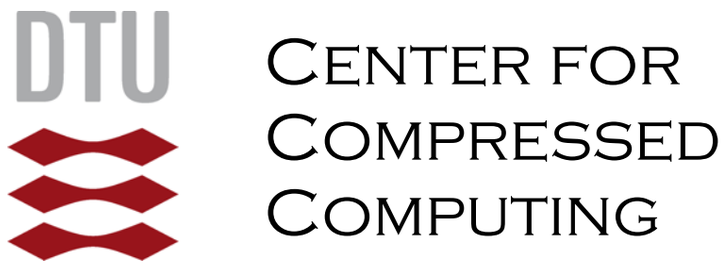
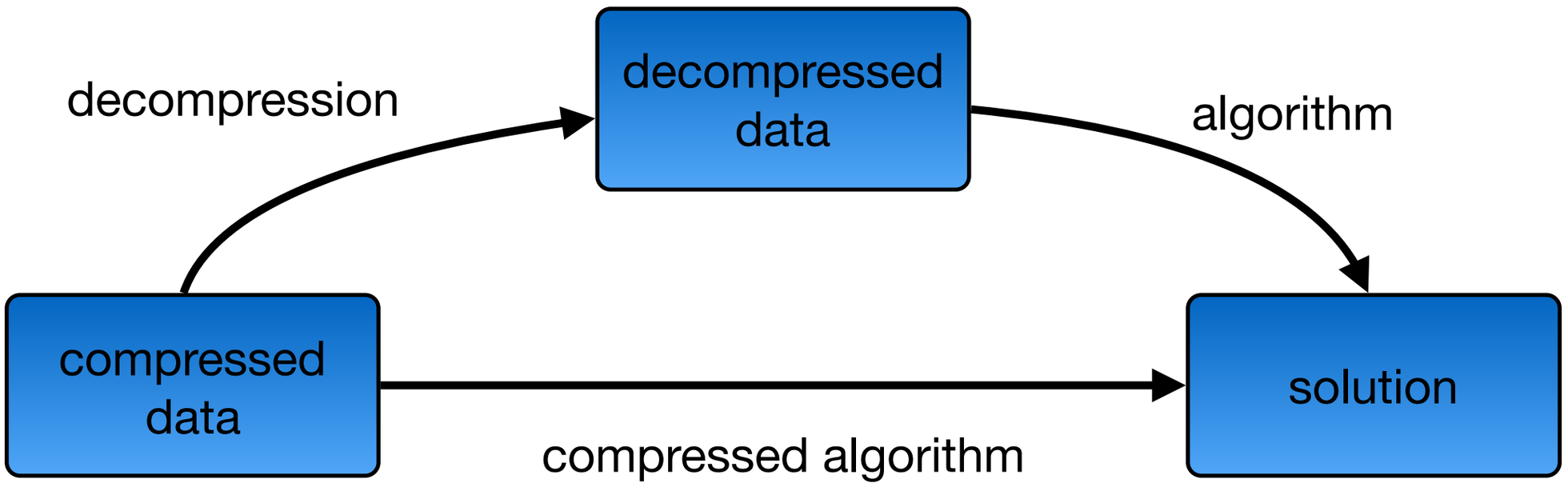
Compressed Computing

Philip Bille



Compressed Computing

- Compressed algorithms
- Compressed data structures
- Computation friendly compression schemes



Plan

- Data compression schemes.
  - Statistical compression
  - Dictionary compression
  - Grammar compression
  - Kolmogorov compression
- Random access
  - Substring decompression
  - Compressed pattern matching
  - Compressed indexing
- Repetitive collections
  - Relative compression
  - Dynamic compression

Data Compression Schemes

- Statistical compression.
  - Huffman, arithmetic encoding,...
- Dictionary compression.
  - Lempel-Ziv, ...
- Grammar based schemes.
  - Repair, sequitur, greedy, bisection, …
- Kolmogorov complexity.

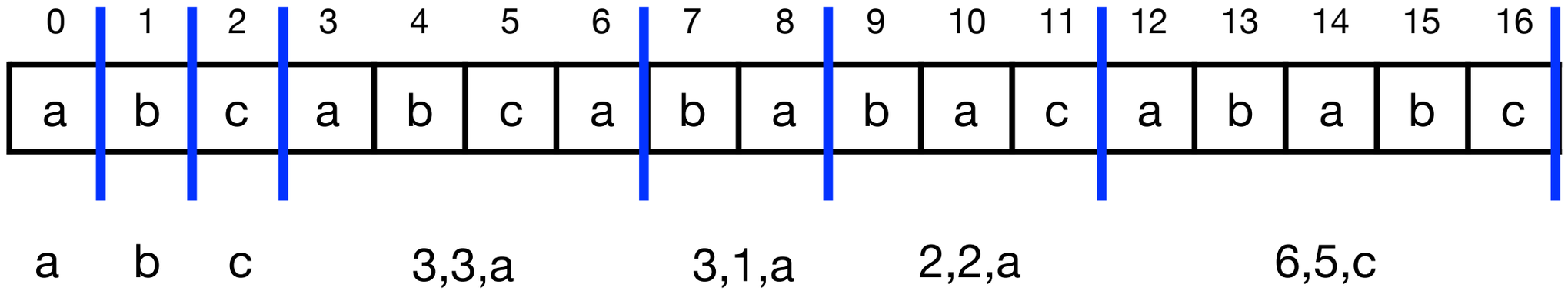
Huffman Compression

abcabcababacababc

char	frequency	code
a	7/17	0
b	6/17	10
c	4/17	11

01011010110100100110100101 1

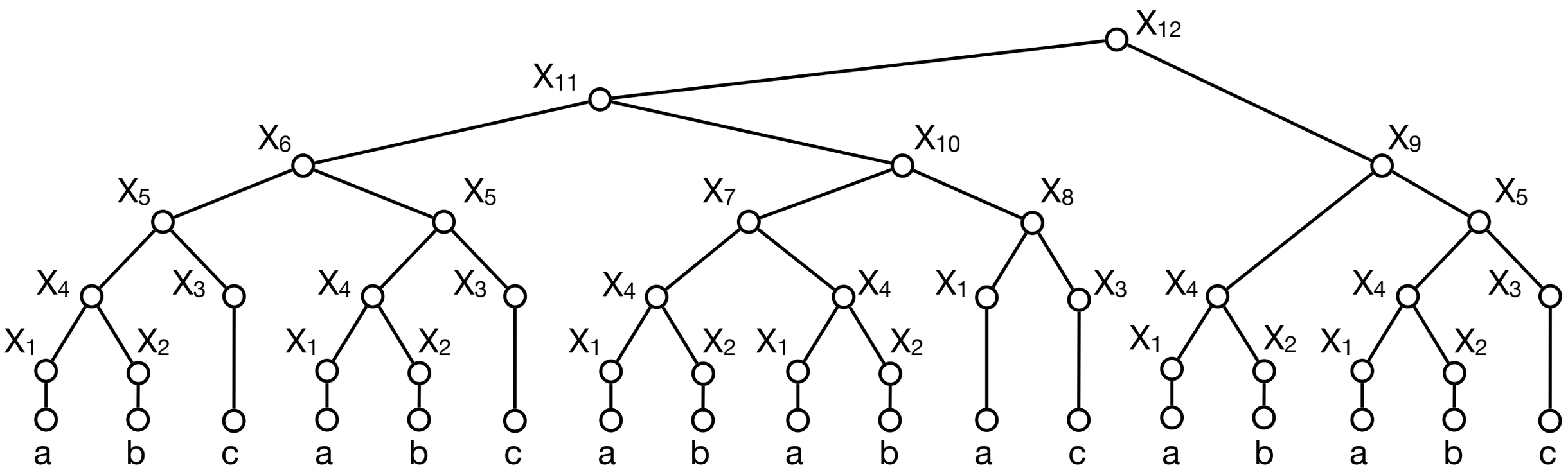
Lempel-Ziv Compression



Grammar Compression

abcabcababacababc

$X_{12} \rightarrow X_{11}X_9$
$X_{11} \rightarrow X_6X_{10}$
$X_{10} \rightarrow X_7X_8$
$X_9 \rightarrow X_4X_5$
$X_8 \rightarrow X_1X_3$
$X_6 \rightarrow X_5X_5$
$X_5 \rightarrow X_4X_3$
$X_4 \rightarrow X_1X_2$
$X_3 \rightarrow c$
$X_2 \rightarrow b$
$X_1 \rightarrow a$



Re-Pair Compression

X_9	
X_8X_6	$X_9 \rightarrow X_8X_6$
$X_3X_7X_6$	$X_8 \rightarrow X_3X_7$
$X_3X_4X_5X_6$	$X_7 \rightarrow X_4X_5$
$X_3X_4X_5X_1X_2$	$X_6 \rightarrow X_1X_2$
$X_3X_4acX_1X_2$	$X_5 \rightarrow ac$
$X_3X_1X_1acX_1X_2$	$X_4 \rightarrow X_1X_1$
$X_2X_2X_1X_1acX_1X_2$	$X_3 \rightarrow X_2X_2$
$X_1cX_1cX_1X_1acX_1X_1c$	$X_2 \rightarrow X_1c$
abcabcababacababc	$X_1 \rightarrow ab$

Lempel-Ziv vs. Grammar Compression

- Smallest grammar is NP-hard.
- LZ is lower bound on the smallest grammar.
- LZ can be converted to grammar with blowup by logarithmic factor.

Random Access

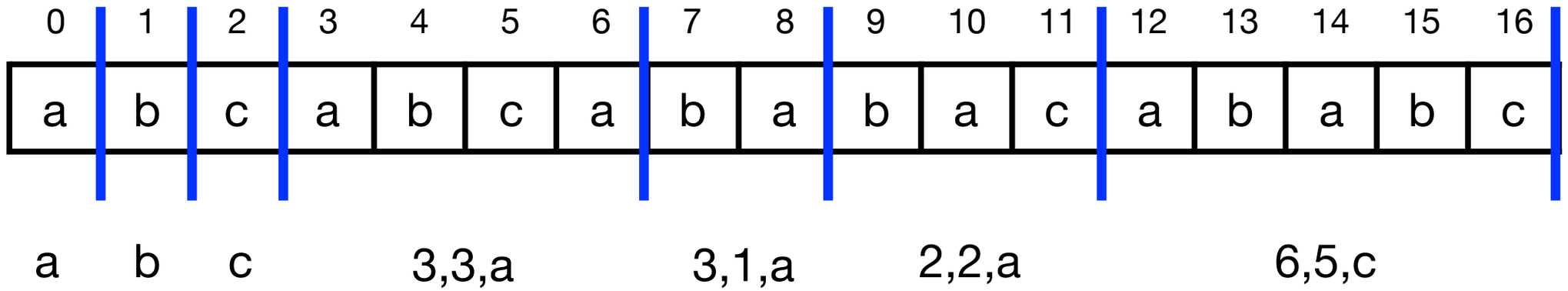
- What character is at position x in S?

(a) (b) (c) (3,3,a) (3,1,a) (2,2,a) (6,5,c)

$X_{12} \rightarrow X_{11}X_9$
$X_{11} \rightarrow X_6X_{10}$
$X_{10} \rightarrow X_7X_8$
$X_9 \rightarrow X_4X_5$
$X_8 \rightarrow X_1X_3$

$X_6 \rightarrow X_5X_5$
$X_5 \rightarrow X_4X_3$
$X_4 \rightarrow X_1X_2$
$X_3 \rightarrow c$
$X_2 \rightarrow b$
$X_1 \rightarrow a$

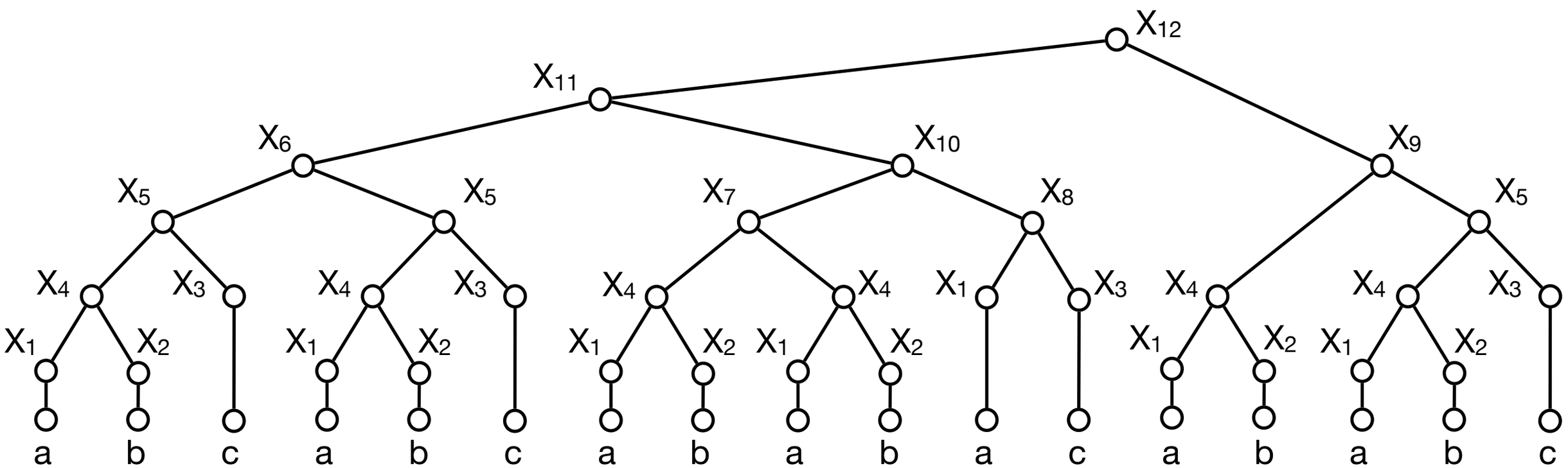
Random Access in LZ



Random Access in Grammars

$X_{12} \rightarrow X_{11}X_9$
$X_{11} \rightarrow X_6X_{10}$
$X_{10} \rightarrow X_7X_8$
$X_9 \rightarrow X_4X_5$
$X_8 \rightarrow X_1X_3$

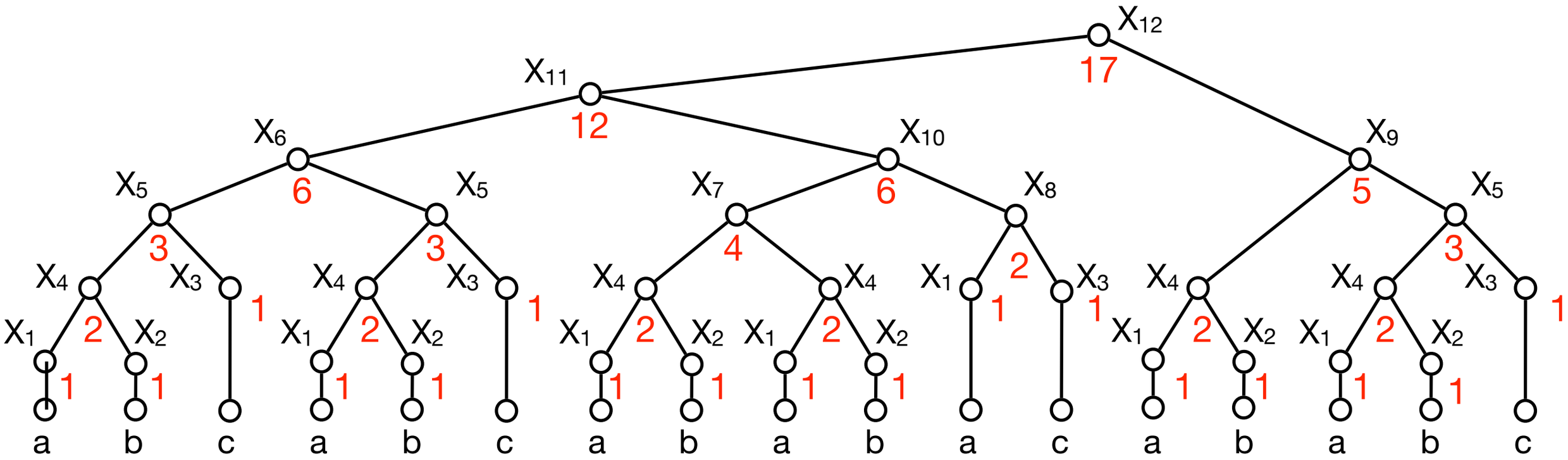
$X_6 \rightarrow X_5X_5$
$X_5 \rightarrow X_4X_3$
$X_4 \rightarrow X_1X_2$
$X_3 \rightarrow c$
$X_2 \rightarrow b$
$X_1 \rightarrow a$



N = length of string
n = size of grammar

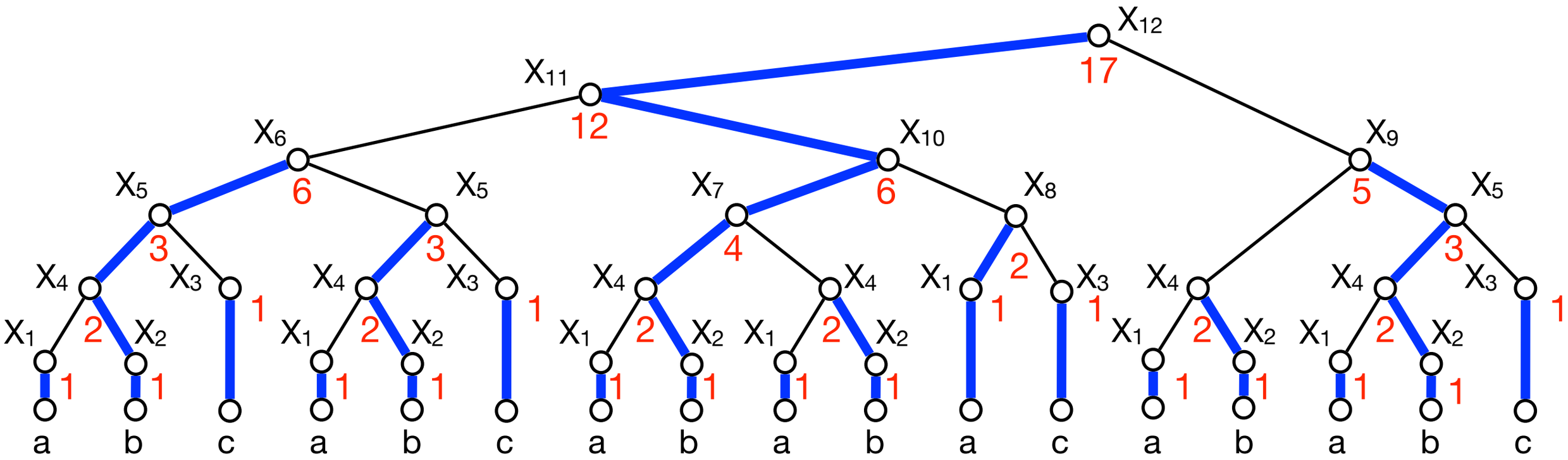
Goal: O(n) space and O(log N) time

Top Down Search



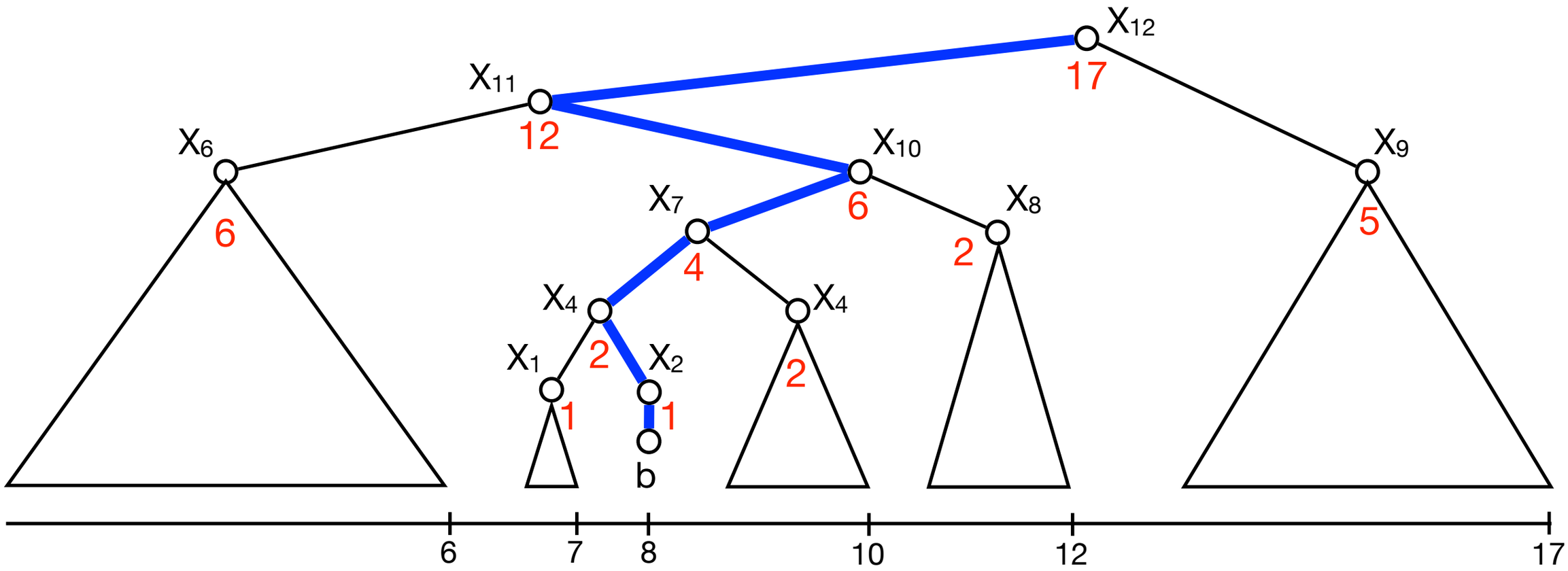
- Time. O(h) = O(n)
- Space. O(n)

Heavy-Path Decomposition



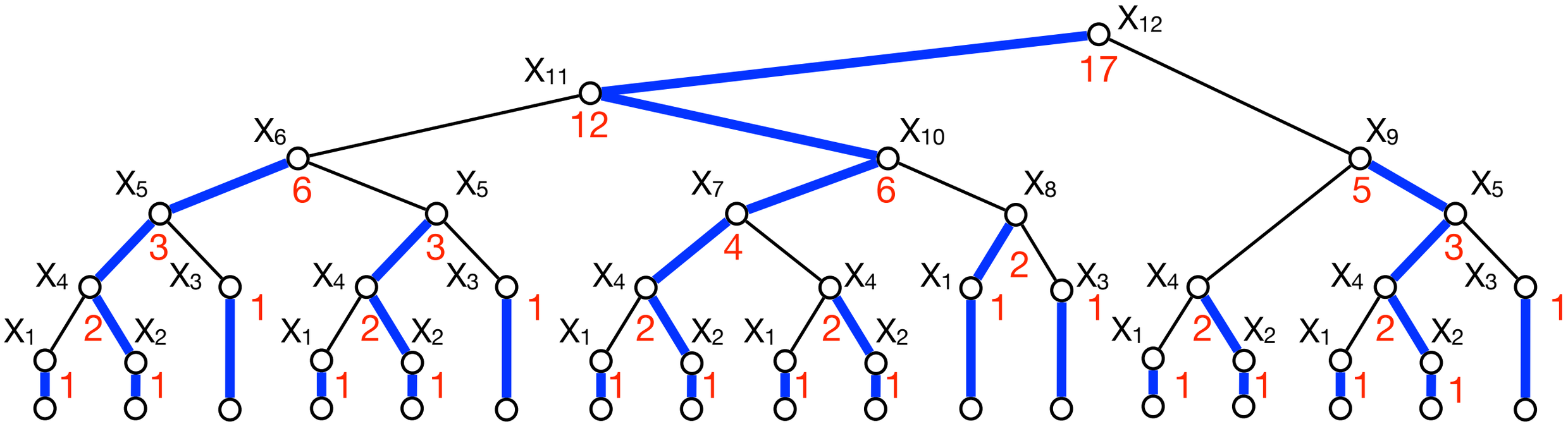
- O(log N) heavy paths on any root-to-leaf path.

Searching a Heavy-Path

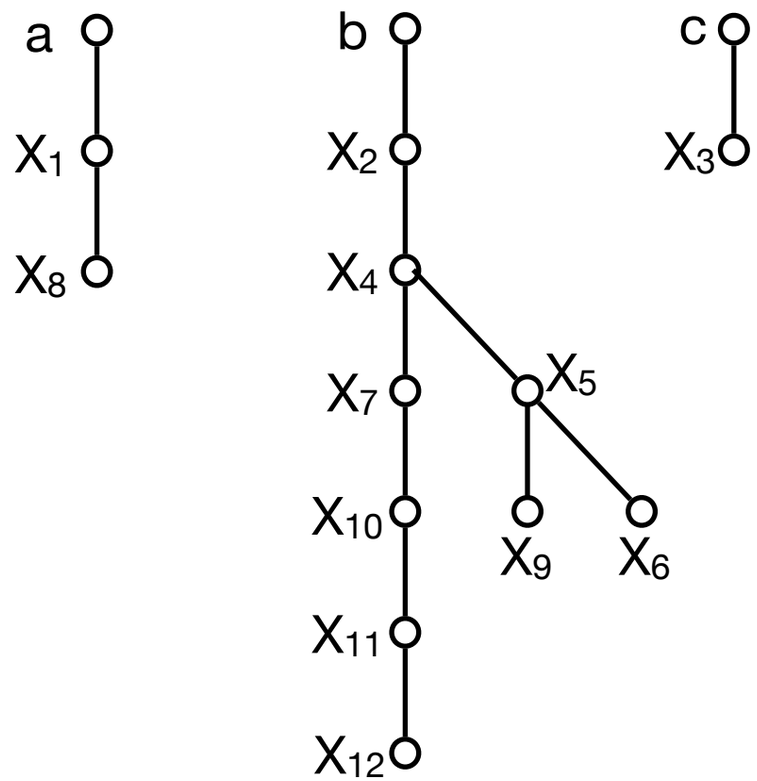


- Time. O(log N log n)
- Space. O(n^2)

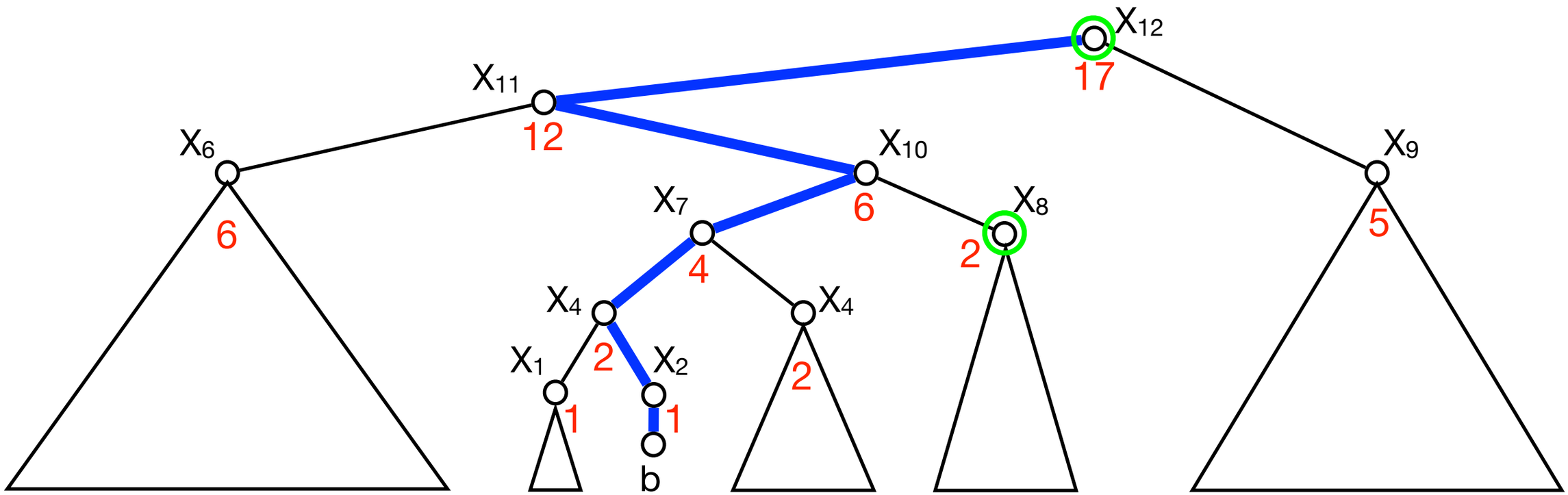
Compact Representation for Heavy-Paths



- Time. O(log N log n)
- Space. O(n)



Biased Search



$$\log\left(\frac{S(X_{12})}{S(X_8)}\right)+\log\left(\frac{S(X_8)}{S(X_5)}\right)+\log\left(\frac{S(X_5)}{S(X_3)}\right)+\cdots$$

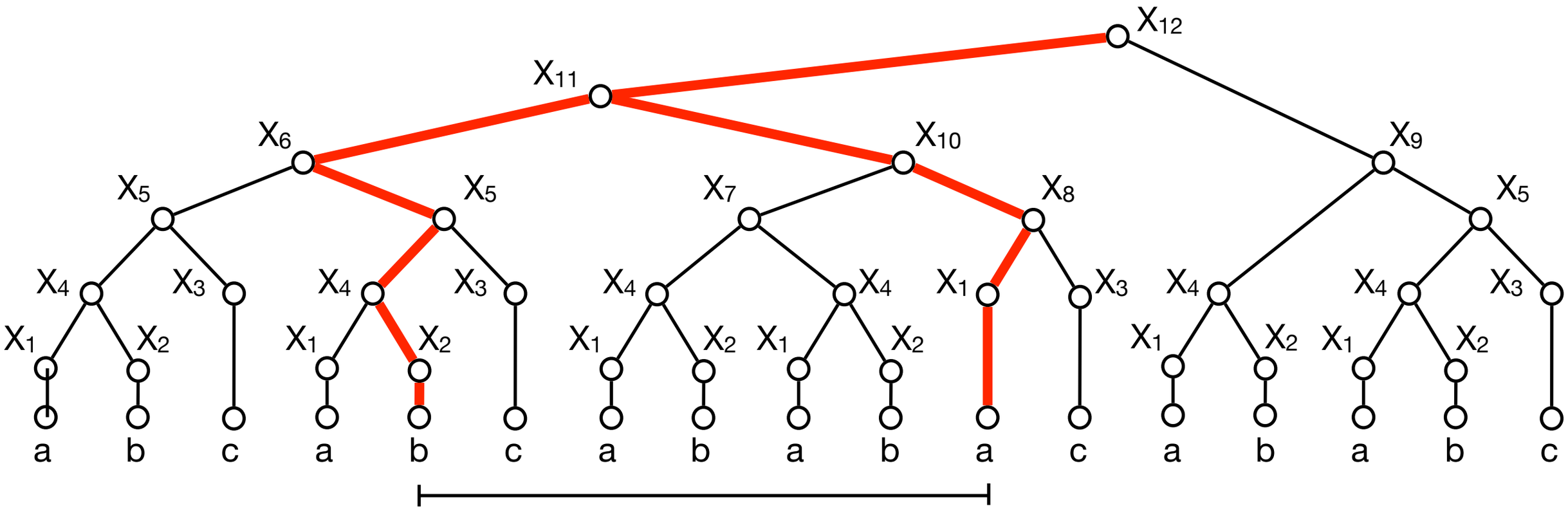
$$=\log(S(X_{12}))-\log(S(X_8))+\log(S(X_8))-\log(S(X_5))+\log(S(X_5))-\log(S(X_3))+\cdots$$

$$=O(\log N)$$

Random Access

Time	Space
O(n)	O(n)
O(log N log n)	$O(n^2)$
O(log N log n)	O(n)
O(log N)	O(n)

Substring Decompression

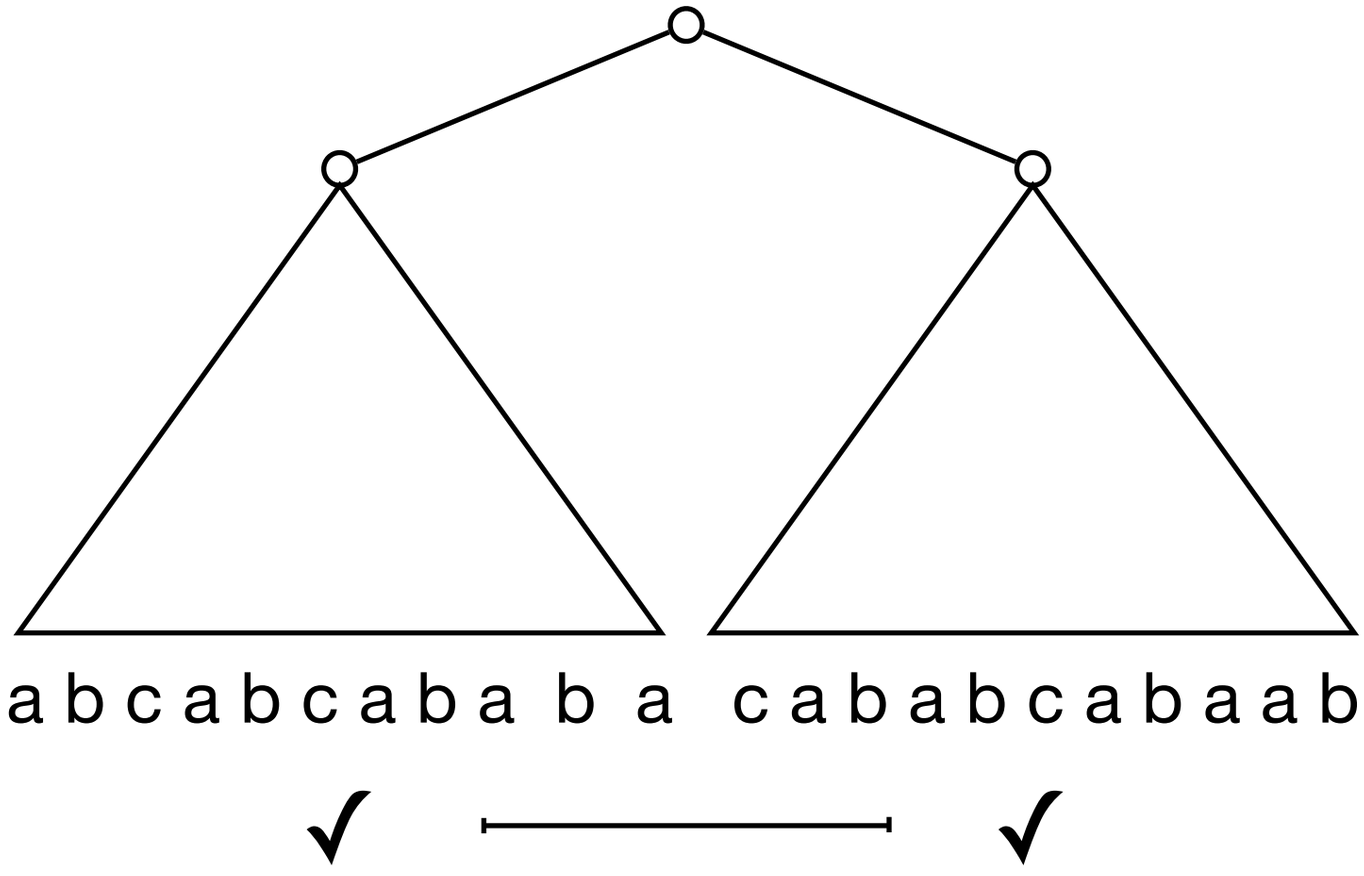


Compressed Pattern Matching

- Does "bac" occur in S (perhaps with k errors)?

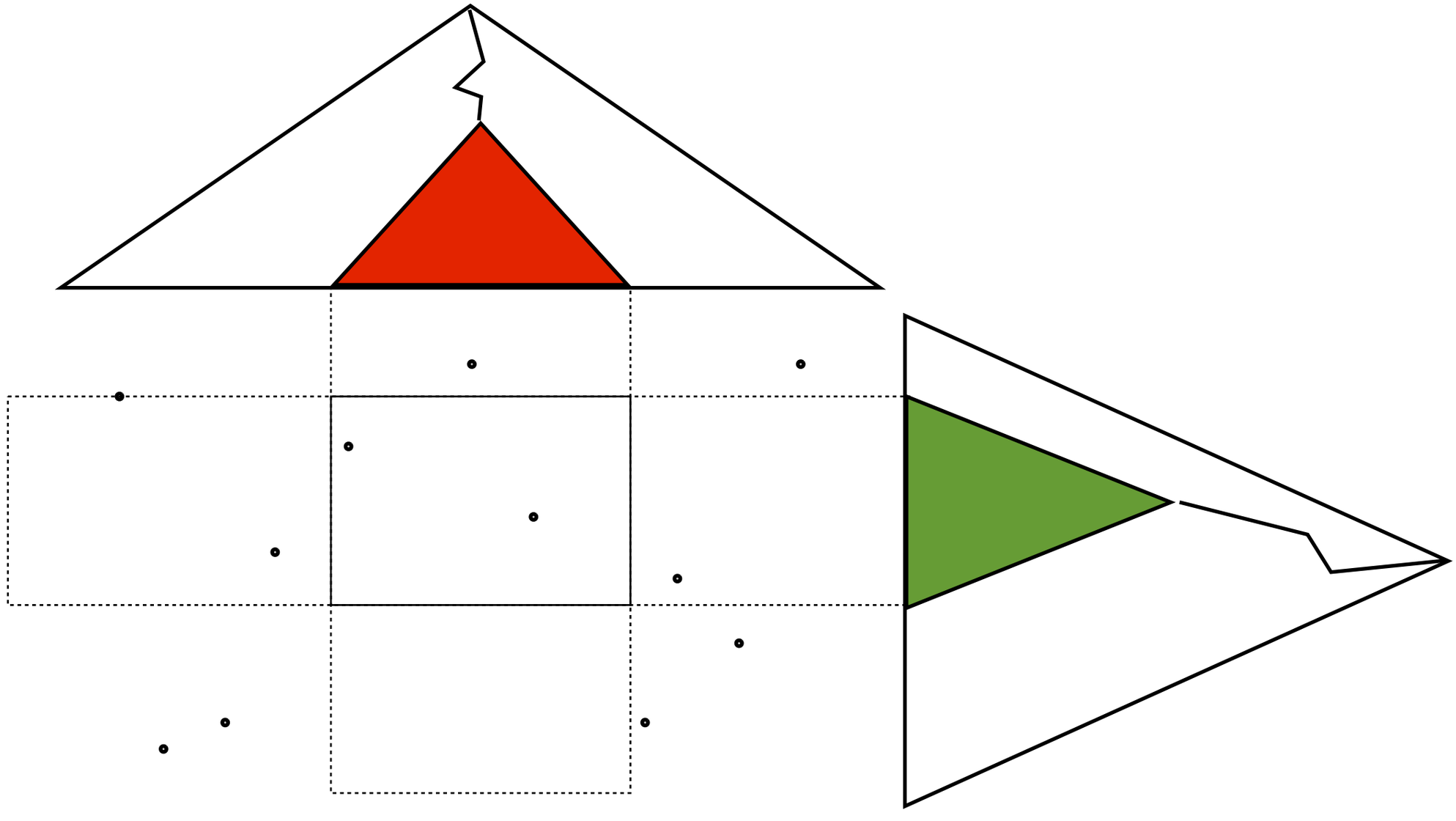
$X_{12} \rightarrow X_{11}X_9$
$X_{11} \rightarrow X_6X_{10}$
$X_{10} \rightarrow X_7X_8$
$X_9 \rightarrow X_4X_5$
$X_8 \rightarrow X_1X_3$

$X_6 \rightarrow X_5X_5$
$X_5 \rightarrow X_4X_3$
$X_4 \rightarrow X_1X_2$
$X_3 \rightarrow c$
$X_2 \rightarrow b$
$X_1 \rightarrow a$

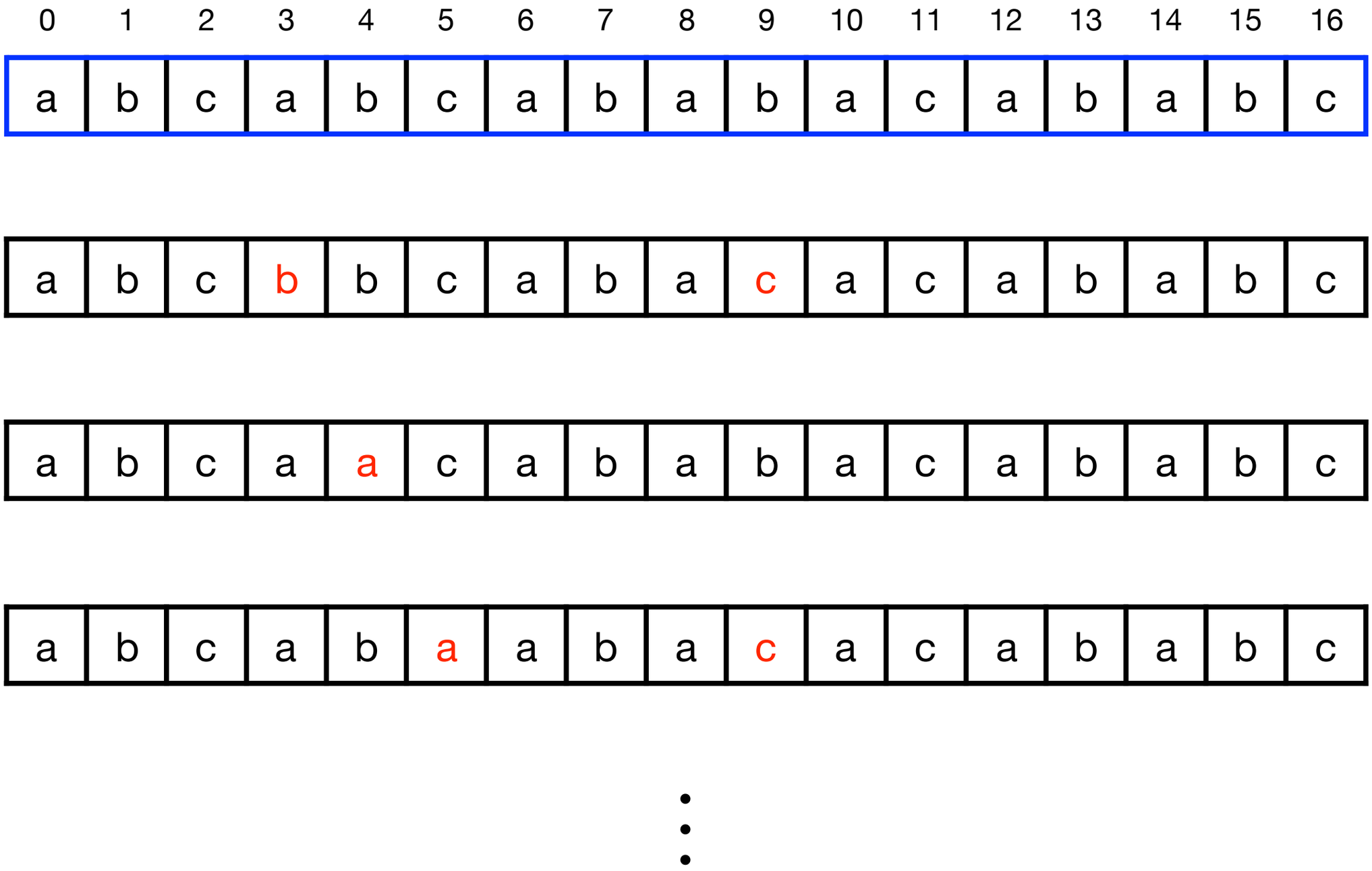


Compressed Indexing

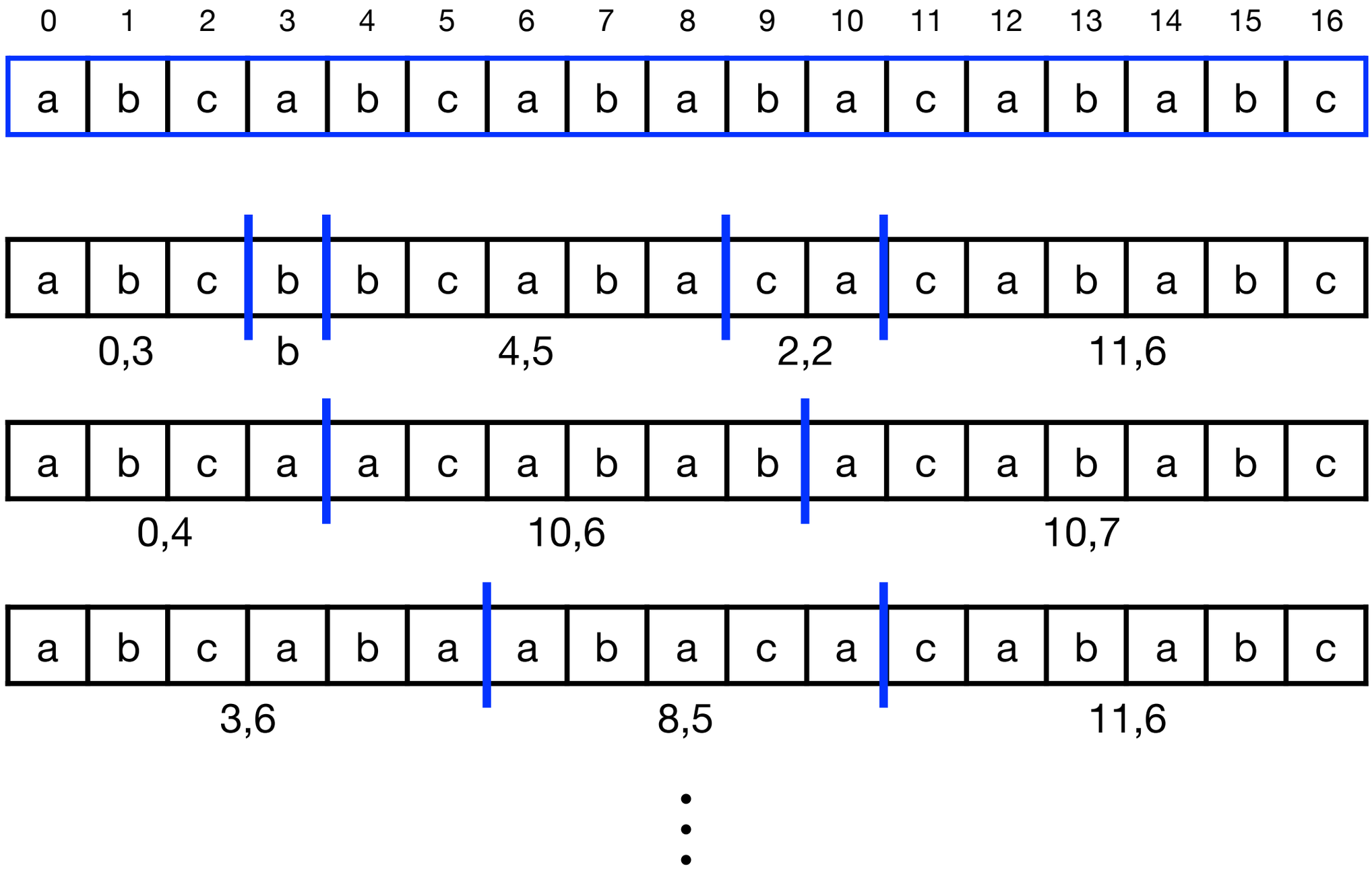
- Index compressed string S for fast lookup.



Repetitive Collections

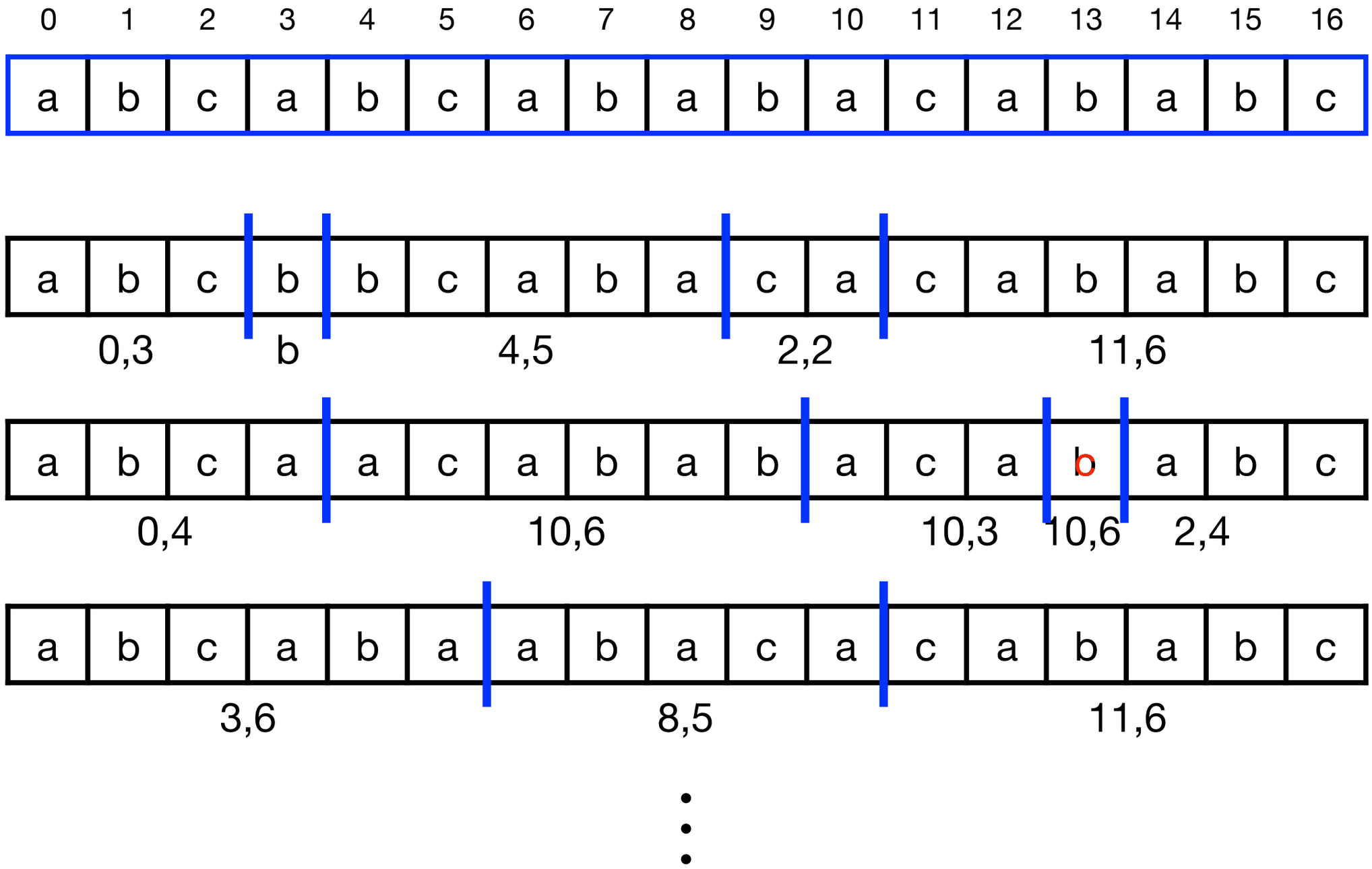


Relative LZ Compression



Dynamic Relative LZ Compression

- Change "b" to "c" at position 45.



Plan

- Data compression schemes.
  - Statistical compression
  - Dictionary compression
  - Grammar compression
  - Kolmogorov compression
- Random access
  - Substring decompression
  - Compressed pattern matching
  - Compressed indexing
- Repetitive collections
  - Relative compression
  - Dynamic compression
- Directions