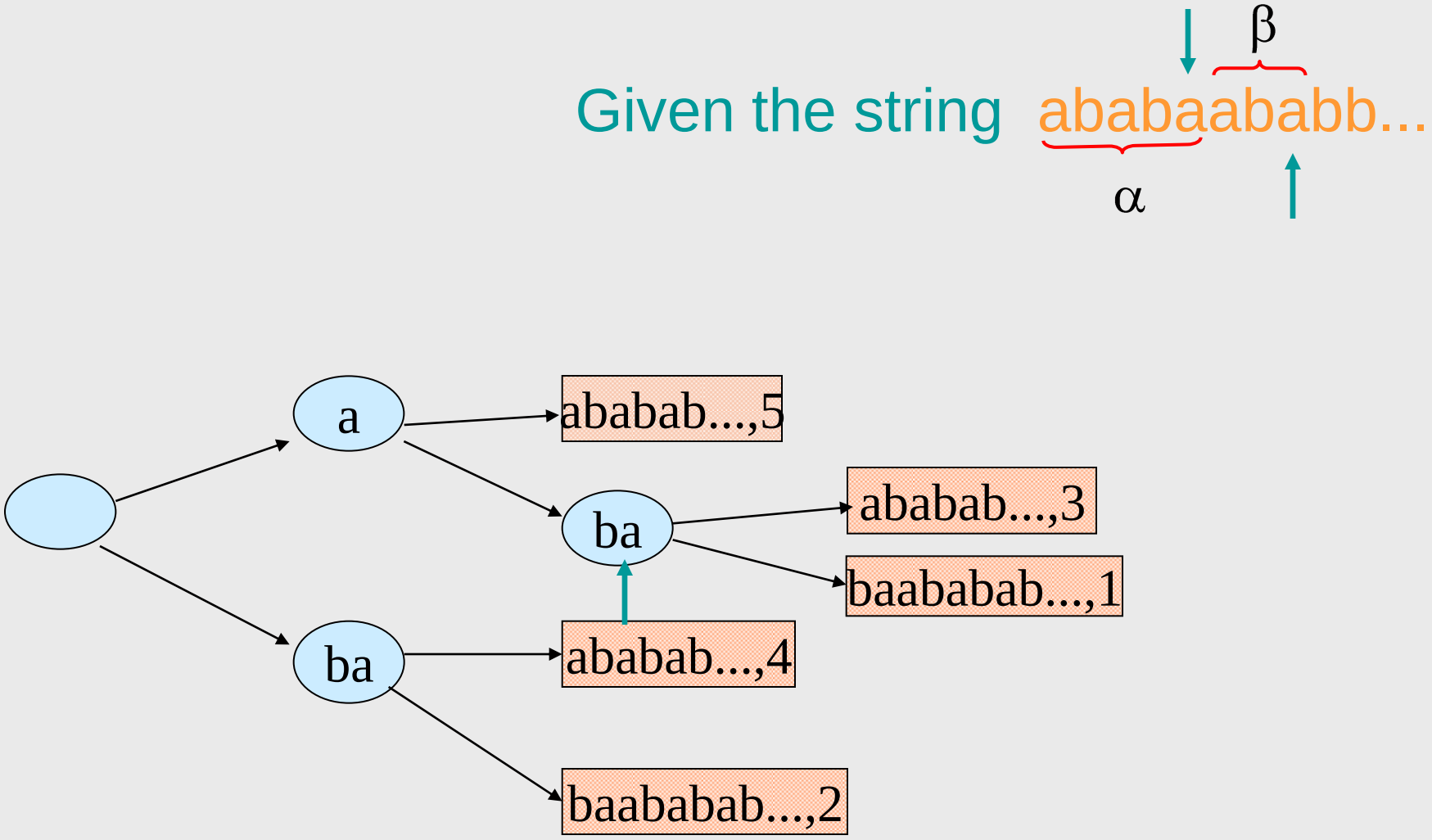
and the suffix-tree



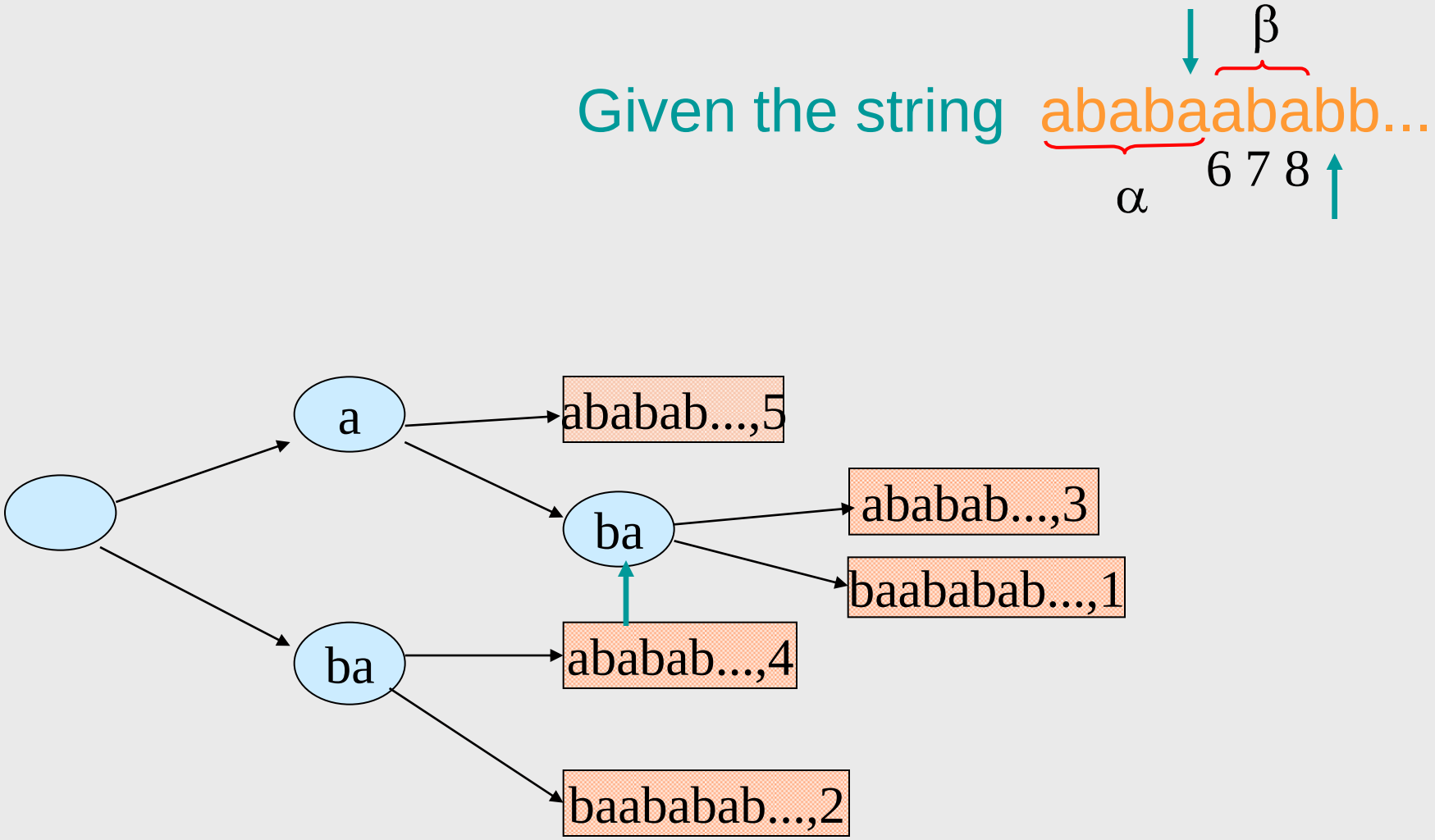
P1: the leaves of suffixes from α have been inserted

P2: the string β is the longest string that can be spelt through the tree.

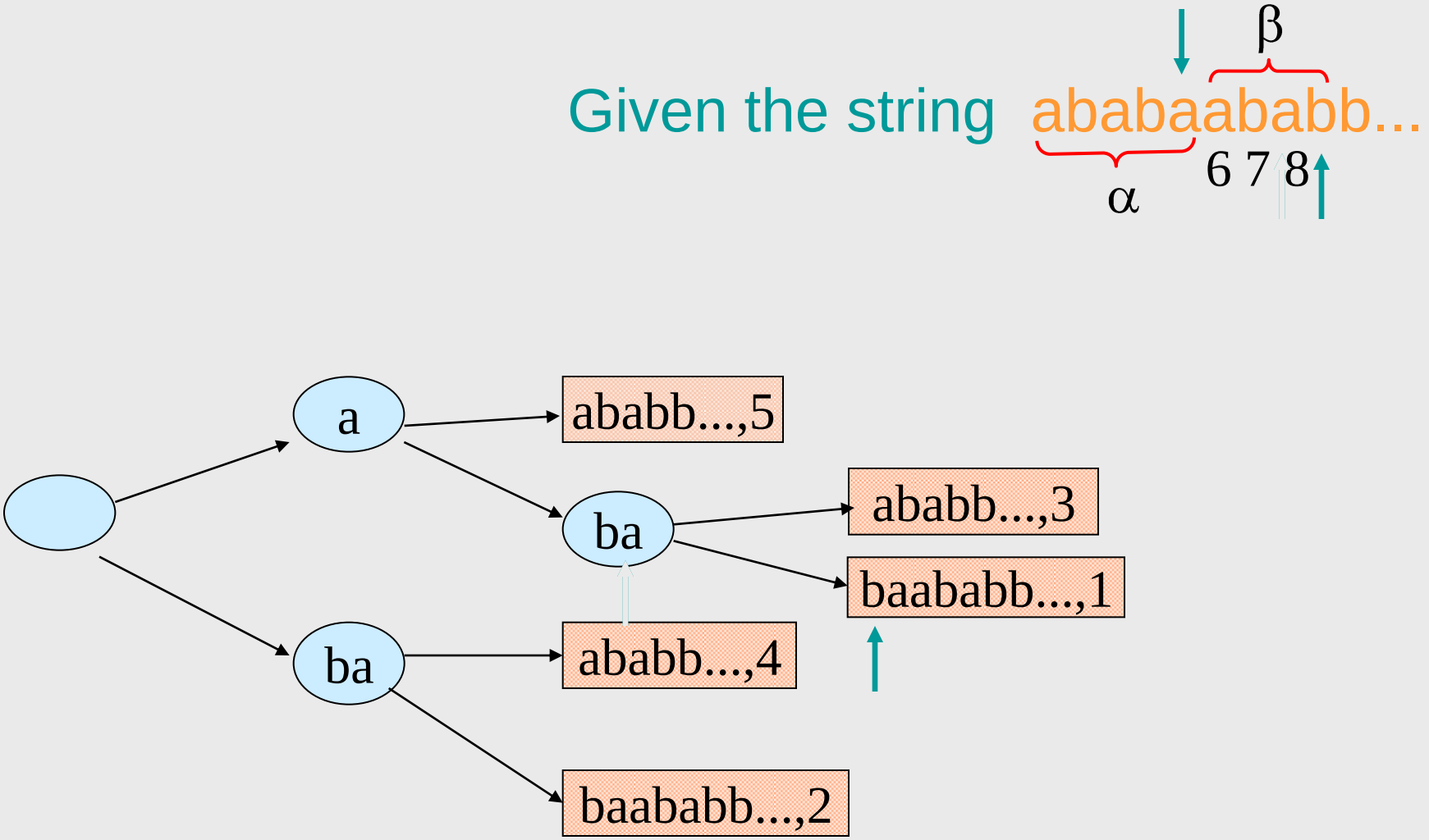
Linear insertion algorithm: example



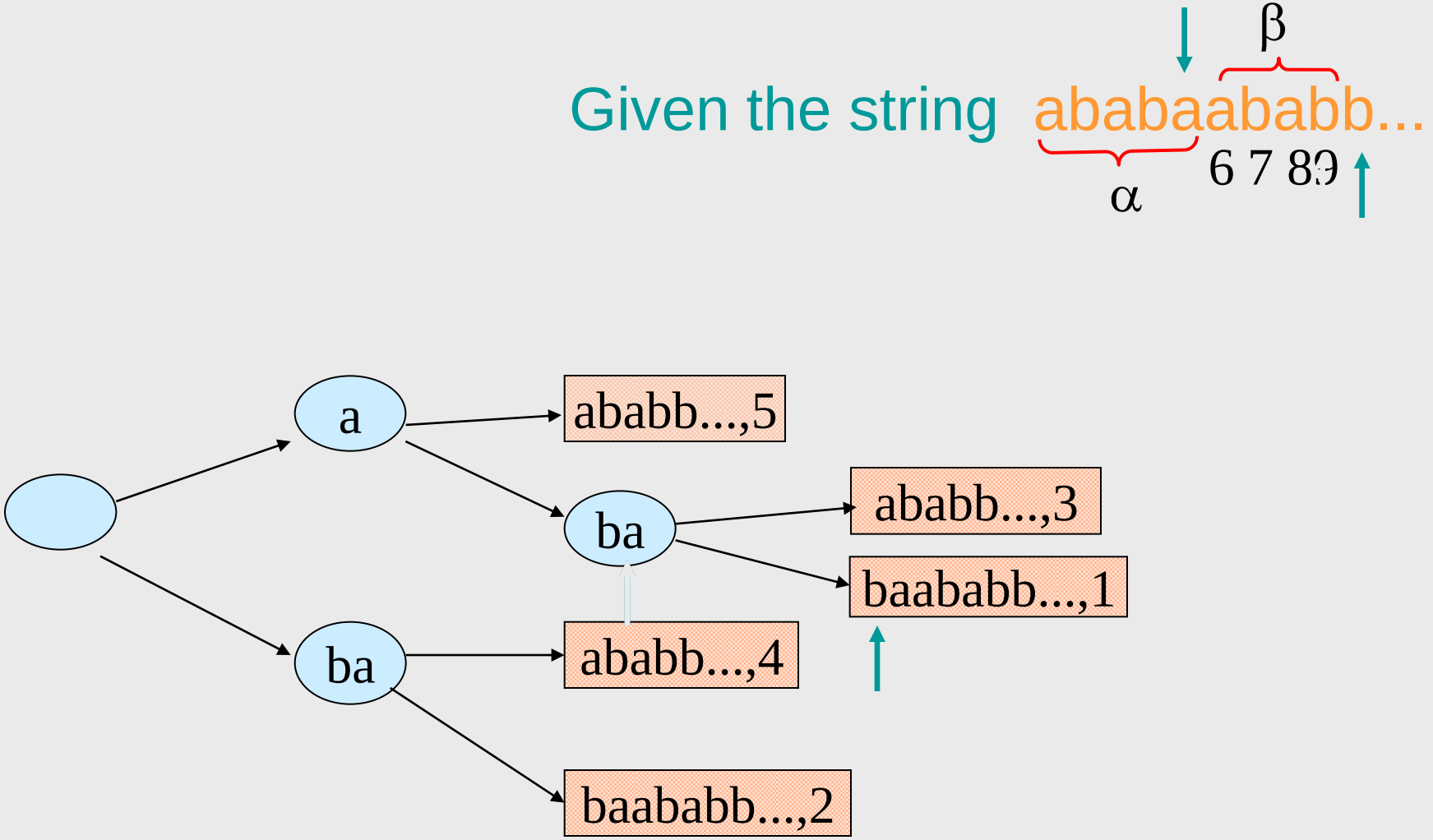
Linear insertion algorithm: example



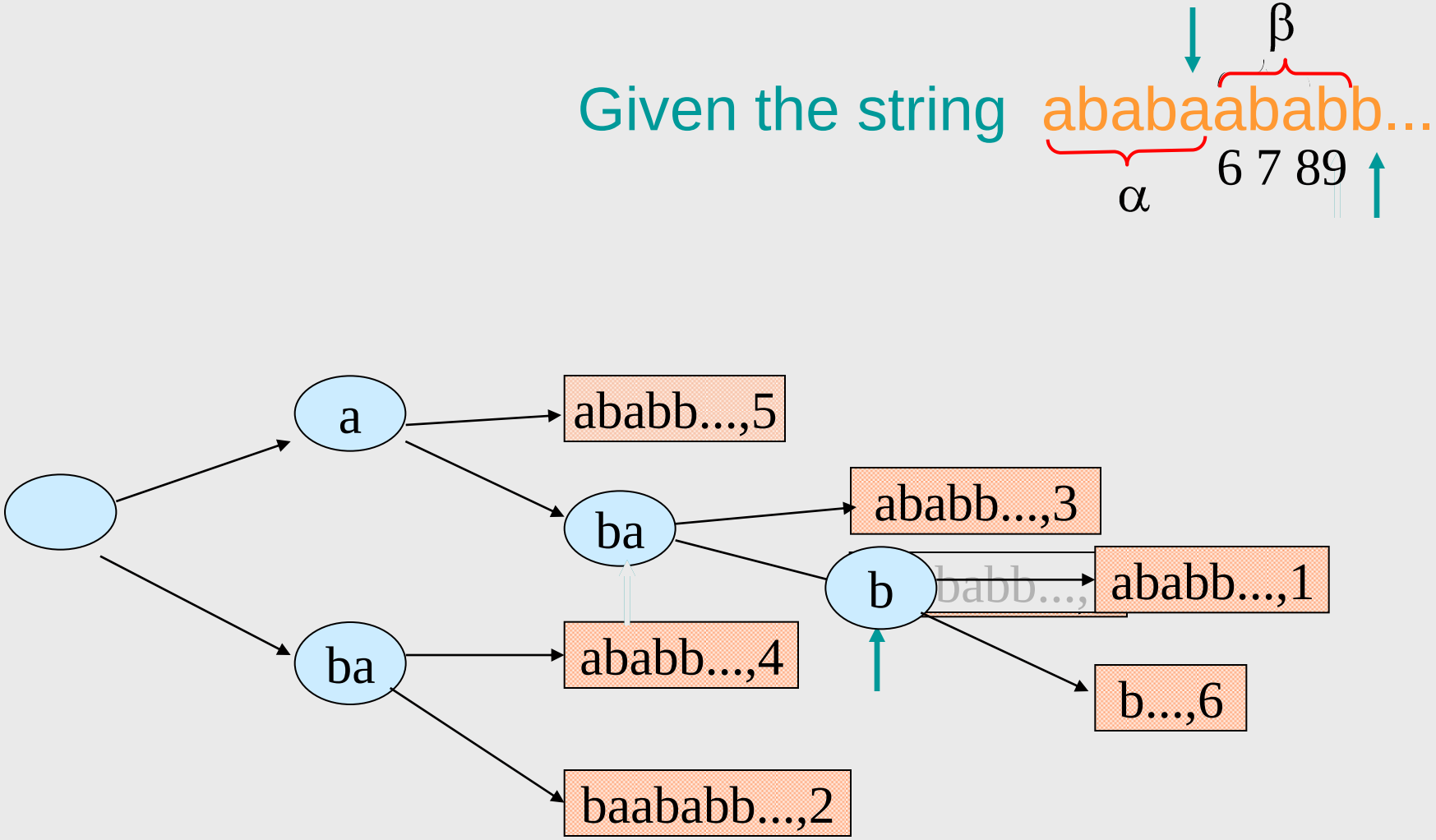
Linear insertion algorithm: example



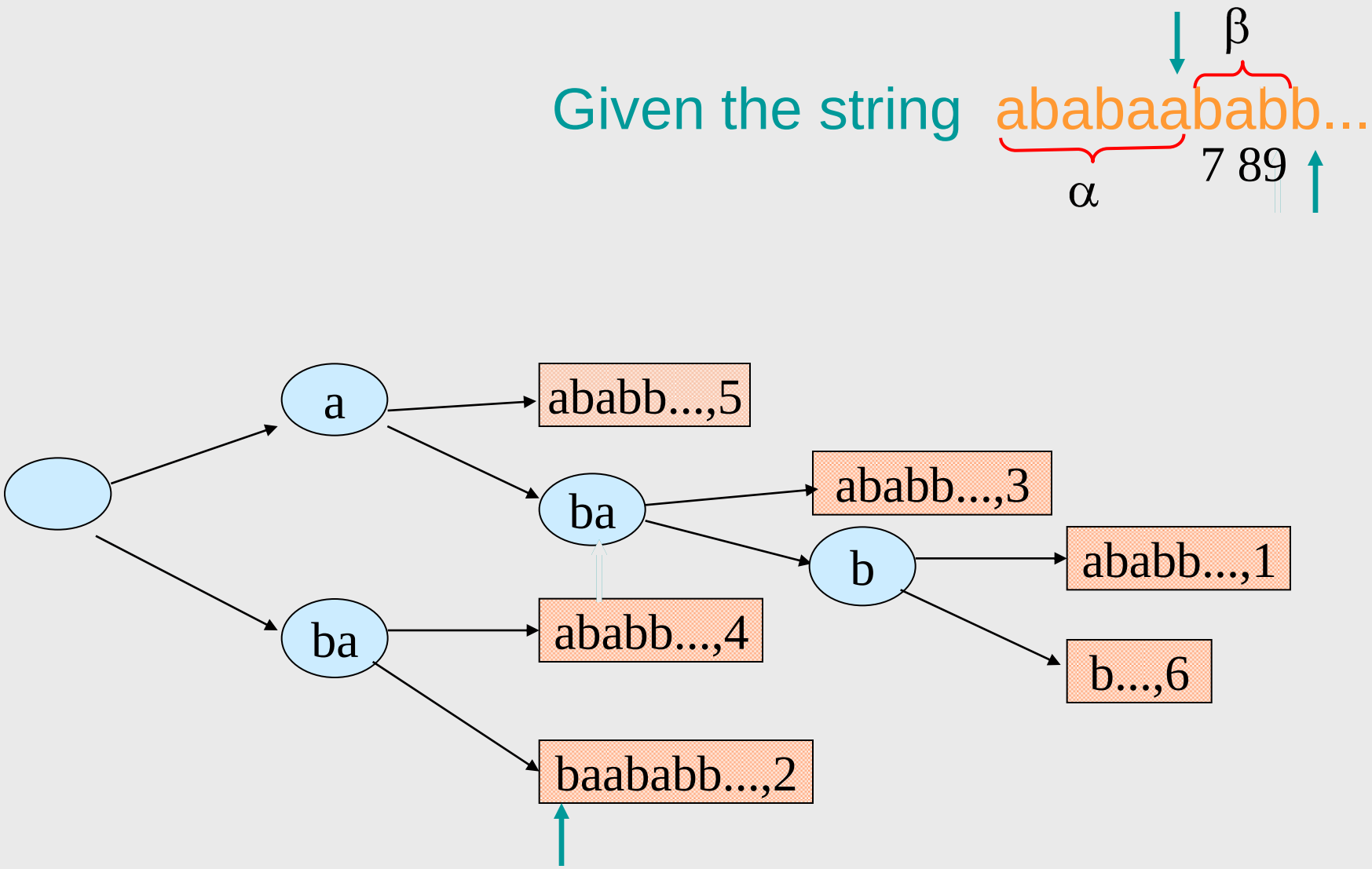
Linear insertion algorithm: example



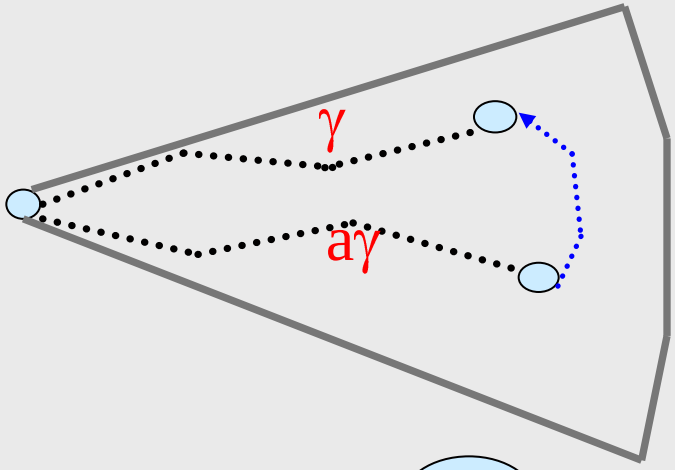
Linear insertion algorithm: example



Linear insertion algorithm: example



Linear insertion algorithm: example

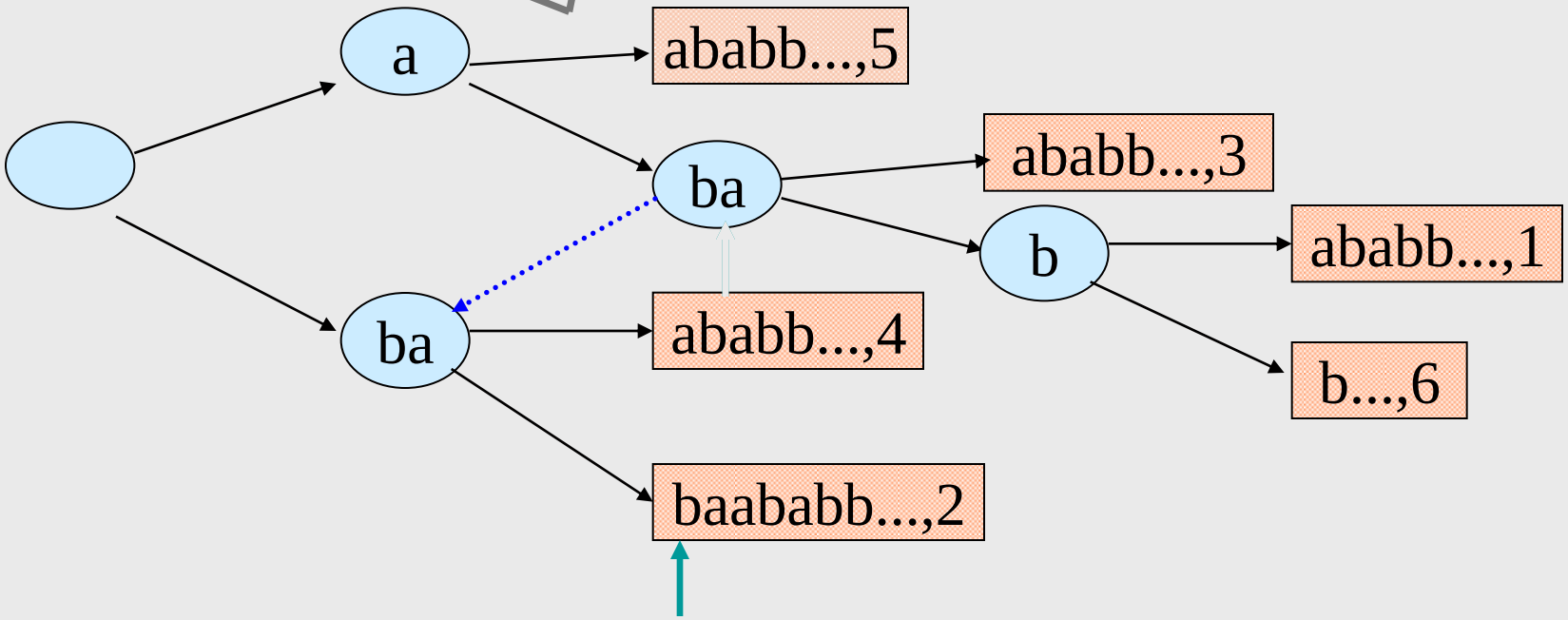


Given the string ababaababb...

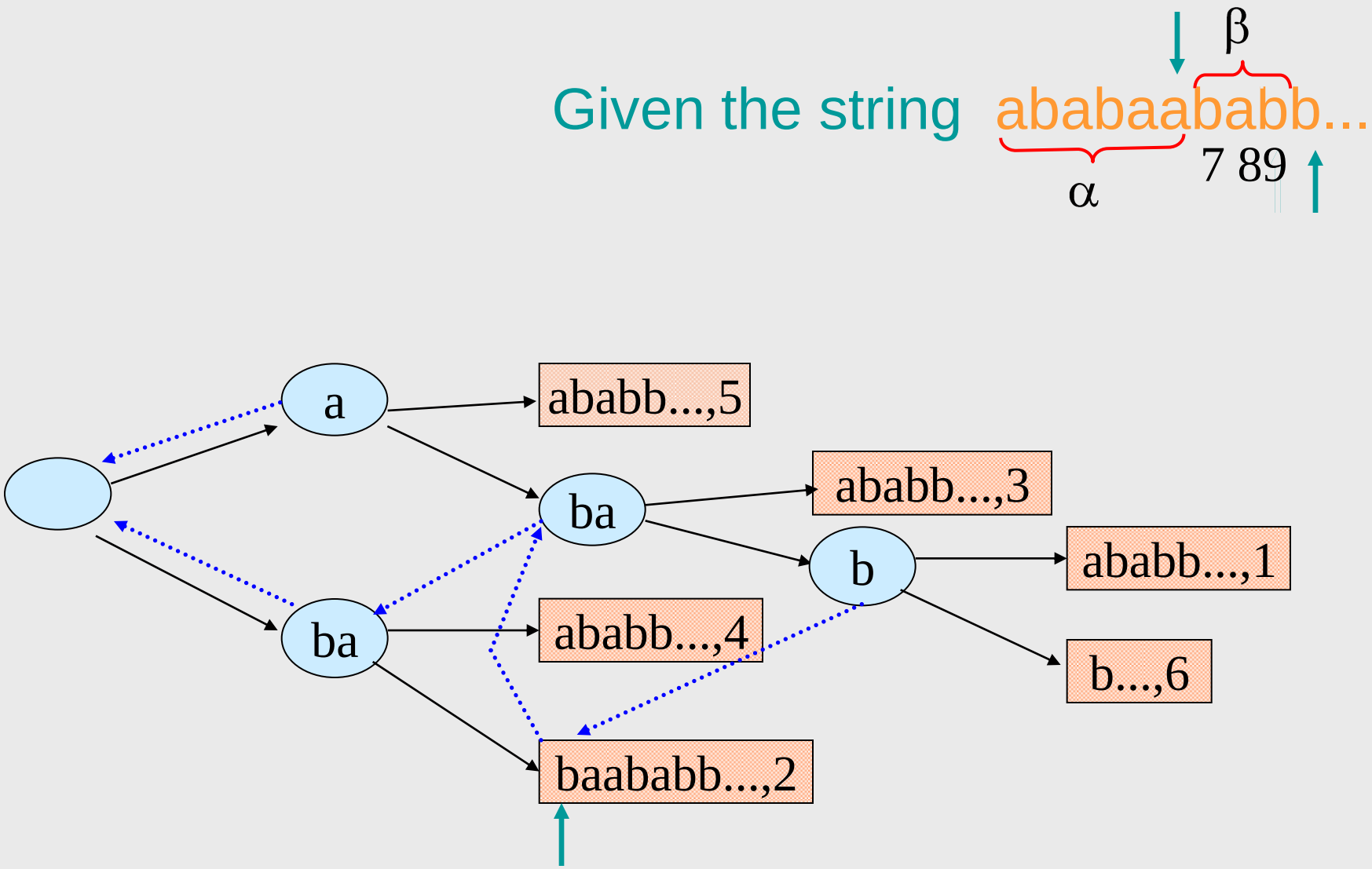
β

α

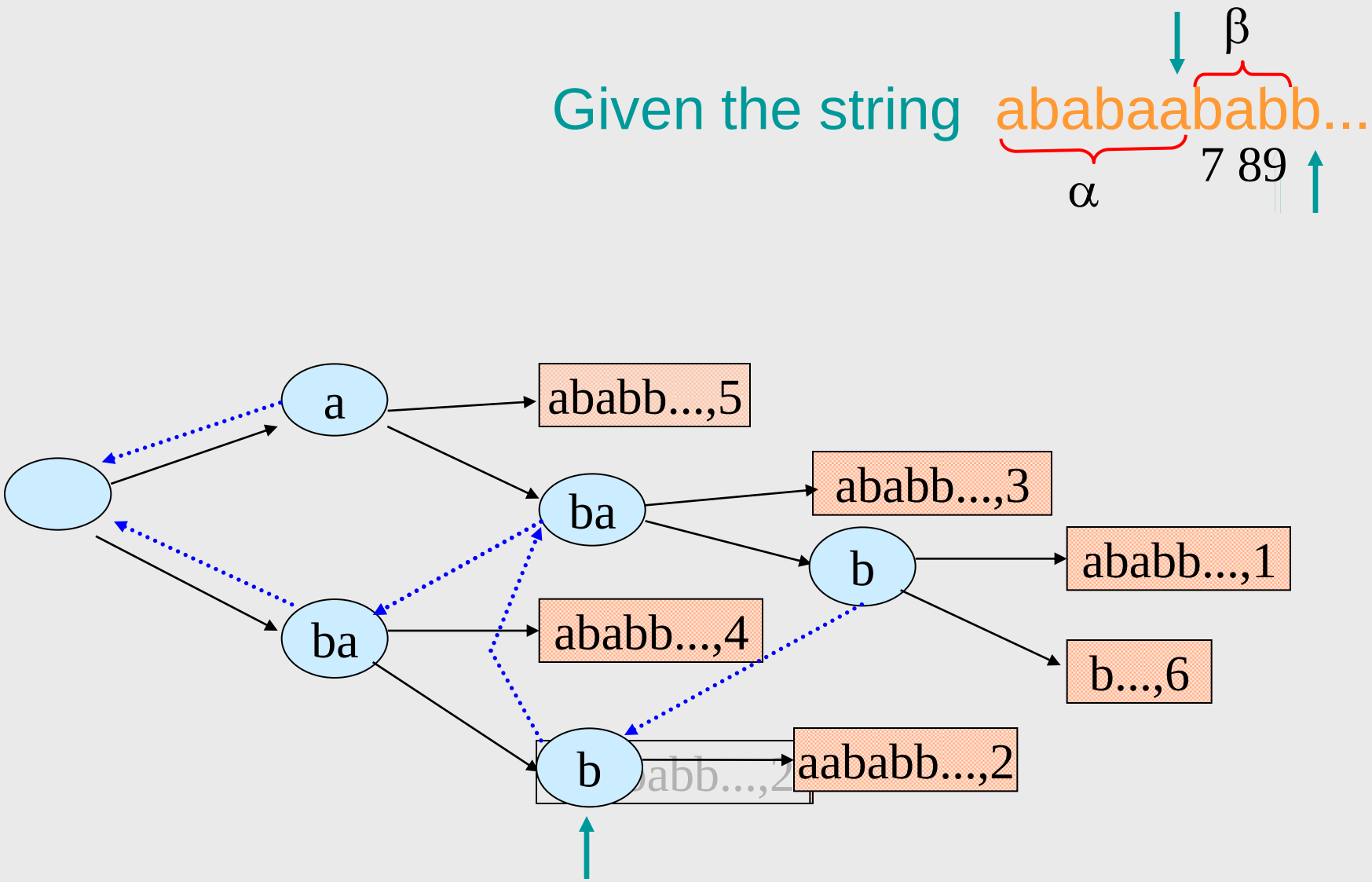
7 89



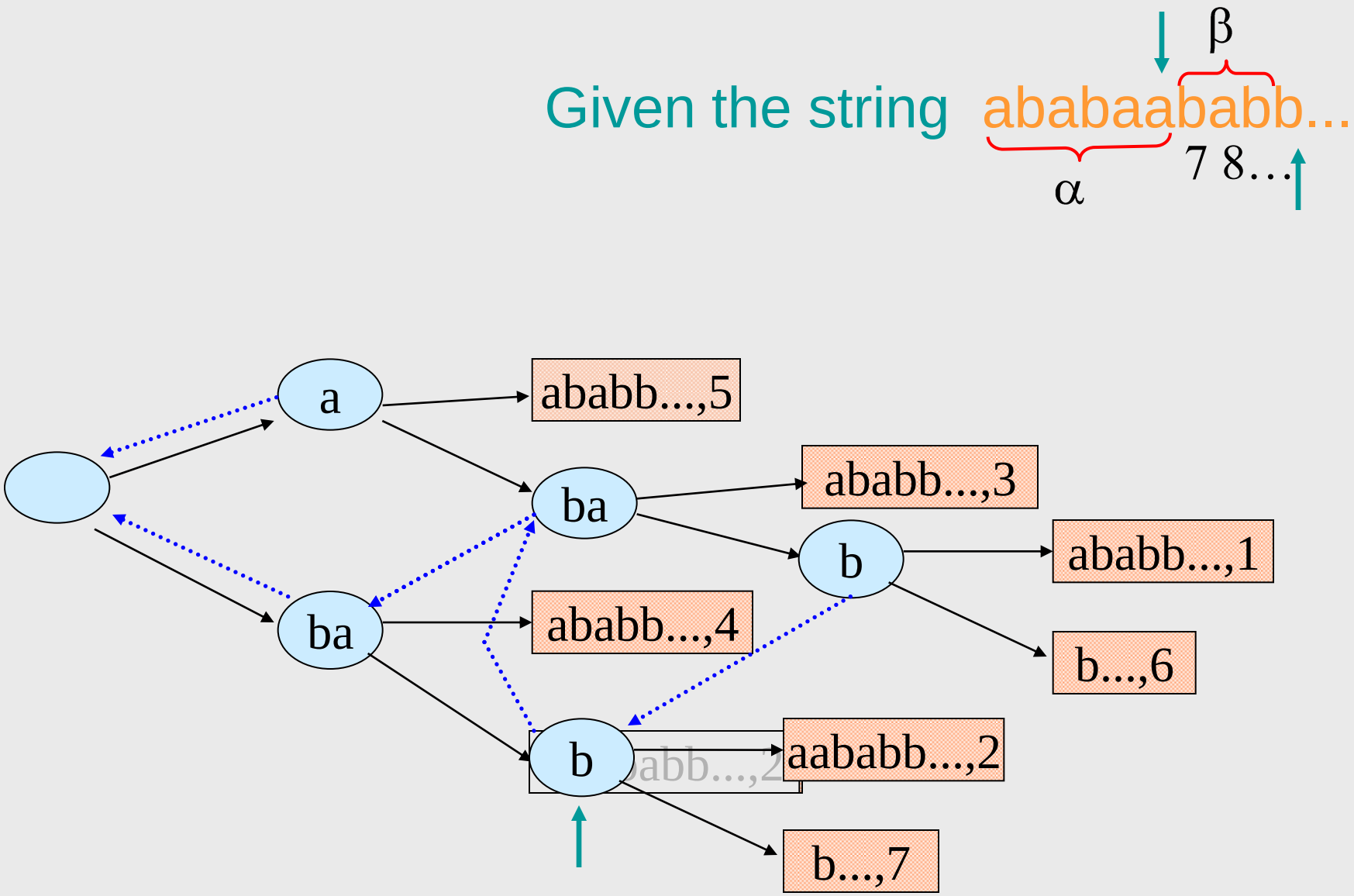
Linear insertion algorithm: example



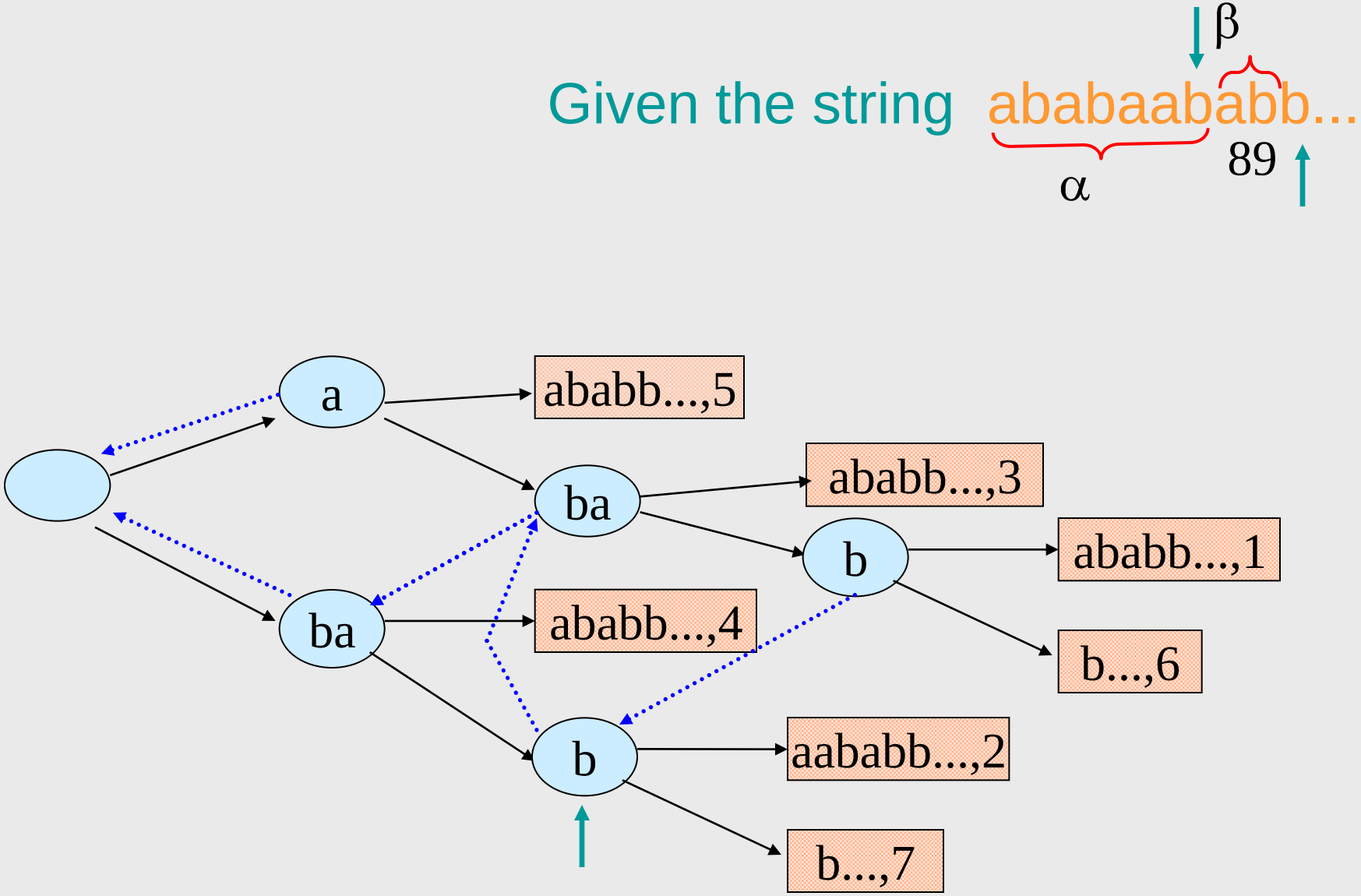
Linear insertion algorithm: example



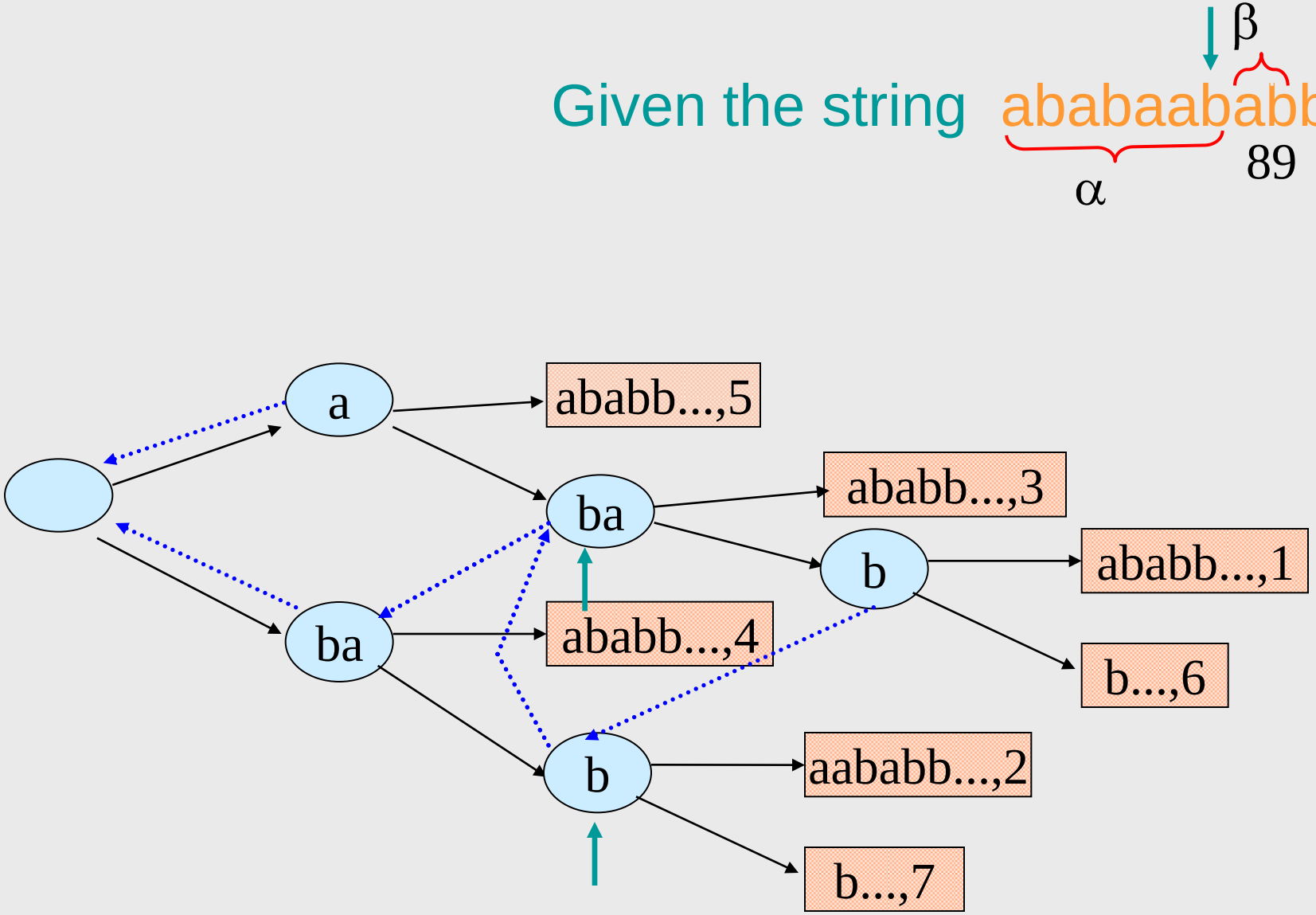
Linear insertion algorithm: example



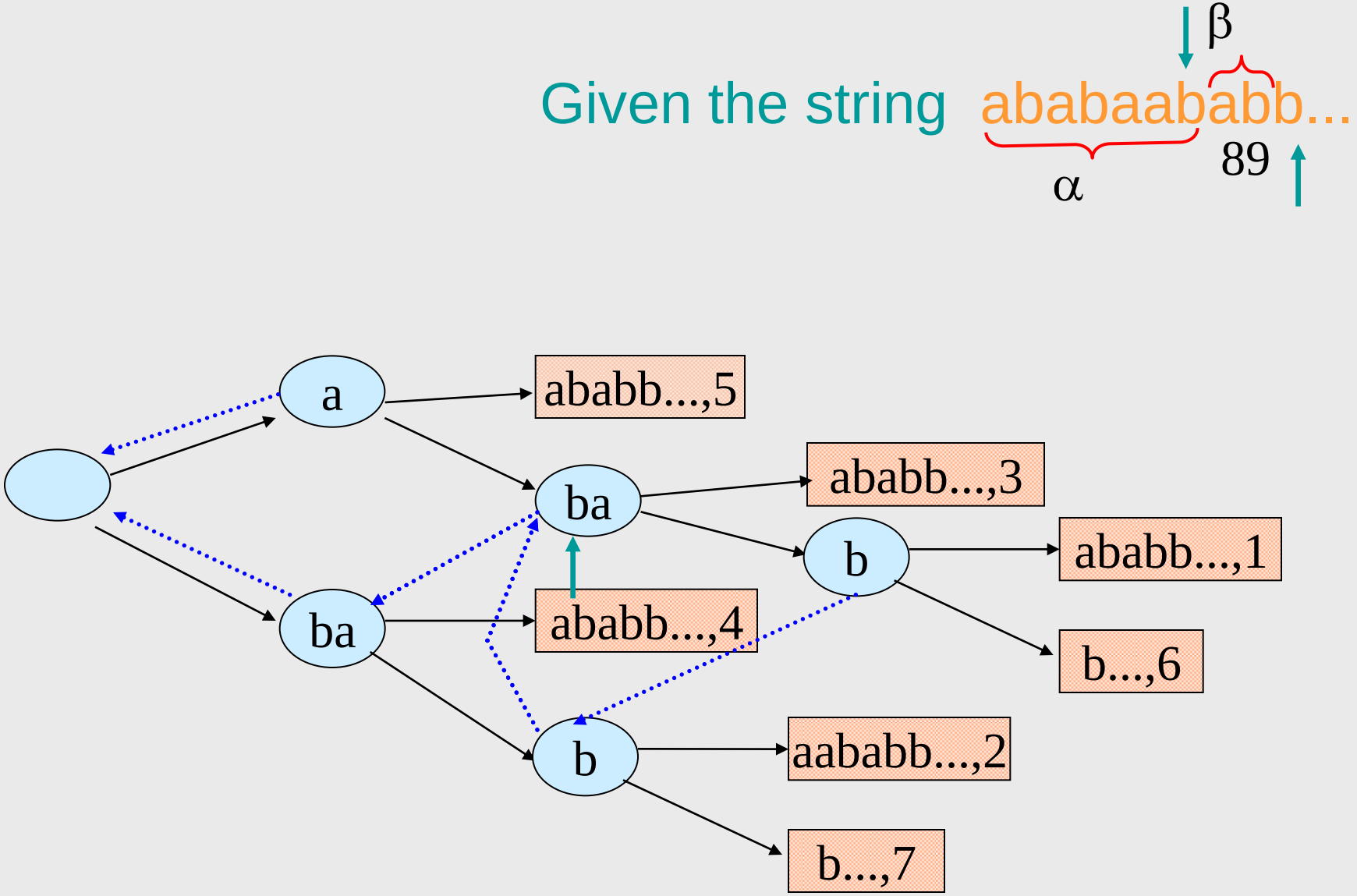
Linear insertion algorithm: example



Linear insertion algorithm: example

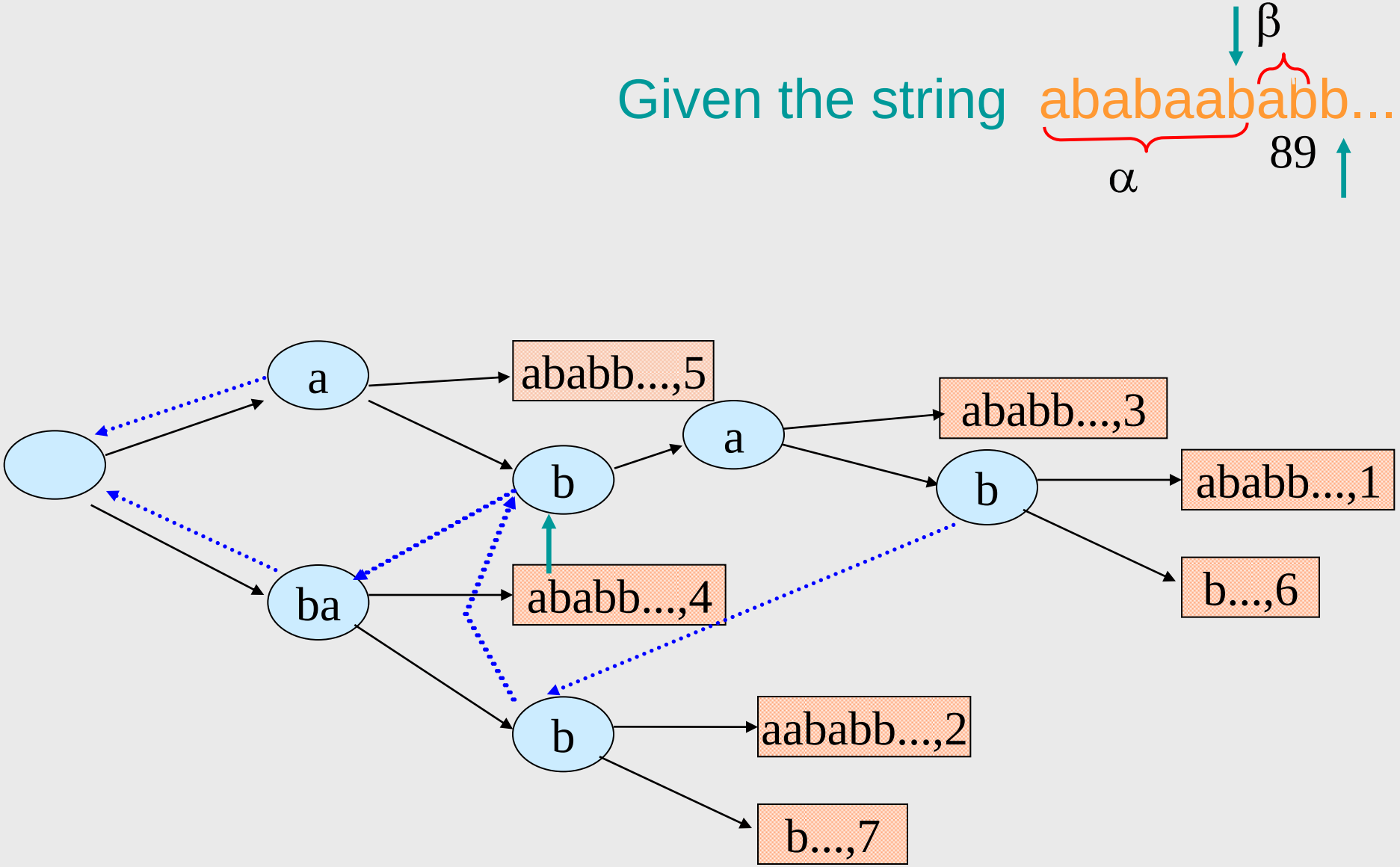


Linear insertion algorithm: example

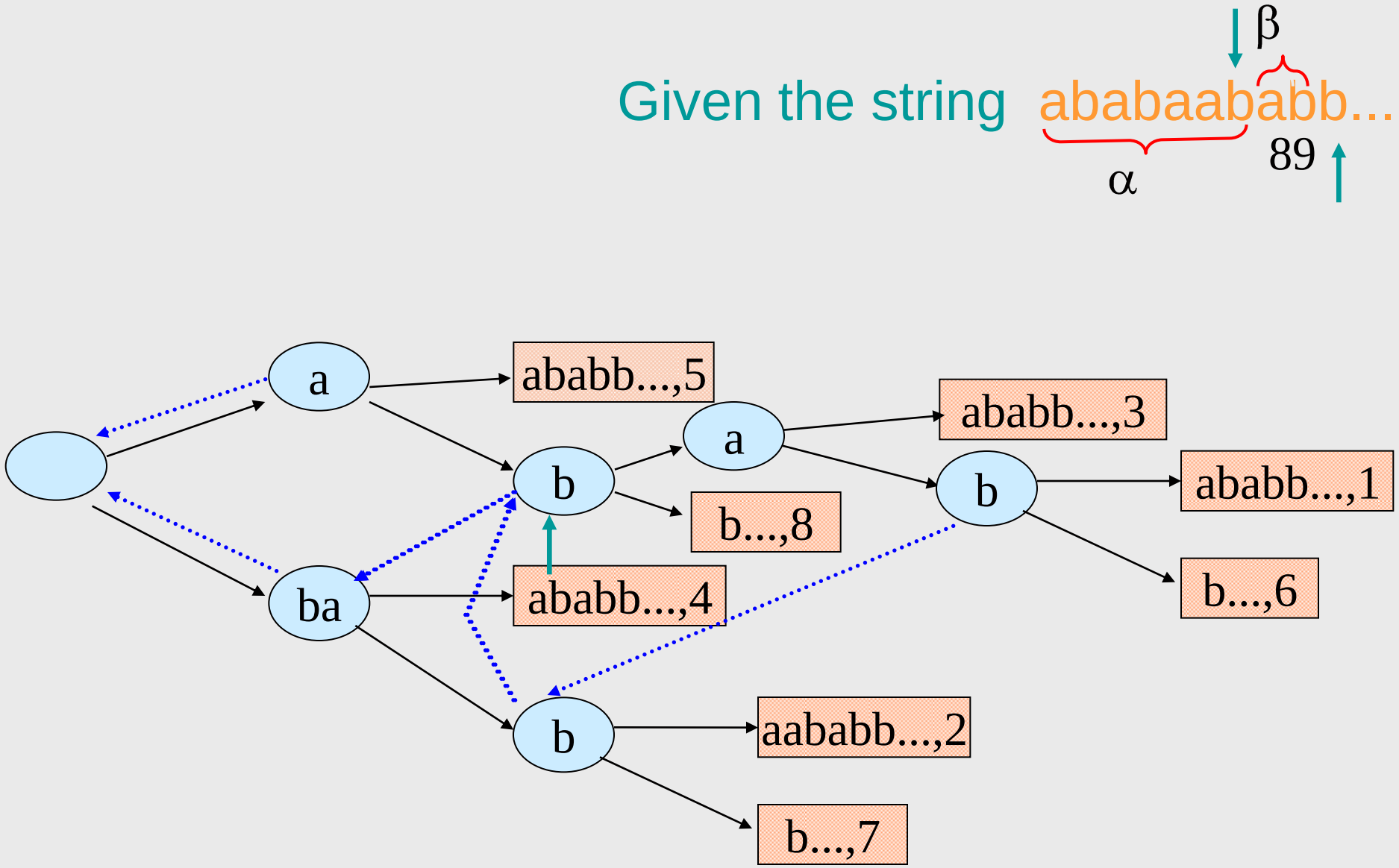


Linear insertion algorithm: example

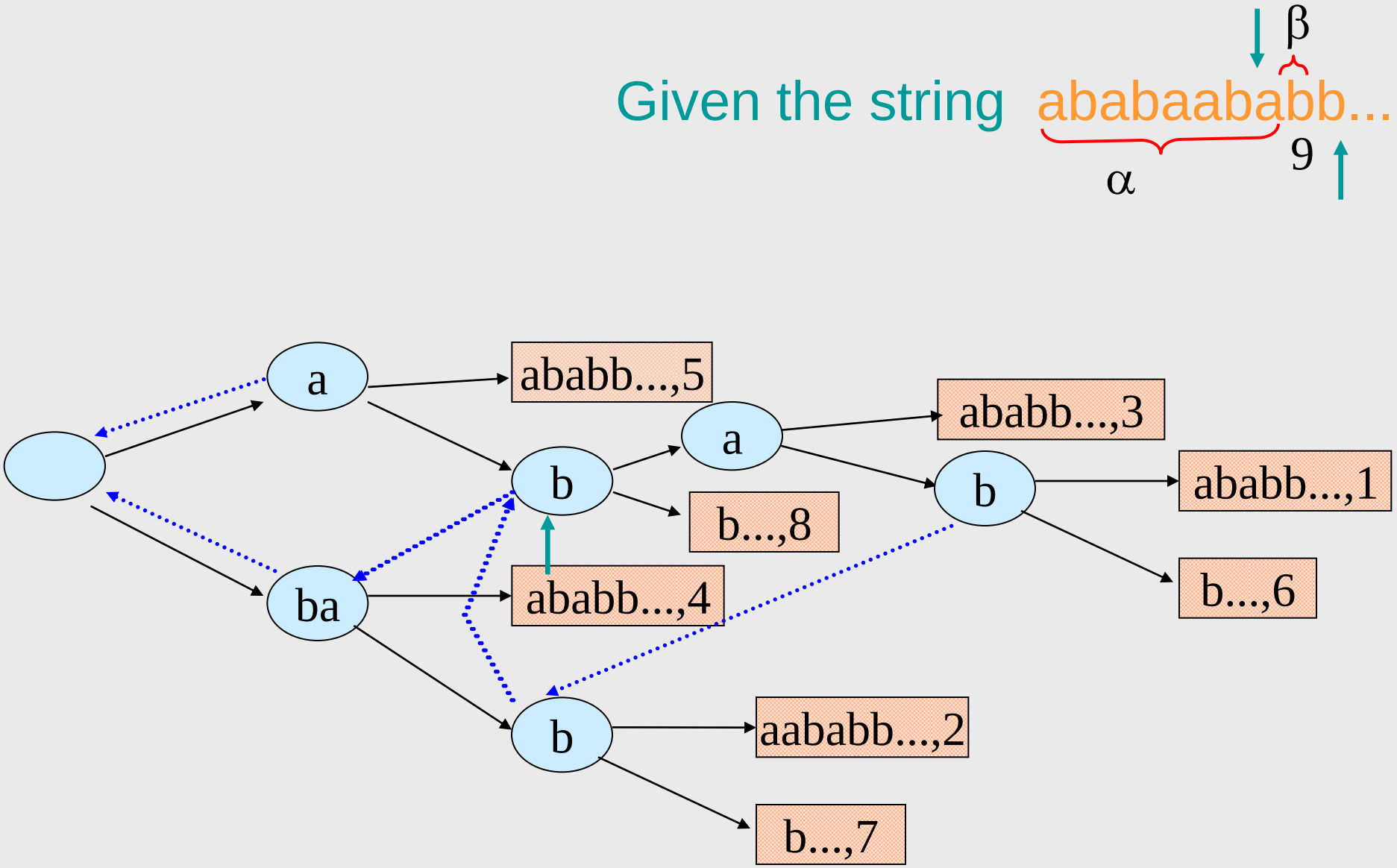
Given the string ababaababb...



Linear insertion algorithm: example

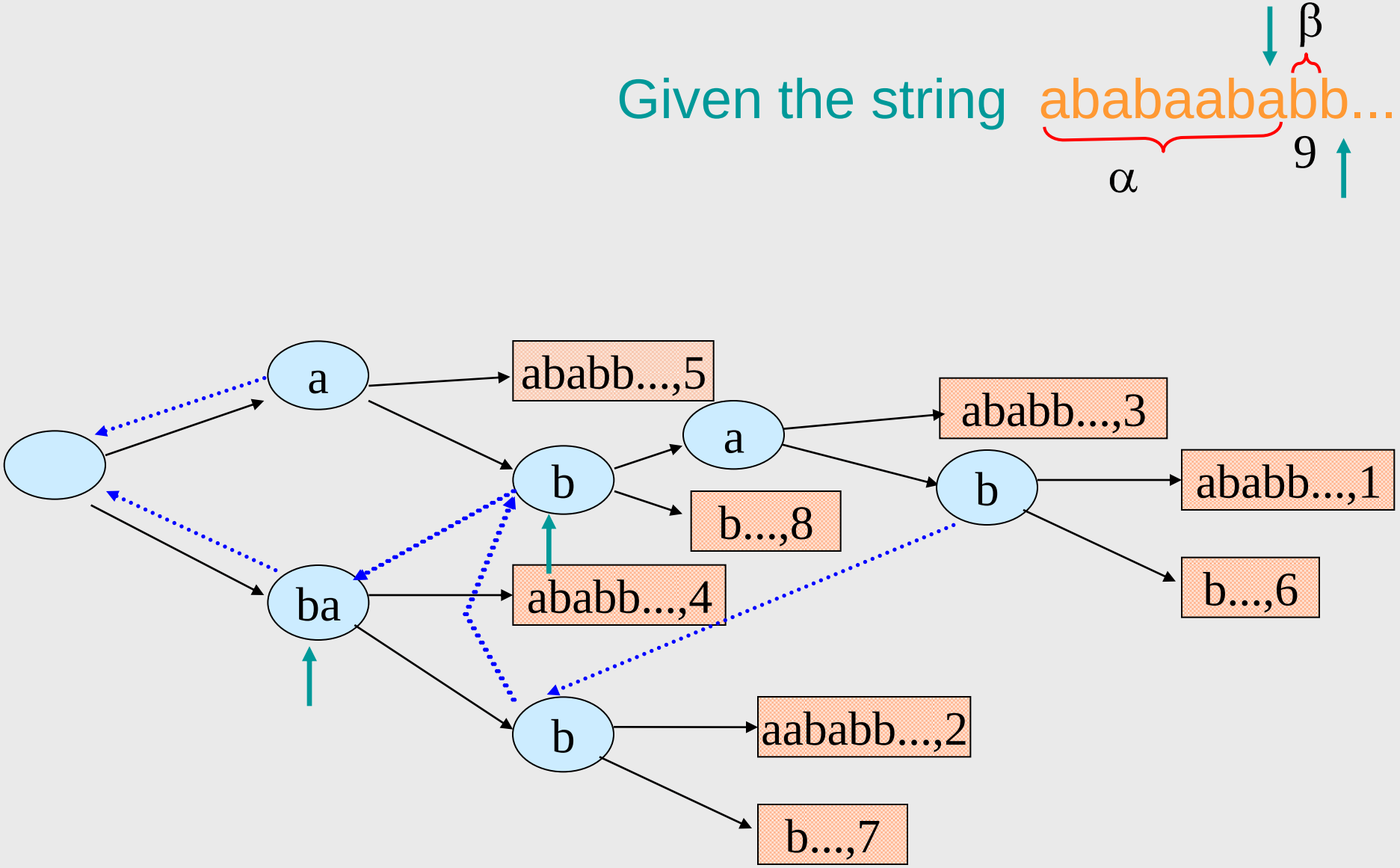


Linear insertion algorithm: example

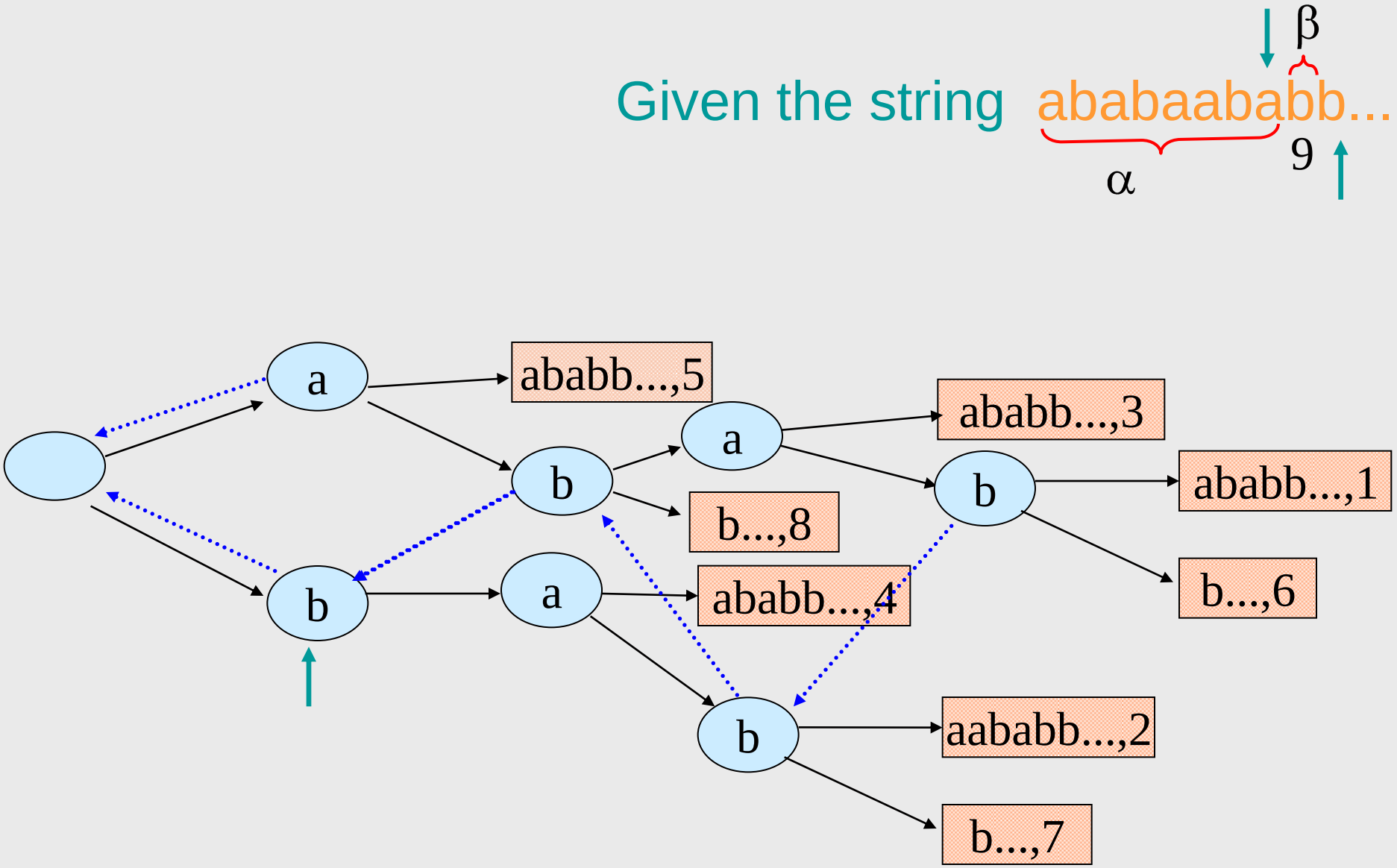


Linear insertion algorithm: example

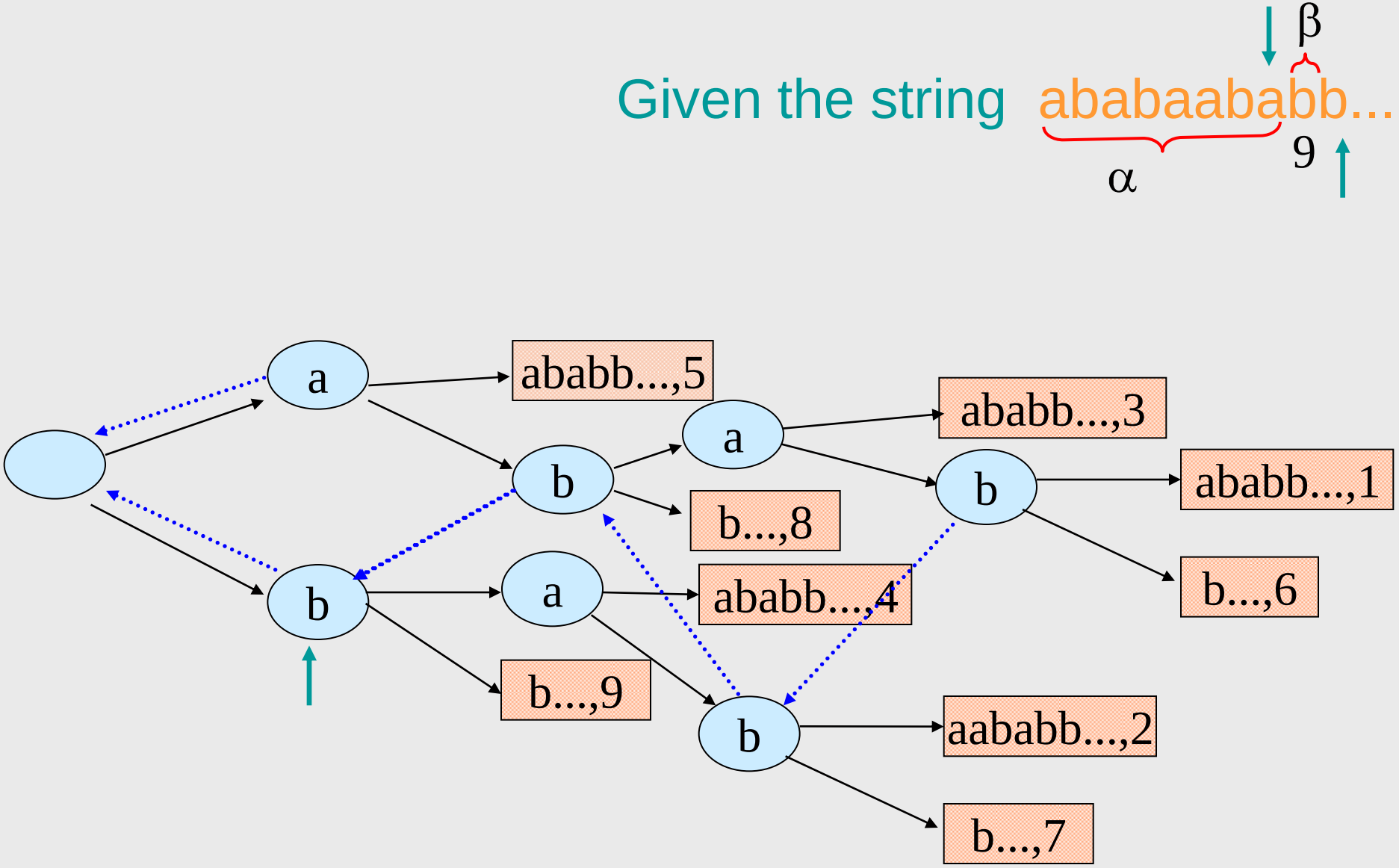
Given the string ababaababb...



Linear insertion algorithm: example



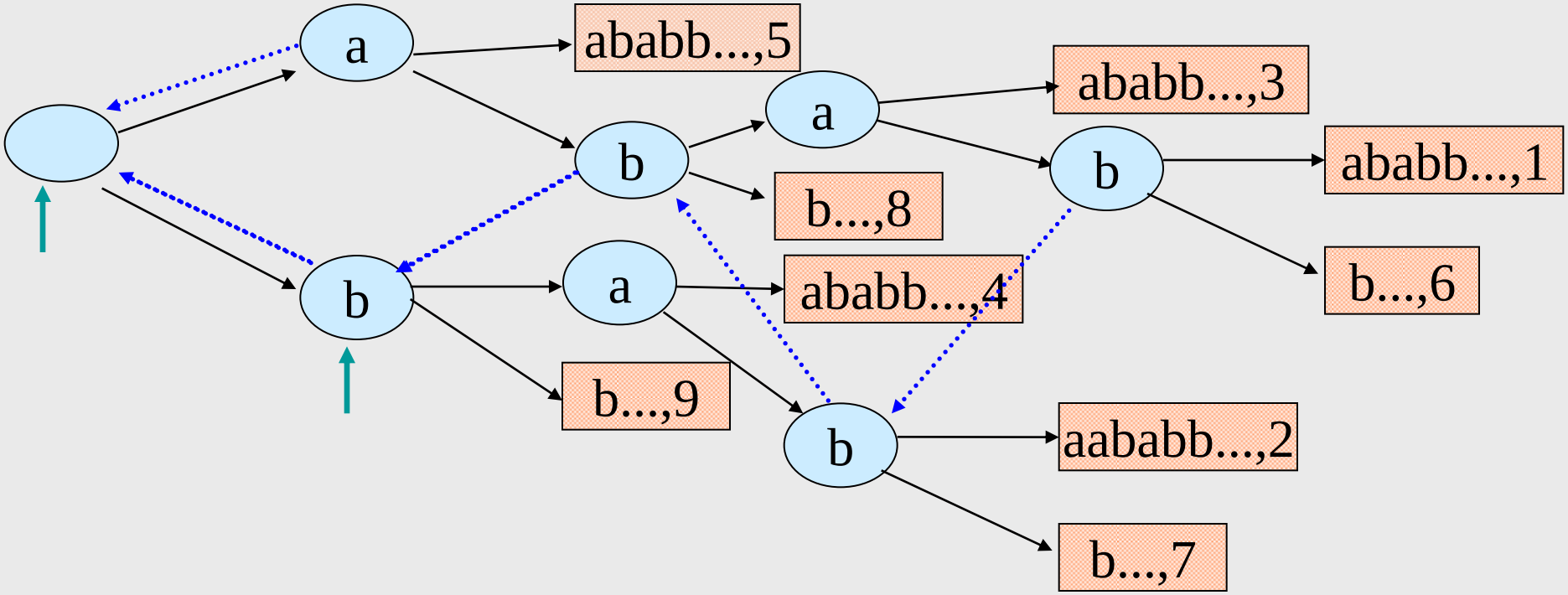
Linear insertion algorithm: example



Linear insertion algorithm: example

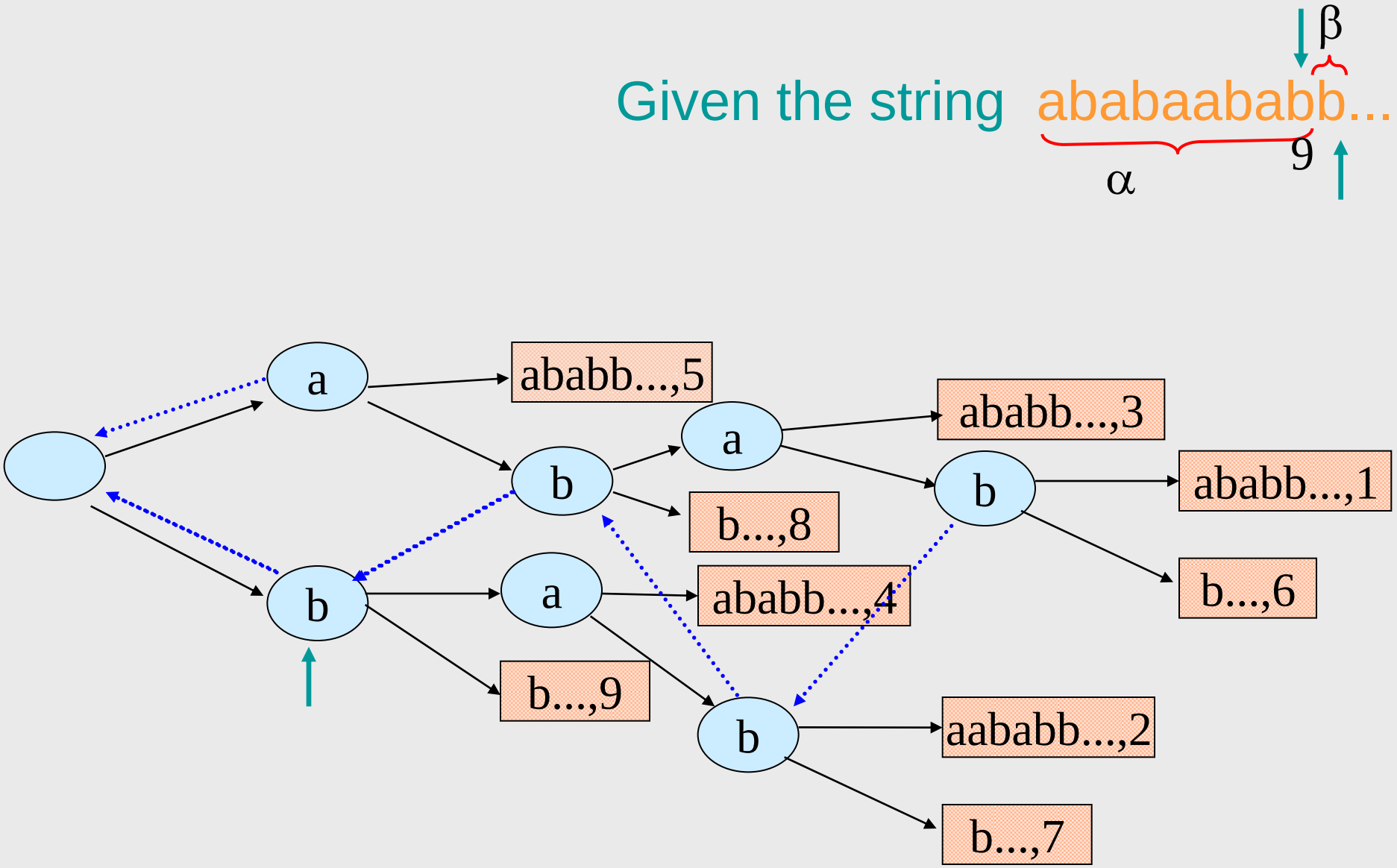
Given the string ababaababb...

α 9



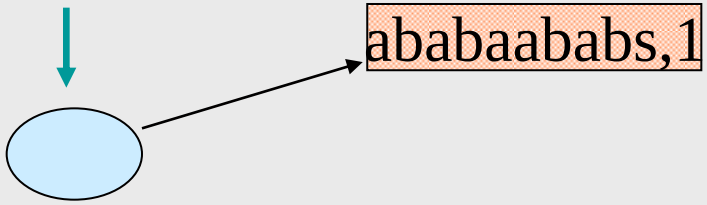
Linear insertion algorithm: example

Given the string ababaababb...



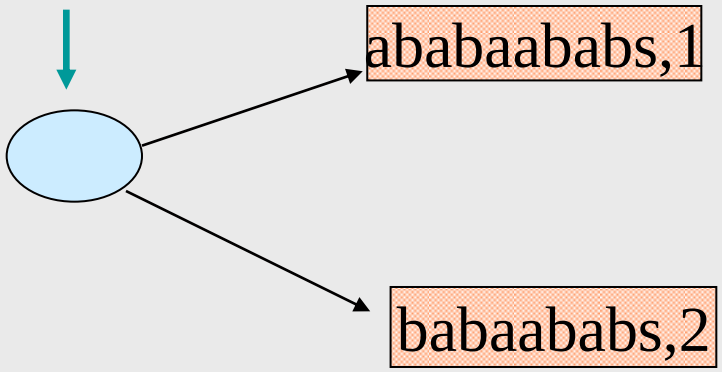
Linear insertion algorithm

Given the string ababaababs



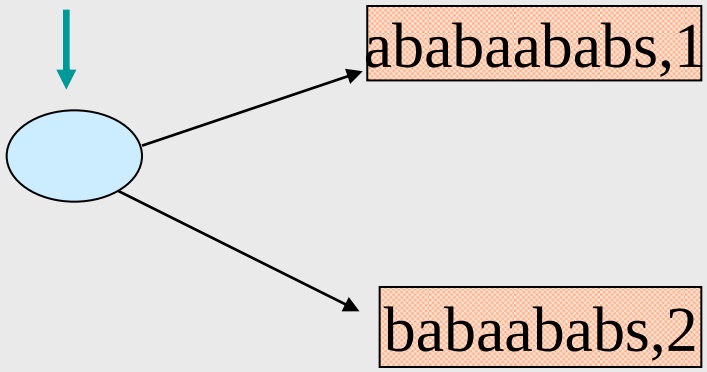
Linear insertion algorithm

Given the string ababaabab s



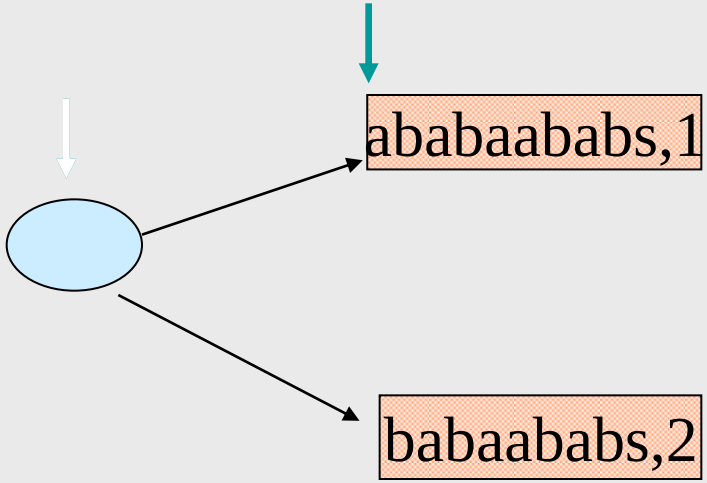
Linear insertion algorithm

Given the string ababaabab**s**



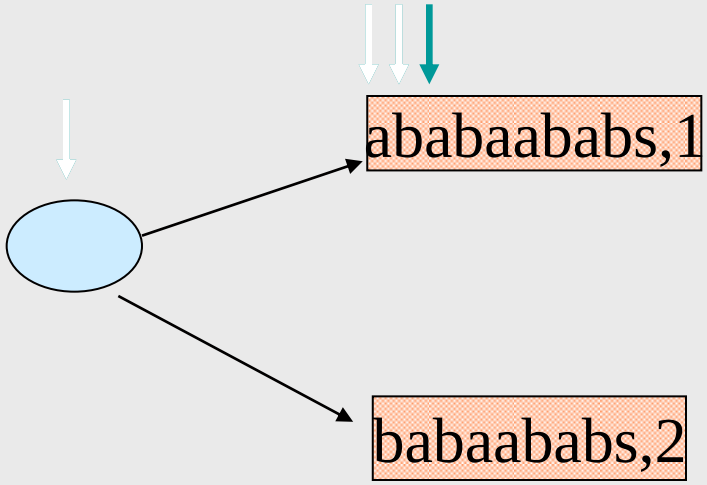
Linear insertion algorithm

Given the string ababaababs



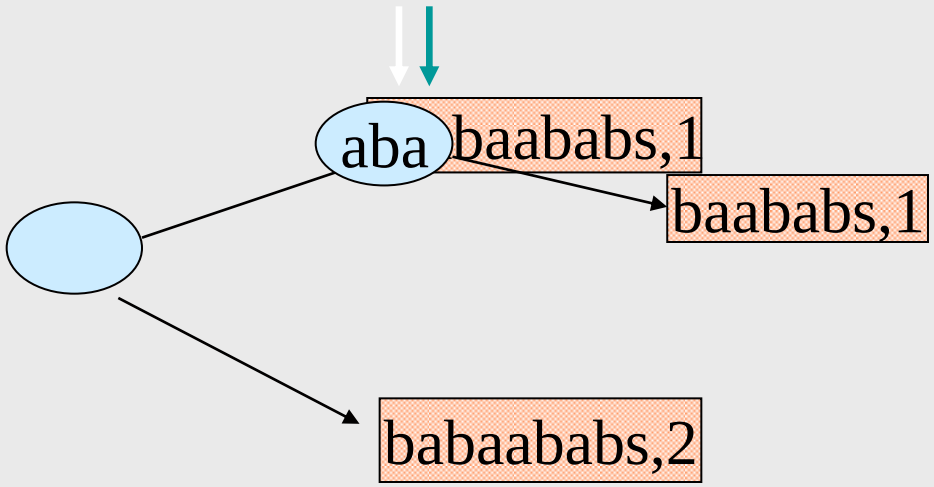
Linear insertion algorithm

Given the string ababaababs



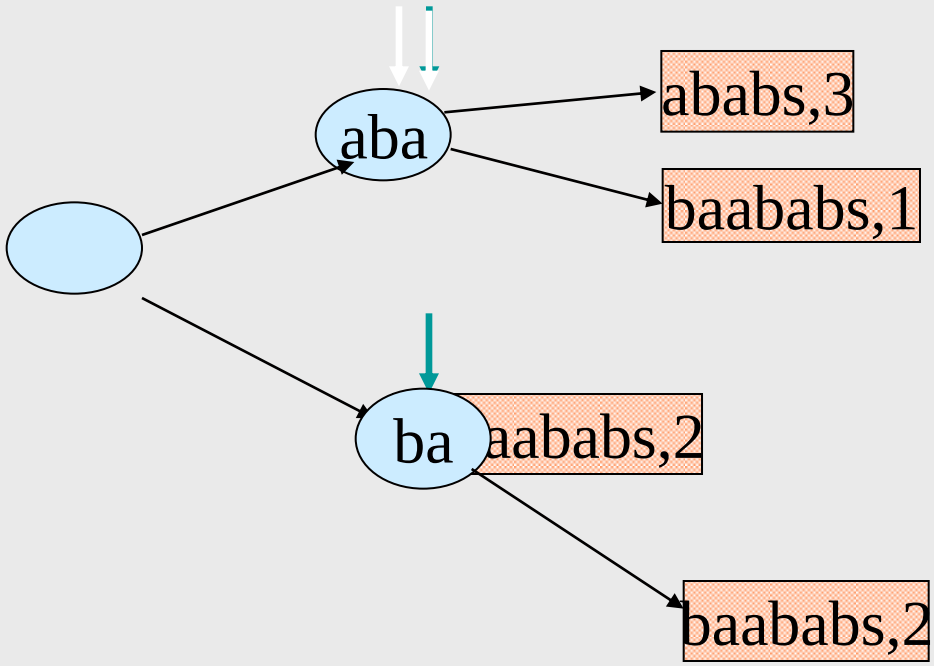
Linear insertion algorithm

Given the string ababaababs



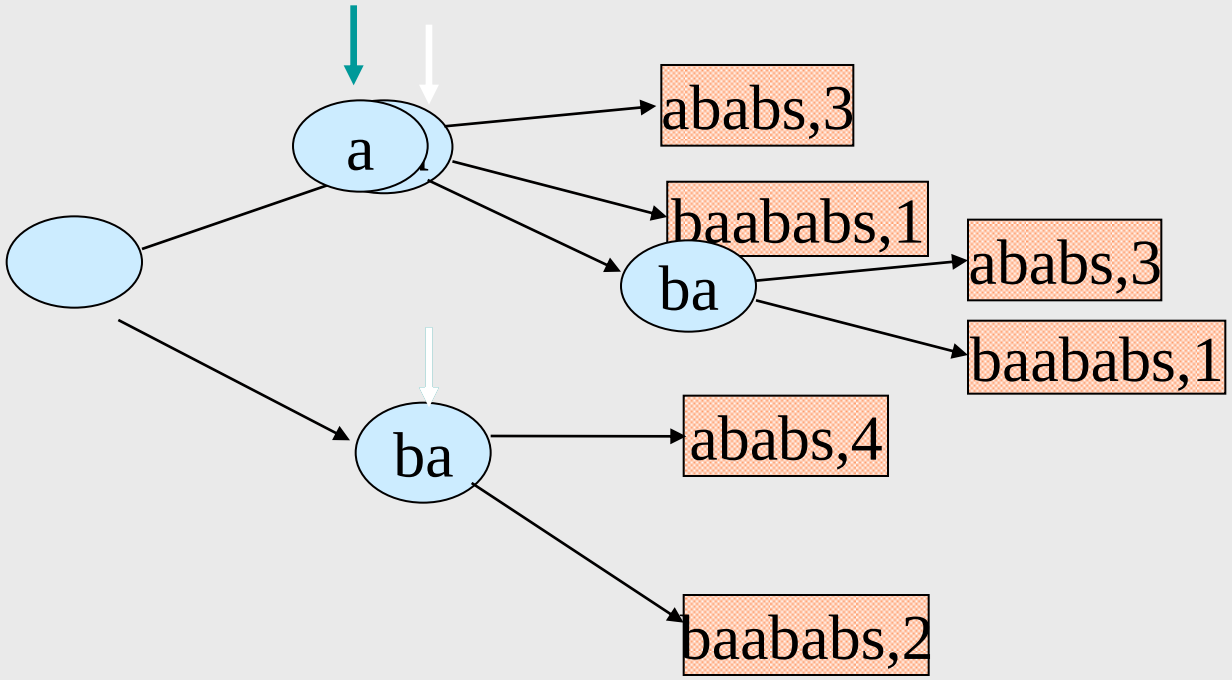
Linear insertion algorithm

Given the string ababaababs



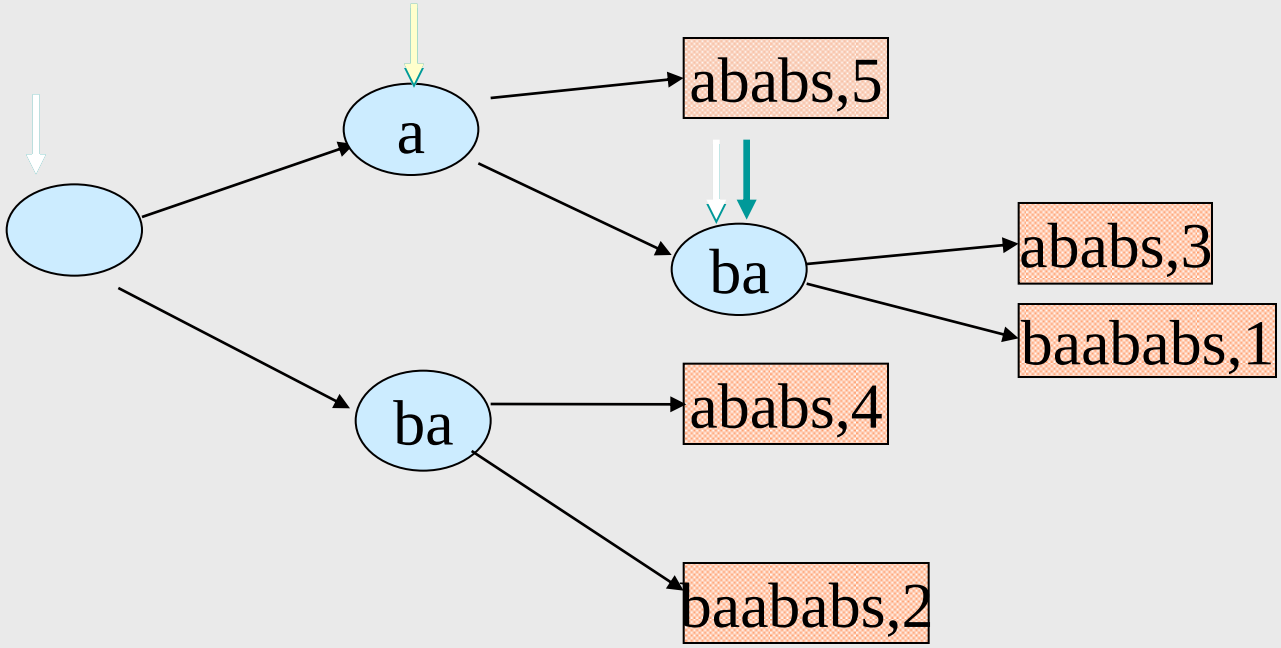
Linear insertion algorithm

Given the string ababaababs



Linear insertion algorithm

Given the string ababaababs



Index

algorithmics genetics group

1a. Part: Suffix trees

Algorithms on strings, trees and sequences,
Dan Gusfield
Cambridge University Press

2a. Part: Suffix arrays

Suffix-arrays: a new method for on-line string searches,
G. Myers, U. Manber

Suffix arrays

Given string ababaa#:

Suffixes:

1: ababaa#
2: babaa#
3: abaa#
4: baa#
5: aa#
6: a#
7: #

… but lexicographically sorted

1	1: #
2	6: a#
3	5: aa#
4	3: abaa#
5	1: ababaa#
6	4: baa#
7	2: babaa#

Which is the cost? O(n log(n))

Applications of suffix arrays

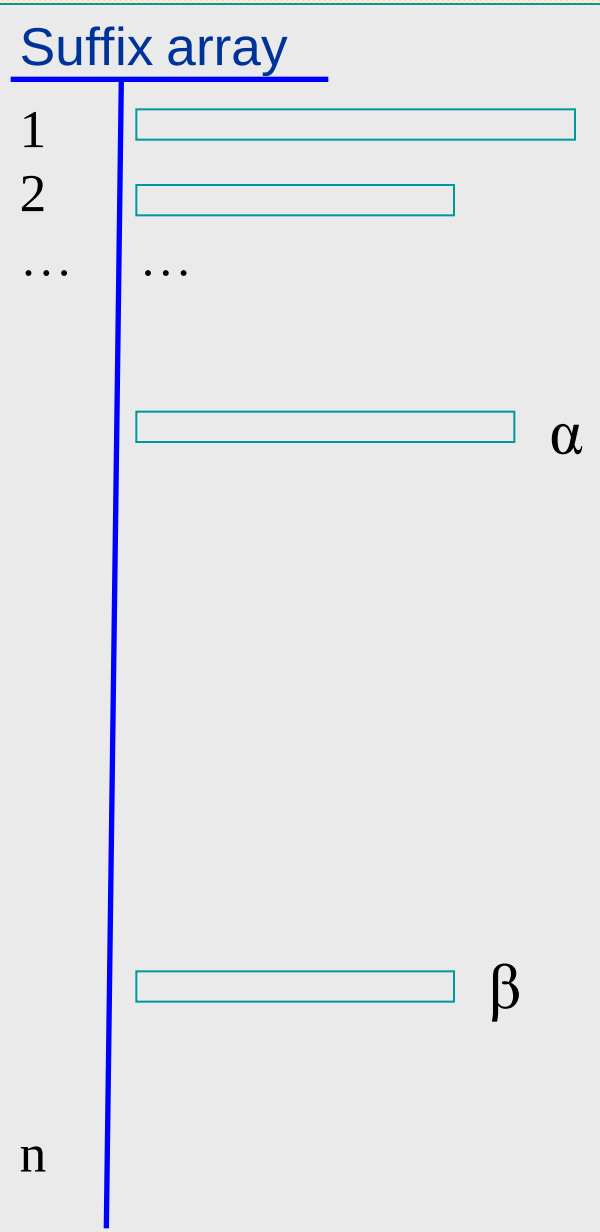
1. Exact string matching

   • Does the sequence abababaas contain any ocurrence of patterns abab, aab, and ab?

1	1: #
2	6: a#
3	5: aa#
4	3: abaa#
5	1: ababaa#
6	4: baa#
7	2: babaa#

Binary search

Search with cost O(log(n) |P|)

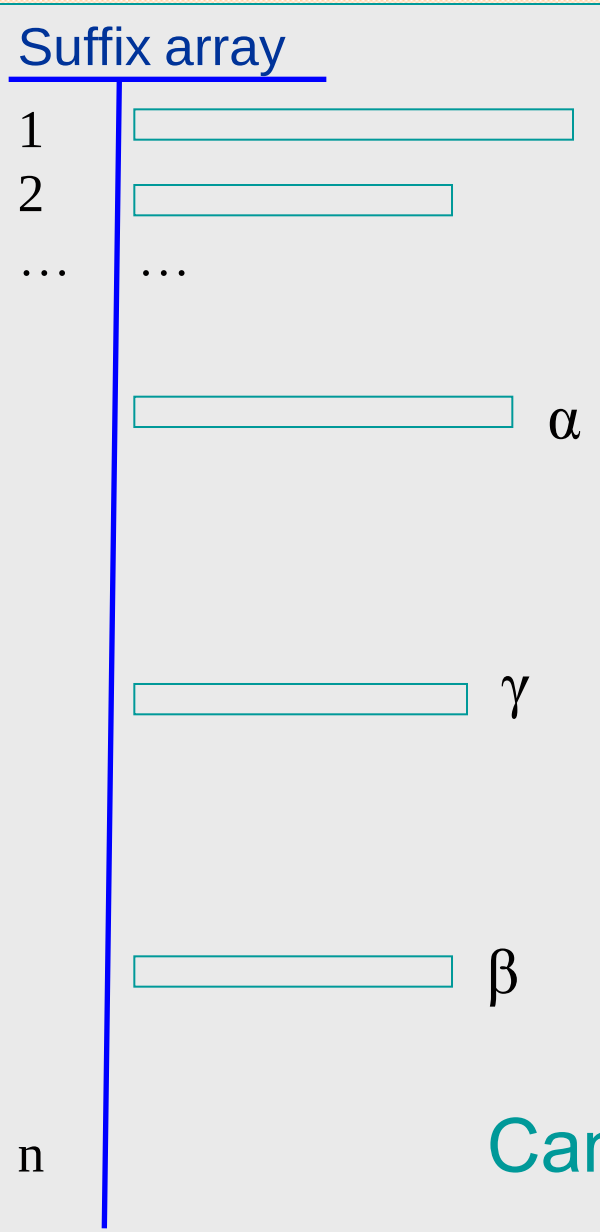


Query:

Invariant Properties:

P1: $\alpha < \text{query} \leq \beta$

Search with cost O(log(n) |P|)



Query:

Invariant Properties:

P1: $\alpha < \text{query} \leq \beta$

Algorithm:

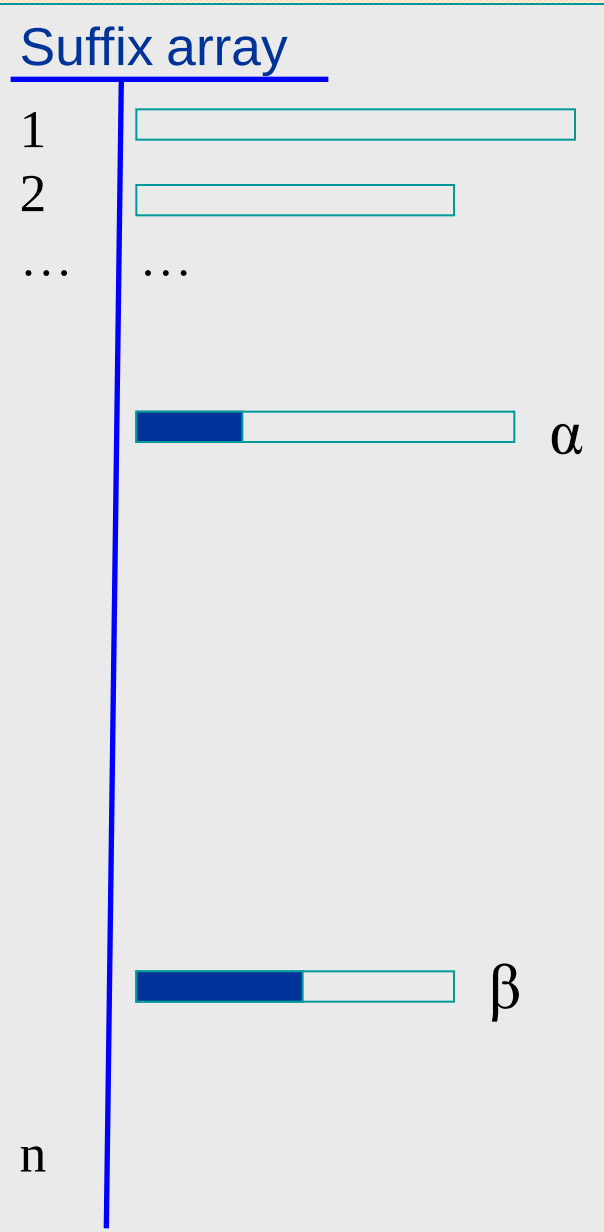
If γ<query then $\alpha = \gamma$

else $\beta = \gamma$

Cost: O(log(n) |P|)

Can it be improved to … O(log(n)+|P|) ?

Fast search with cost O(log(n)+|P|)



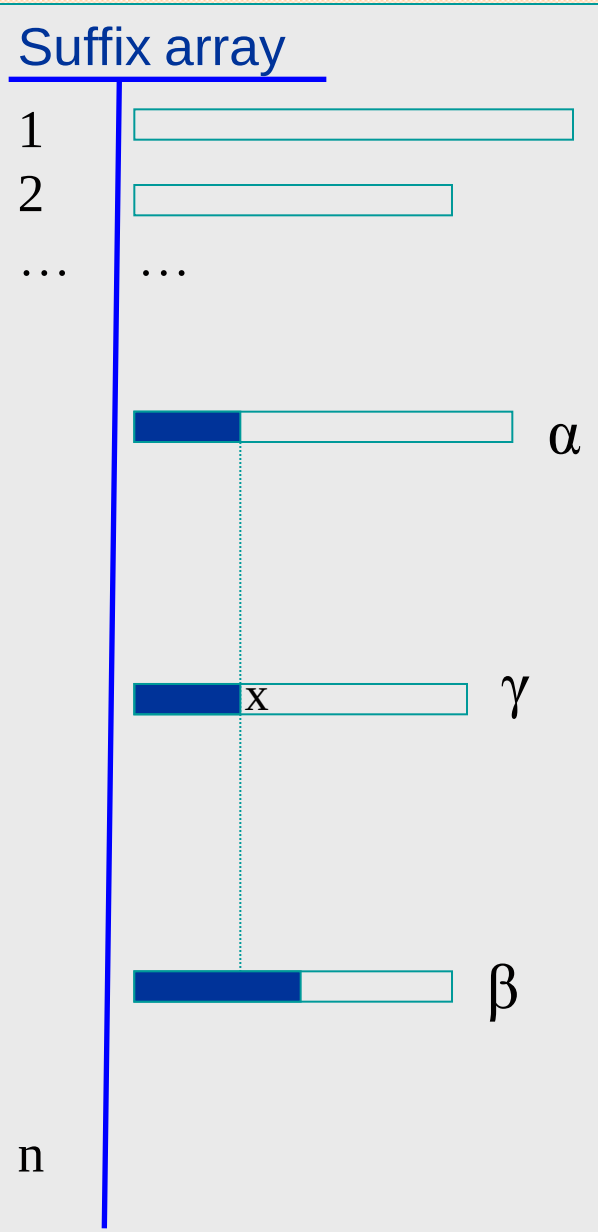
Query:

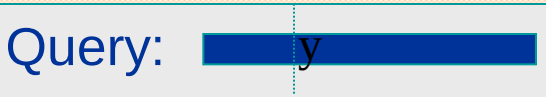
Invariant Properties:

P1: $\alpha < \text{query} \leq \beta$

P2: matches pref(query)

Fast search with cost O(log(n)+|P|)





Invariant Properties:

P1: $\alpha < query \leq \beta$

P2: ▬ matches pref(query)

Algorithm:

If

x<y then $\alpha = \gamma$

x>y then $\beta = \gamma$

x=y then …

fi