Recuperació de la informació

- 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**

Exact string matching: one pattern

How does the string algorithms made the search?

For instance, given the sequence

CTACTACTACGTCTATACTGATCGTAGCTACTACATGC

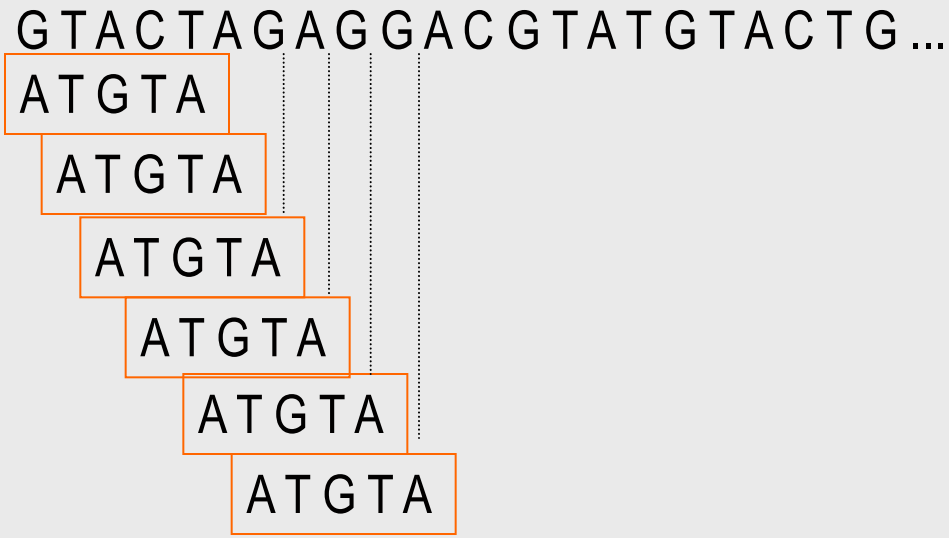
search for the pattern ACTGA.

and for the pattern TACTACGGTATGACTAA

Exact string matching: Brute force algorithm

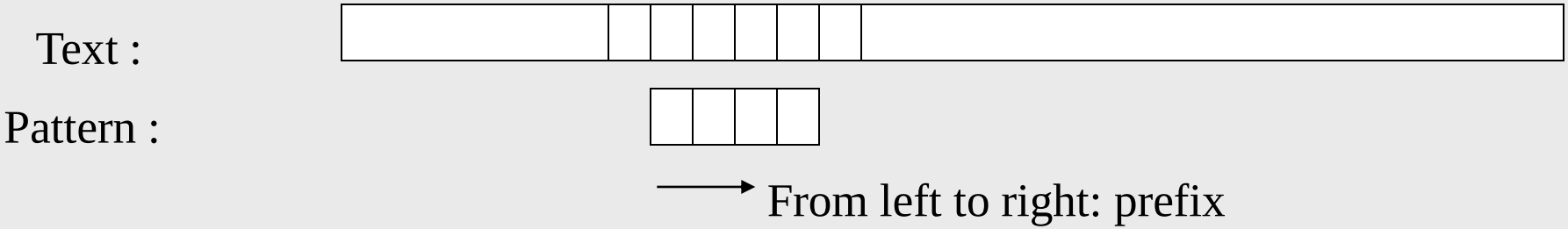
Example:

Given the pattern ATGTA, the search is

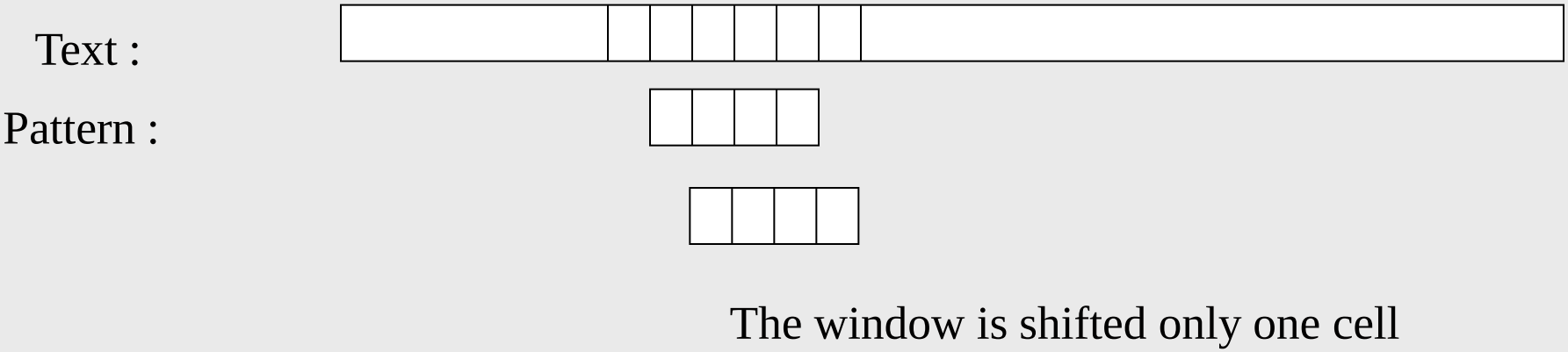


Exact string matching: Brute force algorithm

- How the comparison is made?



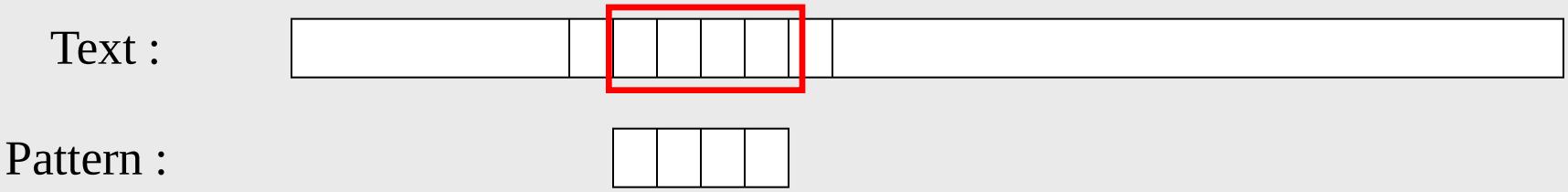
- Which is the next position of the window?



Exact string matching: one pattern

How does the matching algorithms made the search?

There is a sliding window along the text against which the pattern is compared:



At each step the comparison is made and the window is shifted to the right.

Which are the facts that differentiate the algorithms?

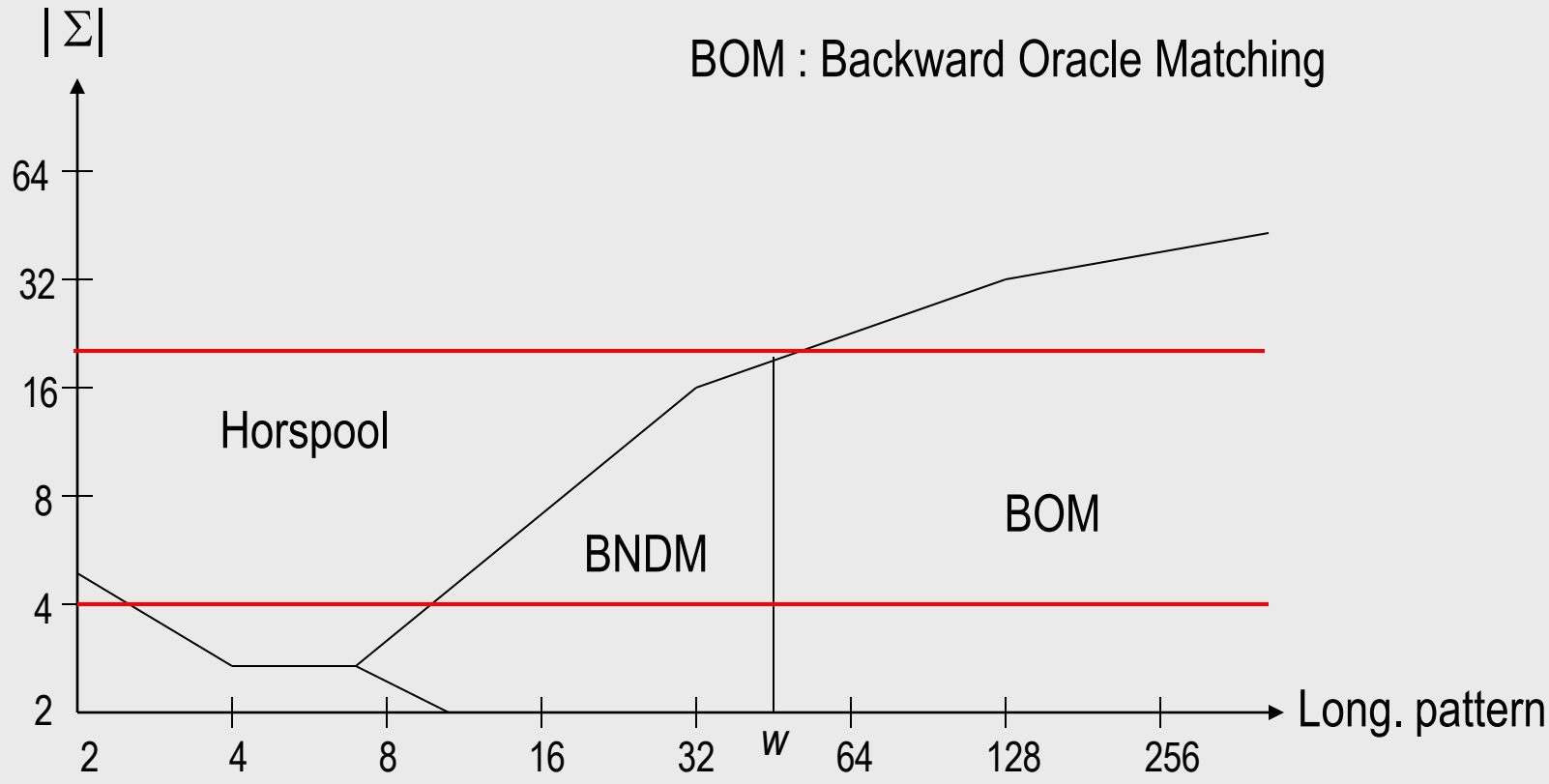
1. How the comparison is made.
2. The length of the shift.

Exact string matching: one pattern (text on-line)

Experimental efficiency (Navarro & Raffinot)

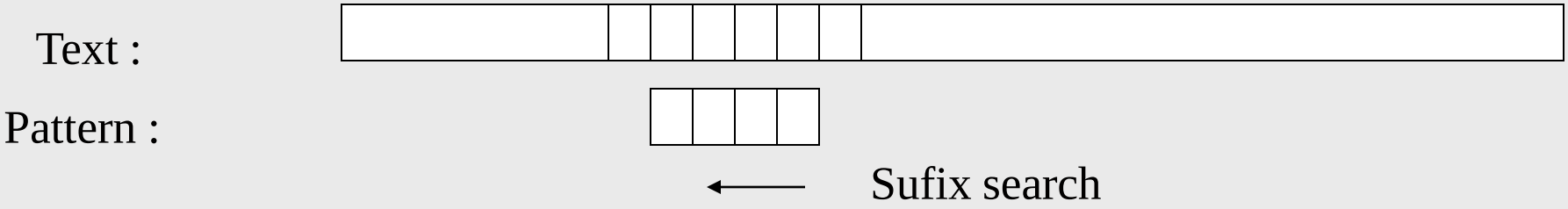
BNDM : Backward Nondeterministic Dawg Matching

BOM : Backward Oracle Matching

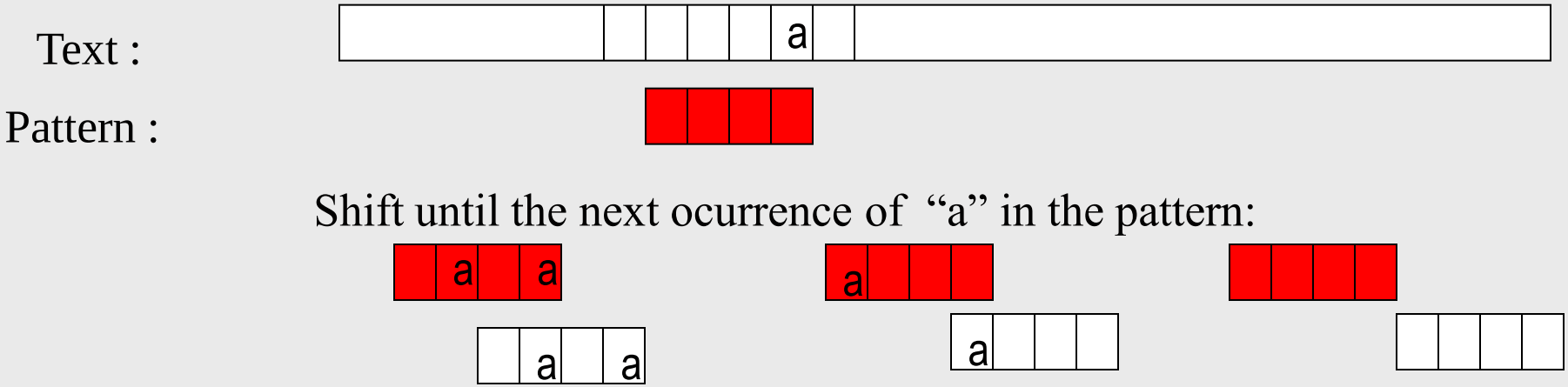


Horspool algorithm

- **How the comparison is made?**



- **Which is the next position of the window?**



We need a preprocessing phase to construct the shift table.

Horspool algorithm : example

Given the pattern ATGTA

- The shift table is:

A	
C	
G	
T	

Horspool algorithm : example

Given the pattern ATGTA

- The shift table is:

A	4
C	
G	
T	

Horspool algorithm : example

Given the pattern ATGTA

- The shift table is:

A	4
C	5
G	
T	

Horspool algorithm : example

Given the pattern ATGTA

- The shift table is:

A	4
C	5
G	2
T	

Horspool algorithm : example

Given the pattern ATGTA

- The shift table is:

A	4
C	5
G	2
T	1

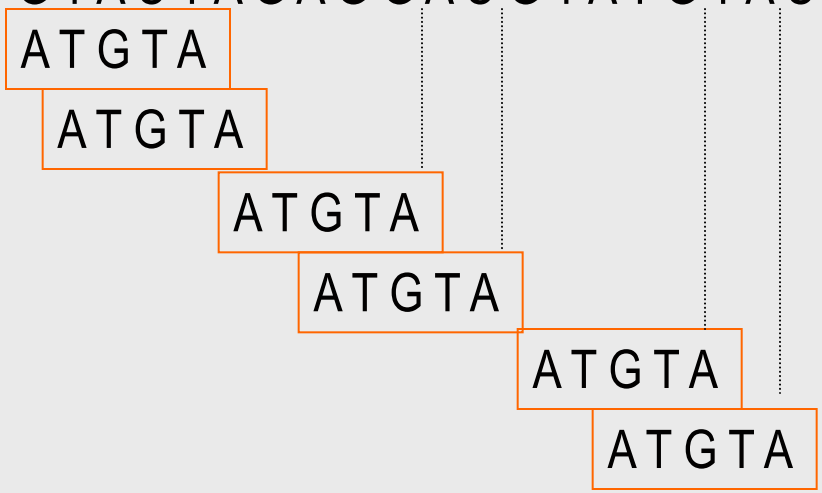
Horspool algorithm : example

Given the pattern ATGTA

- The shift table is:

A	4
C	5
G	2
T	1

- The searching phase: G T A C T A G A G G A C G T A T G T A C T G ...



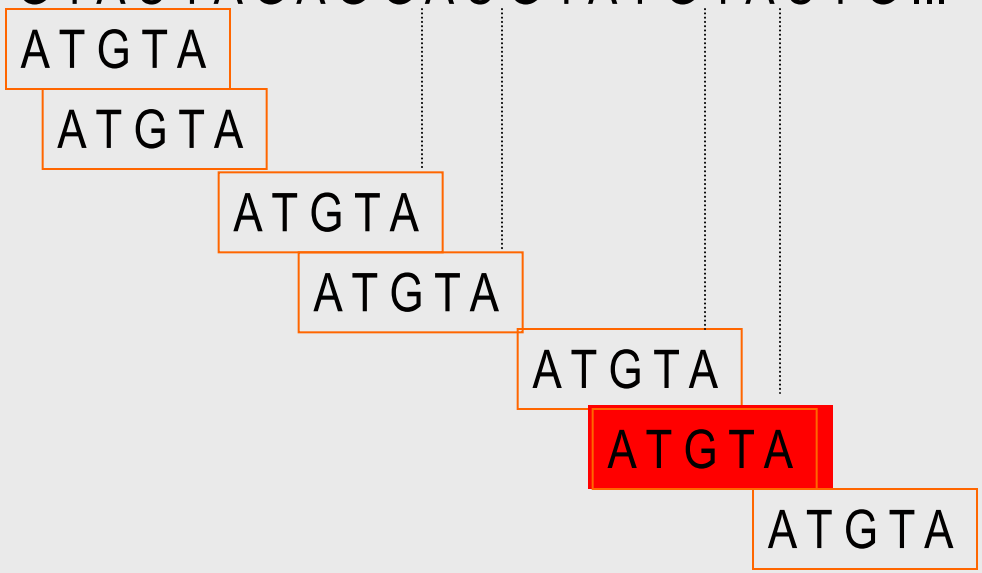
Exemple algorisme de Horspool

Given the pattern ATGTA

- The shift table is:

A	4
C	5
G	2
T	1

- The searching phase: G T A C T A G A G G A C G T A T G T A C T G ...



Qüestions sobre l'algorisme de Horspool

Given the pattern ATGTA, the shift table is

A	4
C	5
G	2
T	1

Given a random text over an equally likely probability distribution (EPD):

1.- Determine the expected shift of the window. And, if the PD is not equally likely?

2.- Determine the expected number of shifts assuming a text of length n.

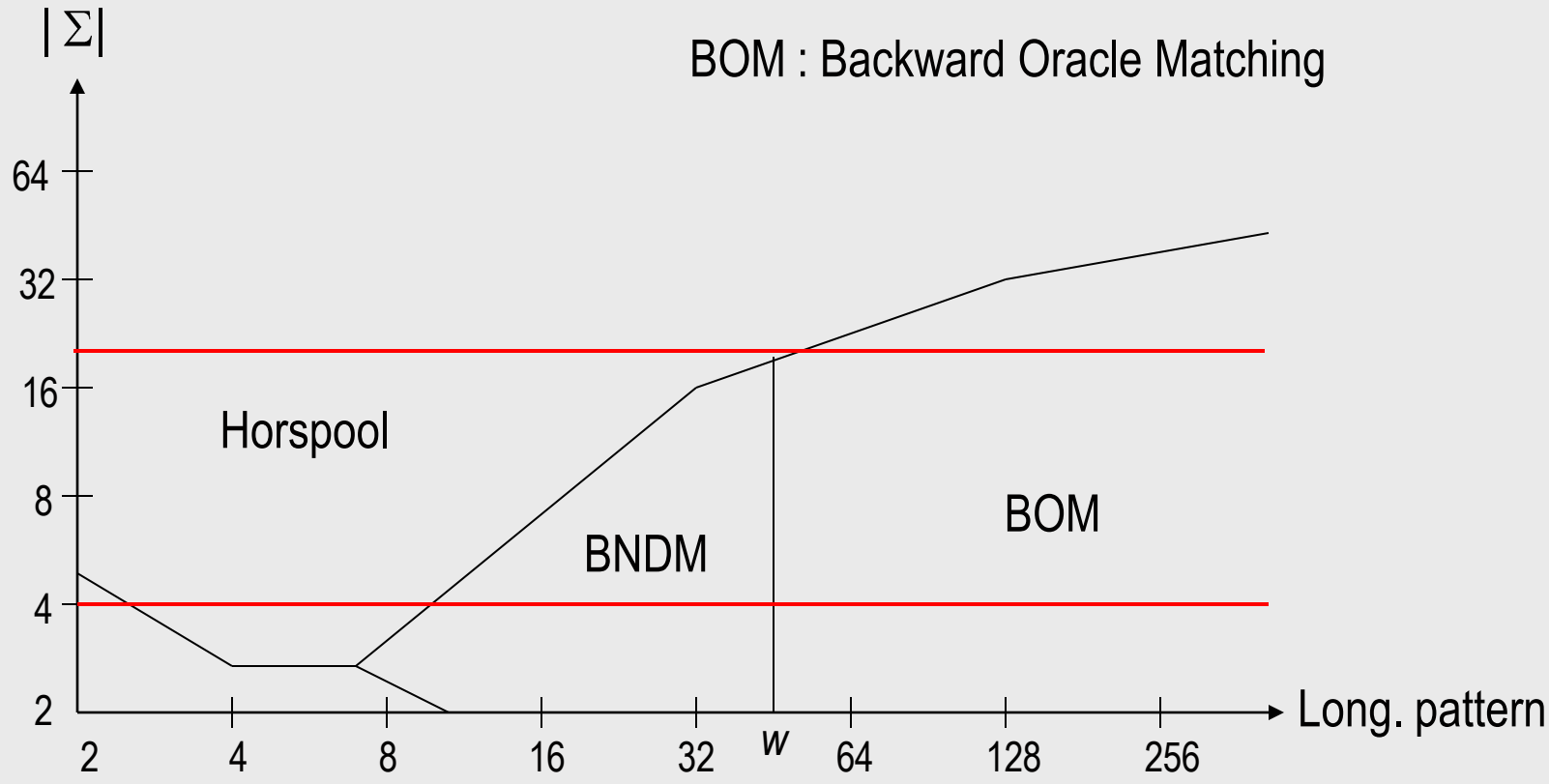
3.- Determine the expected number of comparisons in the suffix search phase

Exact string matching: one pattern (text on-line)

Experimental efficiency (Navarro & Raffinot)

BNDM : Backward Nondeterministic Dawg Matching

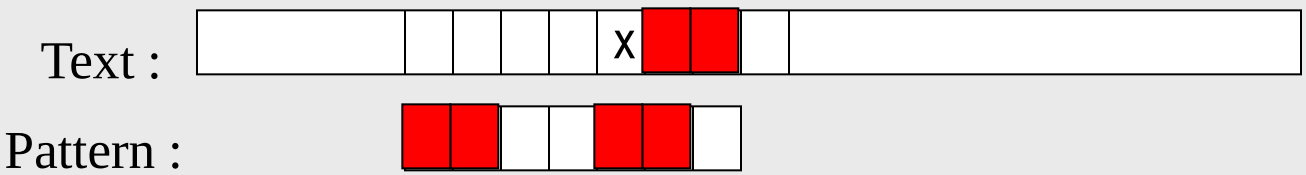
BOM : Backward Oracle Matching



BNDM algorithm

- **How the comparison is made?**

Search for suffixes of T that are factors of



That is denoted as

$D_2 = 1\ 0\ 0\ 0\ 1\ 0\ 0$

Once the next character x is read

$D_3 = D_2 << 1\ \&\ B(x)$

B(x): mask of x in the pattern P.
For instance, if B(x) = (0 0 1 1 0 0 0)

D = (0 0 0 1 0 0 0) & (0 0 1 1 0 0 0) = (0 0 0 1 0 0 0)

- **Which is the next position of the window ?**

Depends on the value of the leftmost bit of D

BNDM algorithm: exaple

Given the pattern ATGTA

- The mask of characters is:

B(A) = (1 0 0 0 1)
B(C) = (0 0 0 0 0)
B(G) = (0 0 1 0 0)
B(T) = (0 1 0 1 0)

- The searching phase: G T A C T A G A G G A C G T A T G T A C T G ...

A T G T A

D_1 = (0 1 0 1 0)
D_2 = (1 0 1 0 0) & (0 0 0 0 0) = (0 0 0 0 0)

A T G T A

D_1 = (0 0 1 0 0)
D_2 = (0 1 0 0 0) & (0 0 1 0 0) = (0 0 0 0 0)

A T G T A

D_1 = (1 0 0 0 1)
D_2 = (0 0 0 1 0) & (0 1 0 1 0) = (0 0 0 1 0)
D_3 = (0 0 1 0 0) & (0 0 1 0 0) = (0 0 1 0 0)
D_4 = (0 1 0 0 0) & (0 0 0 0 0) = (0 0 0 0 0)

A T G T A

Exemple algorisme BNDM

- Given the pattern ATGTA
- The mask of characters is :

B(A) = (1 0 0 0 1)
B(C) = (0 0 0 0 0)
B(G) = (0 0 1 0 0)
B(T) = (0 1 0 1 0)

- The searching phase: G T A C T A G A G G A C G T A T G T A C T G ...

A T G T A

D_1 = (1 0 0 0 1)
D_2 = (0 0 0 1 0) & (0 1 0 1 0) = (0 0 0 1 0)
D_3 = (0 0 1 0 0) & (0 0 1 0 0) = (0 0 1 0 0)
D_4 = (0 1 0 0 0) & (0 1 0 1 0) = (0 1 0 0 0)
D_5 = (1 0 0 0 0) & (1 0 0 0 1) = (1 0 0 0 0)
D_6 = (0 0 0 0 0) & (* * * * *) = (0 0 0 0 0) Trobat!

A T G T A

Exemple algorisme BNDM

Given the pattern ATGTA

- The mask of characters is :

B(A) = (1 0 0 0 1)
B(C) = (0 0 0 0 0)
B(G) = (0 0 1 0 0)
B(T) = (0 1 0 1 0)

How the shif is determined?

- The searching phase: G T A C T A G A A T A C G T A T G T A C T G ...

A T G T A

D_1 = (0 1 0 1 0)
D_2 = (1 0 1 0 0) & (0 0 0 0 0) = (0 0 0 0 0)

A T G T A

D_1 = (0 1 0 1 0)
D_2 = (1 0 1 0 0) & (1 0 0 0 1) = (1 0 0 0 0)
D_3 = (0 0 0 0 0) & (1 0 0 0 1) = (0 0 0 0 0)

A T G T A