Practical fast on-line exact pattern matching algorithms for highly similar sequences

Nadia Ben Nsira Thierry Lecroq Élise Prieur-Gaston

LITIS EA 4108, Normastic FR3638, IRIB, Université de Rouen Normandie, Normandie Université, France

Workshop SeqBio 2018, November 19th, 2018

Table of contents

1. Introduction and notations
2. Search in highly similar sequences

Table of contents

1. Introduction and notations
2. Search in highly similar sequences

Big data

- NGS technologies output numerous individual genomes of the same species
- More than 99% similar

Highly similar sequences

- Differ from the reference by: SNVs (SNPs), indels, CNVs, translocations, ...
- Common and non-common parts

Efficient solutions

- Strong need for efficient indexing and pattern matching

Pattern matching

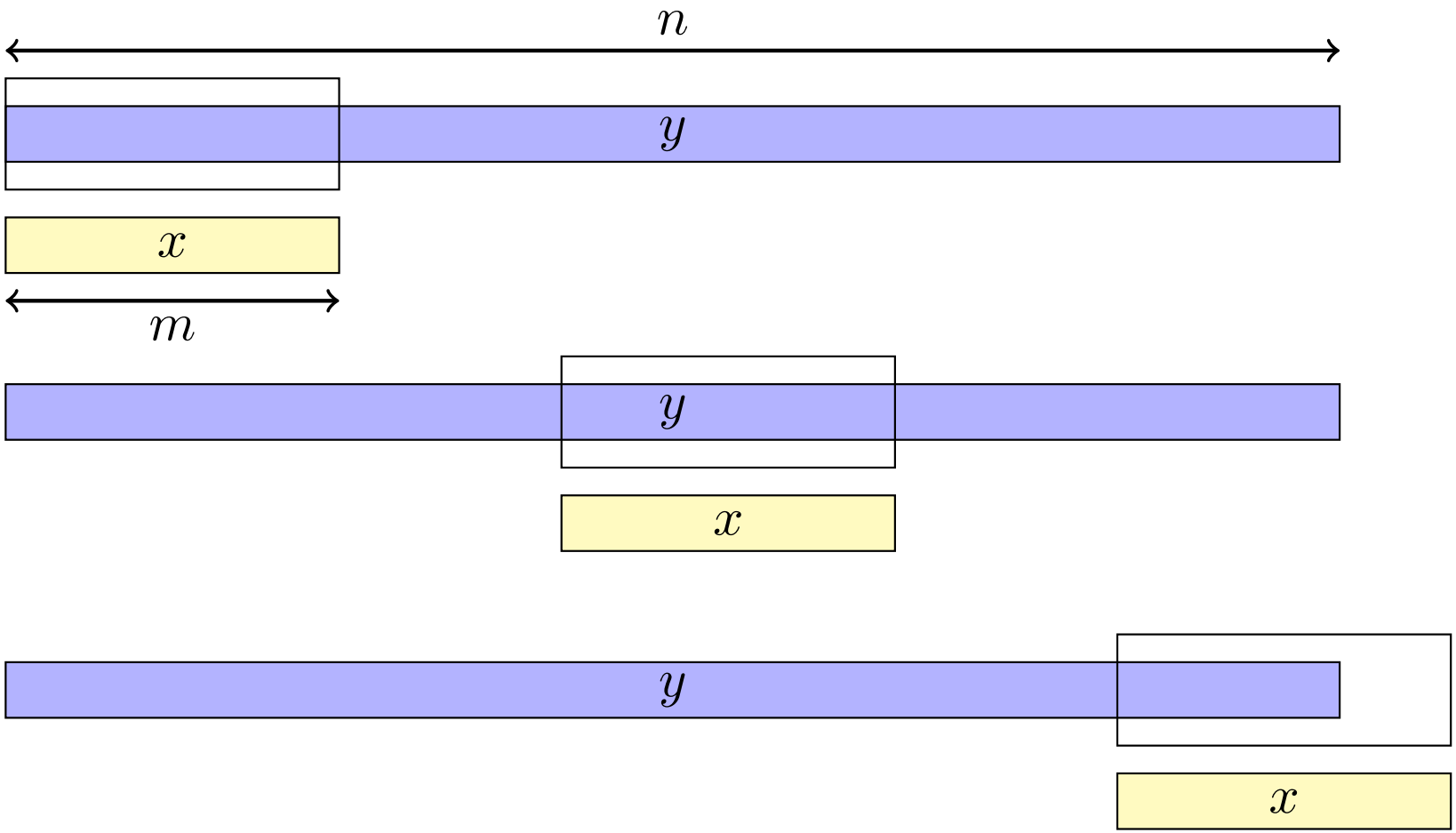
Find one(all the) position(s) of a pattern of length m in a sequence of length n:

- with index $\rightarrow O(m)$
- without index $\rightarrow O(n)$

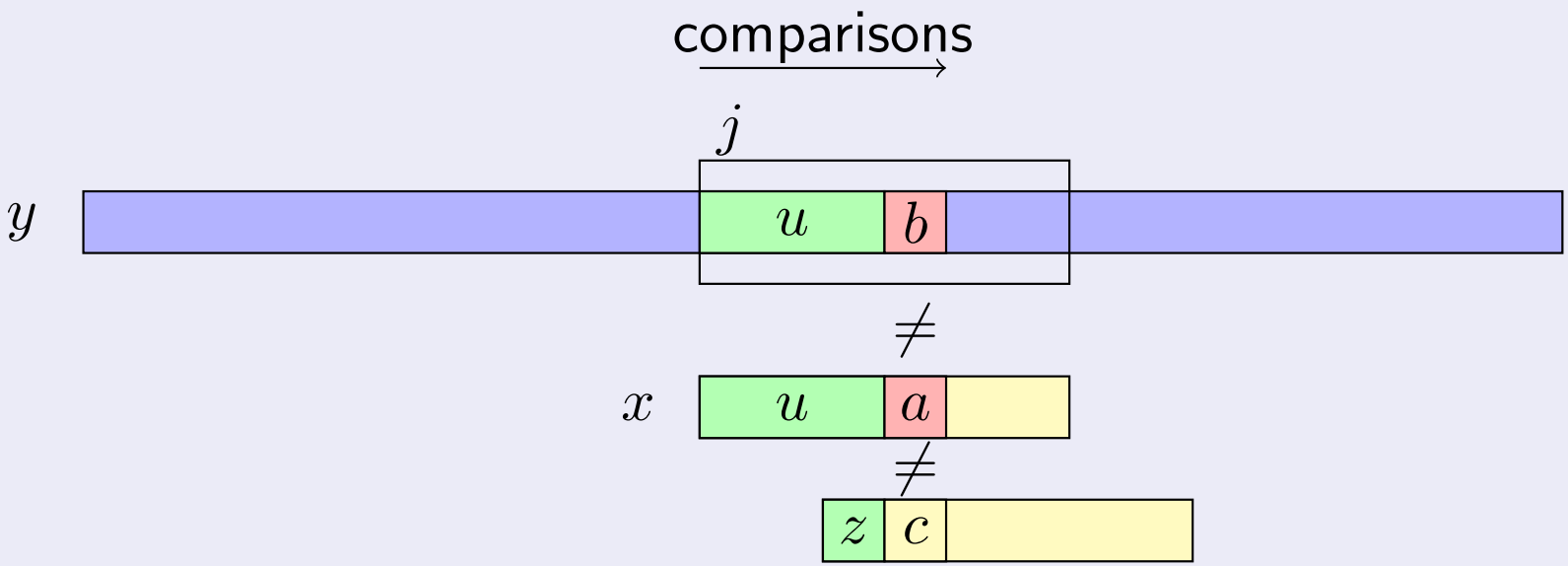
Notations

- finite alphabet Σ
- string $x[0\,..\,m-1]$ on Σ^*
- length $|x| = m$
- $\tilde{x}$ is the reverse of x ($x[m-1]x[m-2]\cdots x[1]x[0]$)
- $x[i\,..\,j]$ is a factor (substring) of x from position i to position j (both inclusive)
- $x[0\,..\,i]$ is a prefix
- $x[i\,..\,m-1]$ is a suffix
- u is a border of x if u is both a prefix and a suffix of x
- *Border*(x) is the longest border of x

Sliding window



Knuth-Morris-Pratt algorithm (1977)



$k = \min\{\ell \mid x[|Border^{\ell}(u)|] \neq a\}$ and $z = Border^{k}(u)$

Boyer-Moore algorithm (1977)

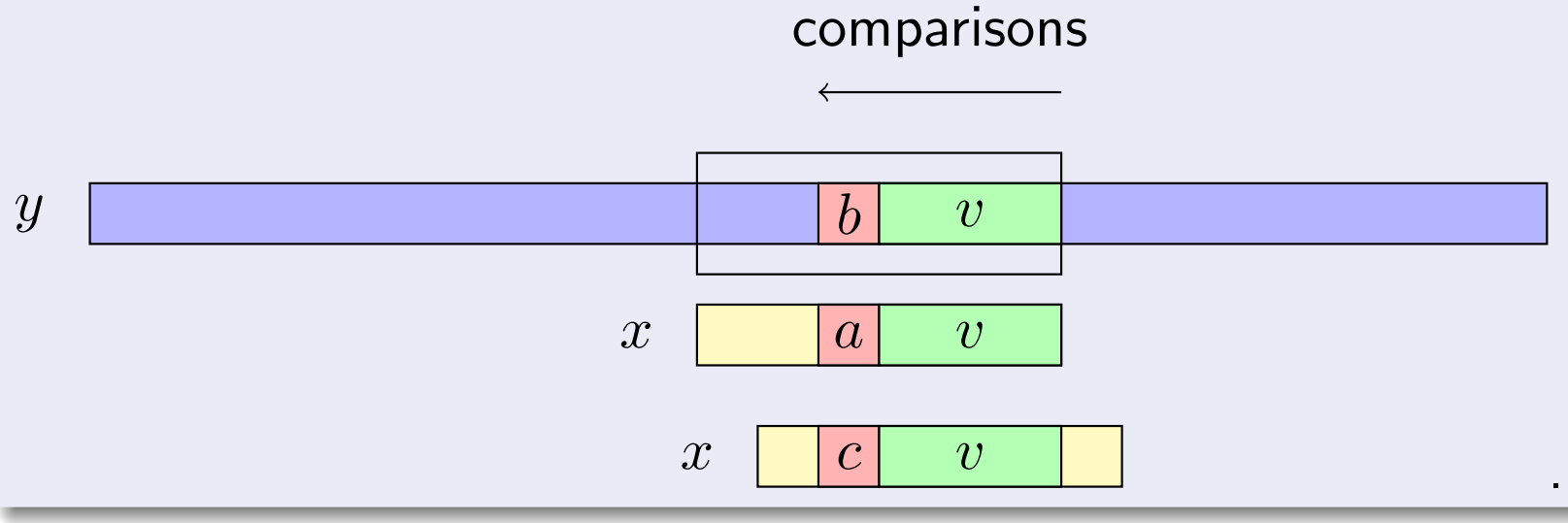


Table of contents

1 Introduction and notations

2 Search in highly similar sequences

Off-line sith an index

- Huang *et al.* 2010: $O(n + N \log N)$ bits where n is the total length of common parts in one string and N is the total length of non-common parts in all sequences
- Kuruppu *et al.* 2010: Relative Lempel-Ziv index
- Na *et al.* 2018: FM-index of an alignment
- BWBBLE, Huang *et al.* 2013: practical solution

Highly similar sequences

r sequences

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
y_0	A	T	G	C	T	A	G	C	A	A	G	A	T	A	C	A	G
y_1	A	T	G	C	T	A	G	C	A	A	C	A	T	A	C	A	G
y_2	A	T	G	C	G	A	G	C	A	A	G	A	T	A	C	A	G
y_3	A	T	G	C	T	A	G	C	A	A	C	A	T	A	C	A	T

Highly similar sequences

r sequences

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
y_0	A	T	G	C	T	A	G	C	A	A	G	A	T	A	C	A	G
y_1	A	T	G	C	T	A	G	C	A	A	C	A	T	A	C	A	G
y_2	A	T	G	C	G	A	G	C	A	A	G	A	T	A	C	A	G
y_3	A	T	G	C	T	A	G	C	A	A	C	A	T	A	C	A	T
y	A	T	G	C	{G, T}	A	G	C	A	A	{C, G}	A	T	A	C	A	{G, T}

Highly similar sequences

r sequences

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
y_0	A	T	G	C	T	A	G	C	A	A	G	A	T	A	C	A	G
y_1	A	T	G	C	T	A	G	C	A	A	C	A	T	A	C	A	G
y_2	A	T	G	C	G	A	G	C	A	A	G	A	T	A	C	A	G
y_3	A	T	G	C	T	A	G	C	A	A	C	A	T	A	C	A	T
y	A	T	G	C	{G, T}	A	G	C	A	A	{C, G}	A	T	A	C	A	{G, T}
					G	A	G	C	A	A	C						

Highly similar sequences

r sequences

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
y_0	A	T	G	C	T	A	G	C	A	A	G	A	T	A	C	A	G
y_1	A	T	G	C	T	A	G	C	A	A	C	A	T	A	C	A	G
y_2	A	T	G	C	G	A	G	C	A	A	G	A	T	A	C	A	G
y_3	A	T	G	C	T	A	G	C	A	A	C	A	T	A	C	A	T
y	A	T	G	C	{G, T}	A	G	C	A	A	{C, G}	A	T	A	C	A	{G, T}
					G	A	G	C	A	A	C						

R. Grossi, C. S. Iliopoulos, C. Liu, N. Pisanti, S. P. Pissis, A. Retha, G. Rosone, F. Vayani, L. Versari

On-Line Pattern Matching on Similar Texts

28th Combinatorial Pattern Matching (CPM), Warsaw, Poland (2017) 9:1–9:14

Highly similar sequences

r sequences

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
y_0	A	T	G	C	T	A	G	C	A	A	G	A	T	A	C	A	G
y_1	A	T	G	C	T	A	G	C	A	A	C	A	T	A	C	A	G
y_2	A	T	G	C	G	A	G	C	A	A	G	A	T	A	C	A	G
y_3	A	T	G	C	T	A	G	C	A	A	C	A	T	A	C	A	T

$$y_0 \text{ et } Z = ((\{2\}, 4, \mathtt{G}), (\{1, 3\}, 10, \mathtt{C}), (\{3\}), 16, \mathtt{T})$$

For highly similar sequences

Hamming distance

For $u, v \in A^*$ such that $|u| = |v|$:

$$\mathit{Ham}(u, v) = \sharp\{i \mid u[i] \neq v[i]\}$$

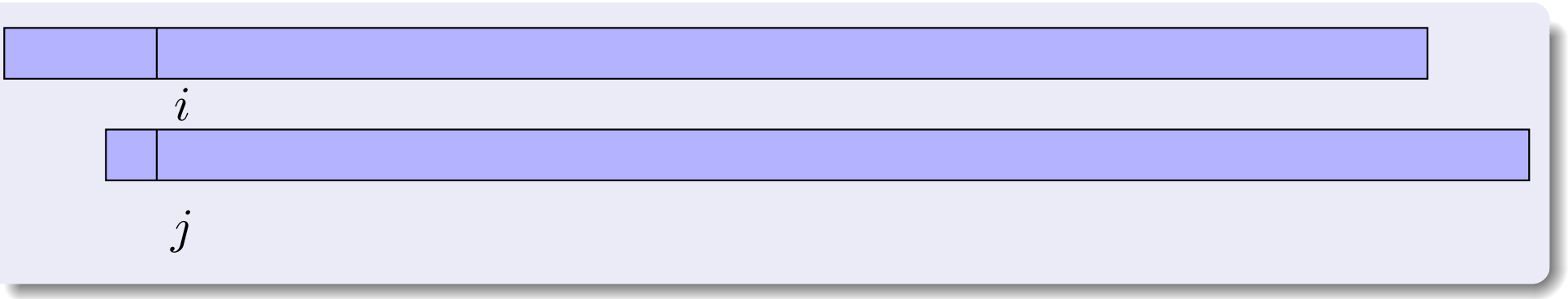
Longest Common Extension

For $x \in A^*$ and $0 \leq i \leq j \leq |x| - 1$:

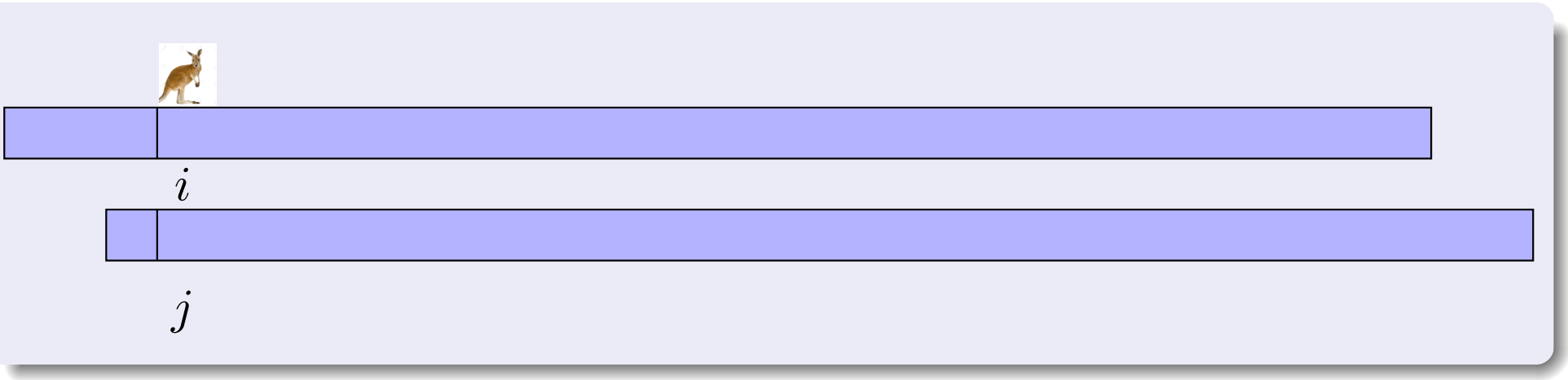
$$LCE_x^k(i, j) = \max\{\ell \mid \mathit{Ham}(x[i \,..\, i + \ell - 1], x[j \,..\, j + \ell - 1]) \leq k\}$$

Kangaroo jumps

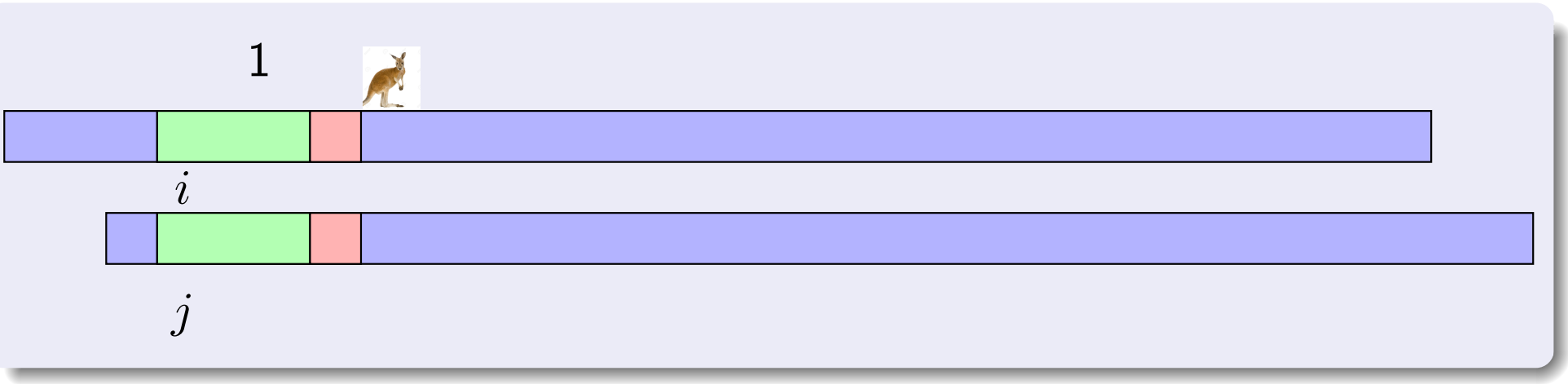
Kangaroo jumps



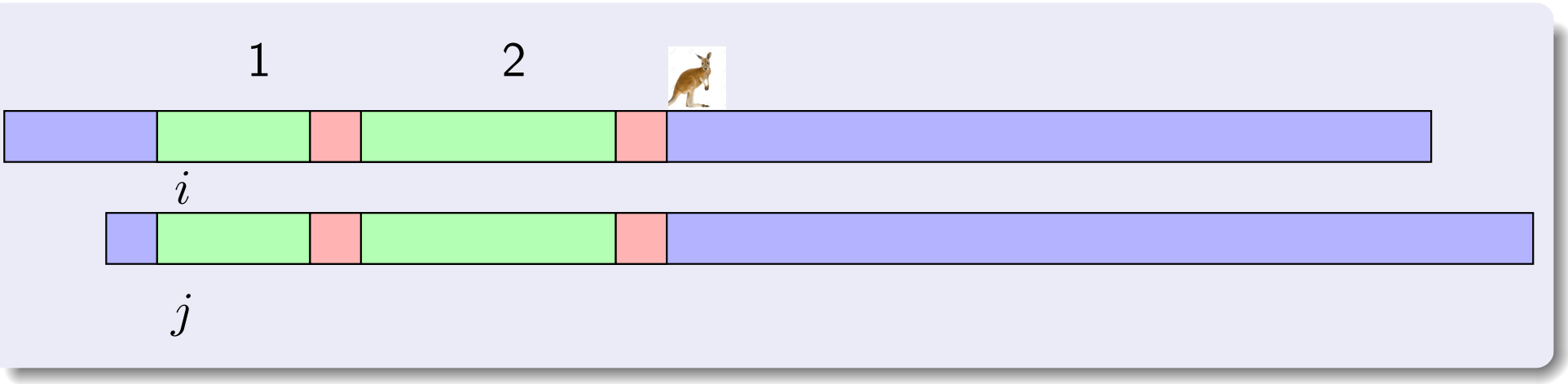
Kangaroo jumps



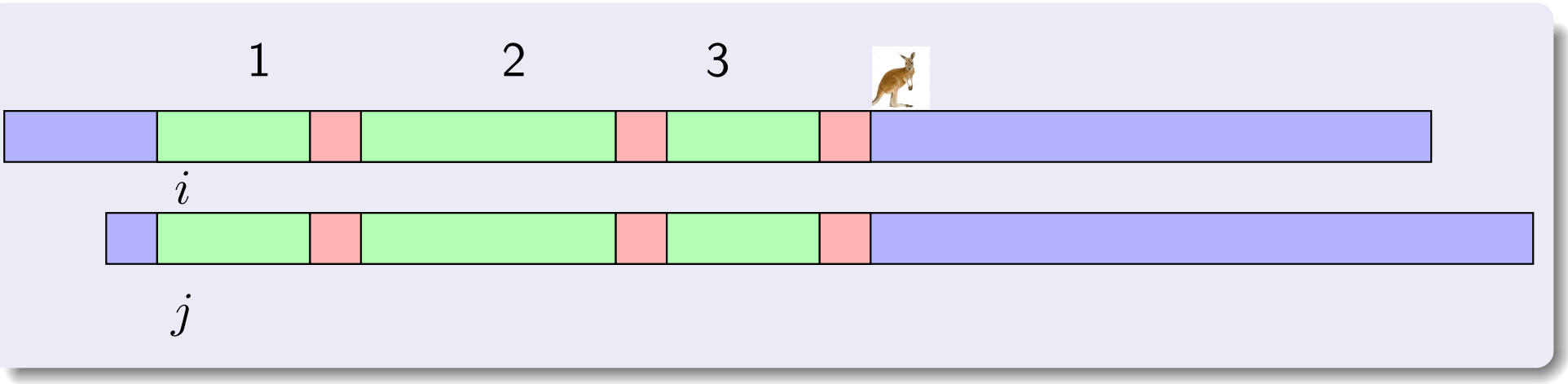
Kangaroo jumps



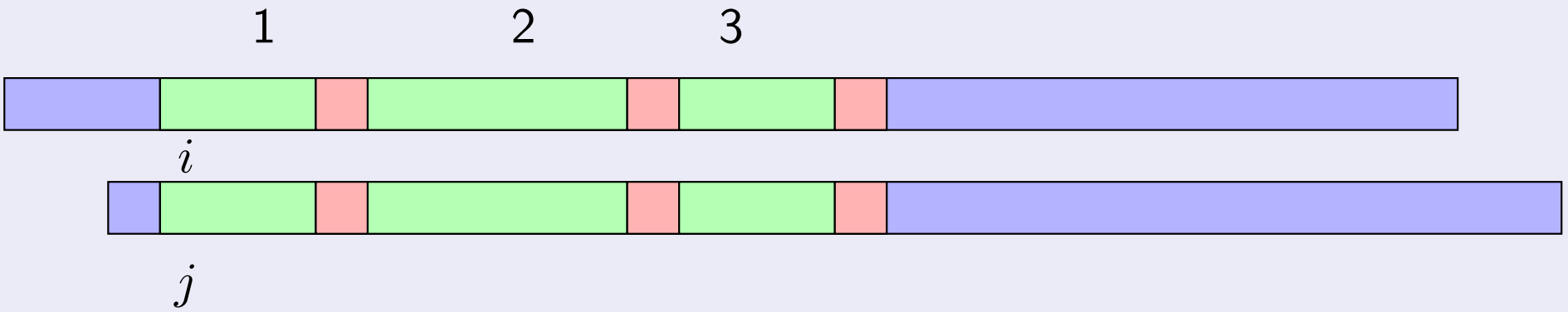
Kangaroo jumps



Kangaroo jumps



Kangaroo jumps



$LCE_x^k(i,j)$ can be computed in $O(k)$ time after $O(n)$ preprocessing time

References

Restriction: 1 variation on a window of size m

Adaptations of KMP and BM without LCE by adapting the shift functions

N. Ben Nsira, T. Lecroq and M. Elloumi

A fast Boyer-Moore type pattern matching algorithm for highly similar sequences

International Journal of Data Mining and Bioinformatics **13**(3) (2015) 266-288

N. Ben Nsira, T. Lecroq and M. Elloumi

On-line String Matching in Highly Similar DNA Sequences

Mathematics in Computer Science **11**(2) (2017) 113–126

2 variants

- relaxing the restriction from 1 to k variations in a window of size m
- searching for a finite set of patterns (still with 1 variation in a window of size m

Single pattern with at most k variations

Applying the Landau-Vishkin algorithm as a filter

Searching with k mismatches in $O(kn)$
When $\mathit{Ham}(x, y_0[j \,..\, j+\ell-1]) = \ell \leq k$

- $\ell = 0$: an exact occurrence of the pattern has been found in y_0 and all the other sequence that do not have a variation comparing to y_0 between position j and position $j + m - 1$ both included.
- $\ell > 0$: let $W = \{i_0, \ldots, i_{\ell-1}\}$ be the set of the ℓ positions such that $y_0[j + i_p] \neq x[i_p]$ with $0 \leq p < \ell$. Then x occurs exactly in y_h if:
  - $(\mathcal{G}, j + i_p, x[i_p]) \in Z$ with $g \in \mathcal{G}$ for all $0 \leq p < \ell$;
  - $\nexists\, (\mathcal{G}, h, c) \in Z$ such that $h \notin W$.

Single pattern with at most k variations

$r = 2$ and $k = 2$

	0	1	2	3	4	5	6	7	8	9	10	
y_0	A	C	C	T	A	C	G	A	C	T	A	
x			C	T	A	C	T	T				$j = 2$ and $W = (4, 5)$
x						C	T	A	C	T	T	$j = 5$ and $W = (1, 5)$
y_1	A	C	C	T	A	C	T	A	C	T	T	$Z = (((\{1\}, 6, \mathtt{T}), (\{1\}, 10, \mathtt{T}))$

Single pattern with at most k variations

$r = 2$ and $k = 2$

	0	1	2	3	4	5	6	7	8	9	10	
y_0	A	C	C	T	A	C	G	A	C	T	A	
x			C	T	A	C	T	T				$j = 2$ and $W = (4, 5)$
x						C	T	A	C	T	T	$j = 5$ and $W = (1, 5)$
y_1	A	C	C	T	A	C	T	A	C	T	T	$Z = (((\{1\}, 6, \mathtt{T}), (\{1\}, 10, \mathtt{T}))$

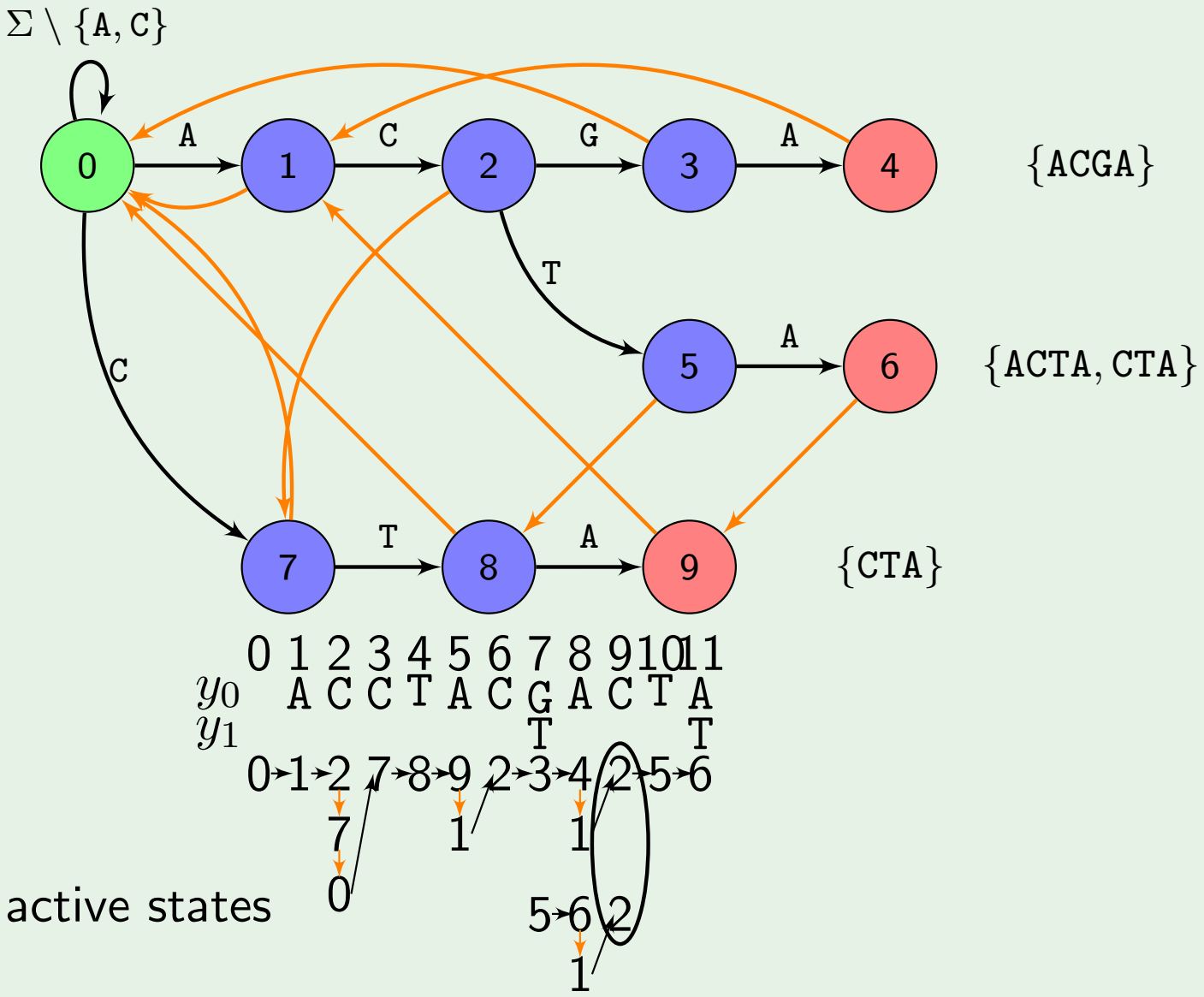
Our solution runs in time $O(knr)$

Multiple patterns with at most 1 variation

- Build a classical trie of the patterns
- Scan the highly similar sequences with at most 2 active states

Multiple patterns with at most 1 variation

$X = \{\texttt{ACGA}, \texttt{ACTA}, \texttt{CTA}\}$ and $r = 2$ sequences

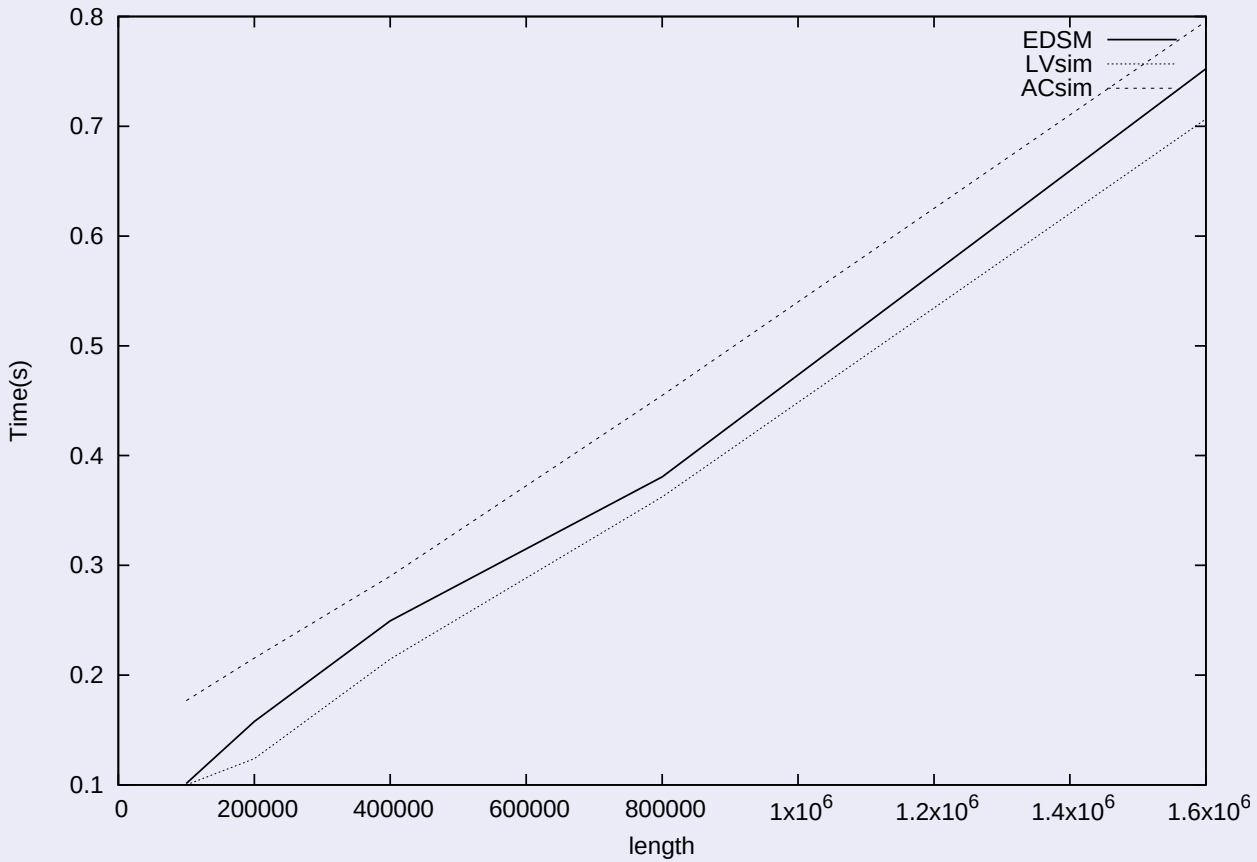


Multiple patterns with at most 1 variation

Our solution runs in time $O(n)$ for the searching phase and in time $O(s)$ for the preprocessing phase where $s = \sum |x|$ for all $x \in X$

Experiments

Similar sequences of different lengths with patterns of length 16



Perspectives

- Do more experiments
- Adapt other pattern matching techniques
- Relax the restrictions
- Adaptive analysis

Thank you for your attention!