

# **Différentes applications de l'oracle des facteurs d'un mot**

# Plan

- " Introduction et notations
- " Oracle des facteurs
- " Répétitions
- " Compression (sans perte)

# Notations

- $A$  : **alphabet** (ensemble de **lettres**)
- $\square$  : **concaténation**
- $A^*$  : ensemble des suites finies de lettres
- $w \in A^*$  : **mot**
- $|w|$  : **longueur** du mot  $w$
- $w = w[0] \square w[1] \square \dots \square w[|w|-1] = w[0..|w|-1]$
- $i$  ( $0 \leq i \leq |w|-1$ ) : **position** sur le mot  $w$
- $\varepsilon$  : **mot vide** ( $|\varepsilon| = 0$ )

# Notations

"  $u$  est un préfixe  
"  $v$  est un suffixe  
"  $z$  est un facteur

} de  $w$  si  $w = uzv$

"  $Fact(w)$  : ensemble des facteurs de  $w$

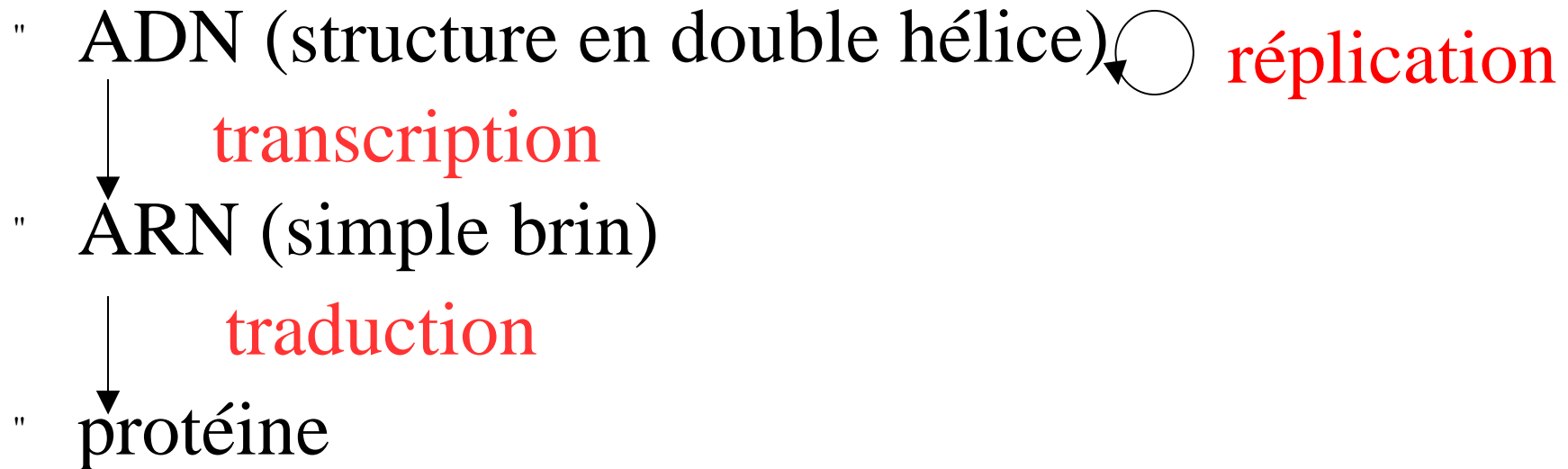
"  $Alph(w)$  : ensemble des lettres de  $w$

# Séquences biologiques

- " ADN :  $A = \{a, c, g, t\}$  (acides nucléiques, paires de base)
- " ARN :  $A = \{a, c, g, u\}$
- " protéine :  $\text{card } A = 20$  (acides aminés)

génom : ensemble de chromosomes

# Dogme central de la biologie moléculaire



# Répétitions

- " **M. Crochemore**, An optimal algorithm for computing the repetitions in a word, *Information Processing Letters* 12 (5) (1983) 244 - 250.
- " **R. Kolpakov & G. Kucherov**, On maximal repetitions in words, *Journal of Discrete Algorithms* 1 (1) (2000) 159 - 186.
- " **W. F. Smyth**, Repetitive perhaps, but not boring, *Theoretical Computer Science* 249 (5) (2000) 345 - 355.

# Structures d'index

- " Arbre des suffixes
- " Arbre compact des suffixes
- " Table des suffixes
- " Automate des suffixes (DAWG)
- " Automate compact des suffixes (CDAWG)

# Oracle des facteurs

Allauzen, Crochemore, Raffinot, 1999

- " Pour un mot  $w$  de longueur  $m$  l'oracle des facteurs de  $w$  est un automate  $(Q, q_0, F, \delta)$  où :
- "  $Q = \{0, 1, \dots, m\}$  est l'ensemble des états
- "  $q_0 = 0$  est l'état initial
- "  $F = Q$  est l'ensemble des états terminaux
- $\delta$  est la fonction de transition

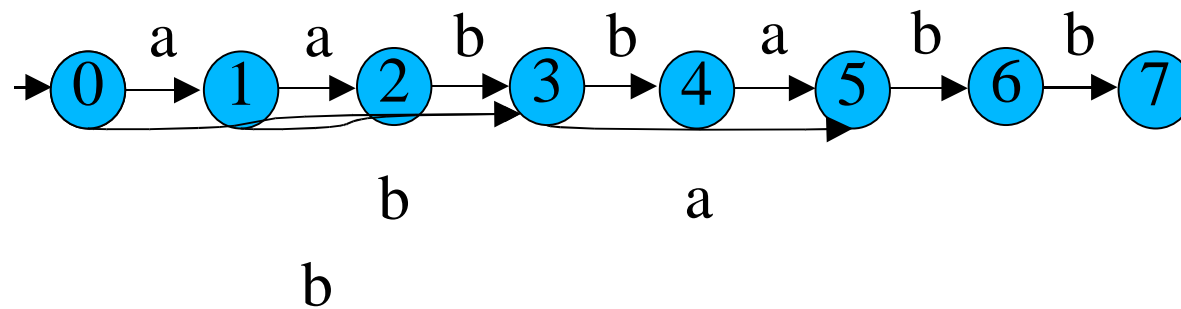
# Oracle des facteurs

Allauzen, Crochemore, Raffinot, 1999

- " Pour un mot  $w$  de longueur  $m$  l'oracle des facteurs de  $w$  :
- " reconnaît au moins tous les facteurs de  $w$
- " possède exactement  $m + 1$  états
- " possède entre  $m$  et  $2m - 1$  transitions

# Exemple d'oracle des facteurs

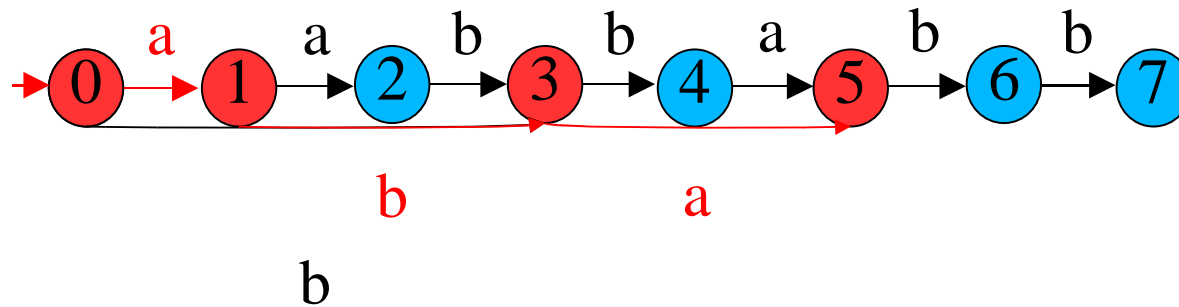
$w = aabbabb$



$Fact(w) = \{ \varepsilon, a, b, aa, ab, ba, bb, aab, abb, bab, bba, aabb, abba, babb, bbab, aabba, abbab, bbabb, aabbab, abbabb, aabbabb \}$

# Langage de l'oracle

$w = aabbabb$



**aba** est reconnu bien qu'il ne soit pas un facteur de aabbabb

# Oracle des facteurs d'un mot $w$

bijection entre :

- les longueurs des préfixes de  $w$
- les états de l'oracle des facteurs de  $w$

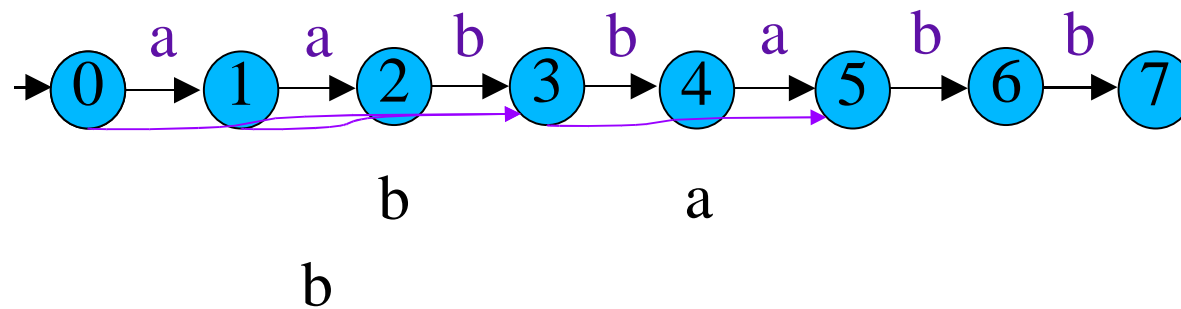
toutes les transitions qui mènent à l'état  $i$  sont étiquetées par la lettre  $w[i-1]$

deux types de transitions :

- transitions internes :  $\delta(i, w[i]) = i+1$  pour  $0 \leq i \leq m-1$
- transitions externes :  $\delta(i, w[j-1]) = j$  avec  $j-i > 1$  pour  $0 \leq i < j \leq m$

# Représentation de l'oracle des facteurs

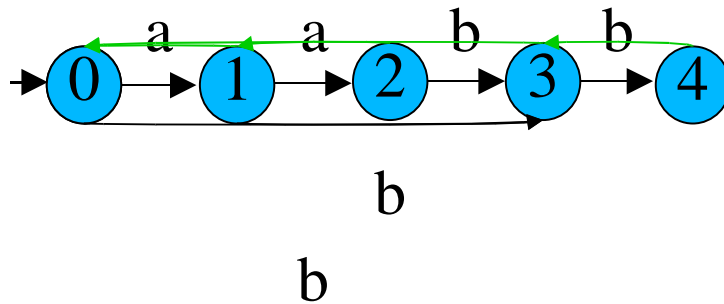
$w = aabbabb$



$aabbabb, (0,3), (1,3), (3,5)$   
représentation indépendante de l'alphabet

# Construction de l'oracle des facteurs

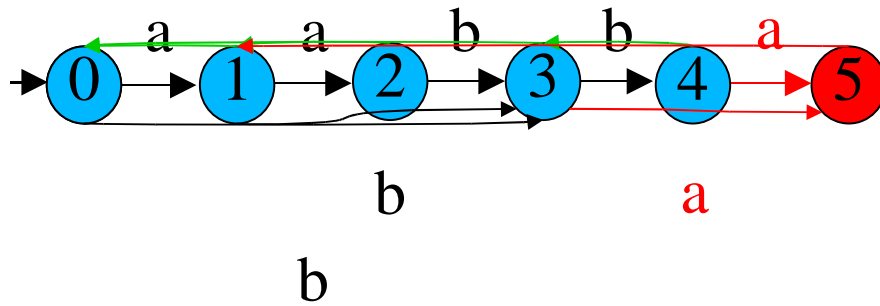
$w = aabb$



Lien suffixe de l'état  $i$  :  $suf(i) = \delta(0, x)$  où  $x$  est le plus long suffixe répété de  $w[0..i-1]$

# Construction de l'oracle des facteurs

$w = aabba$



# Construction de l'oracle des facteurs

Théorème [ACR 99] :

L'oracle des facteurs d'un mot de longueur  $m$  peut être construit en temps et en espace  $O(m)$ .

# Recherche exacte de mot

Trouver toutes les occurrences d'un mot  $x$  de longueur  $m$  dans un texte  $y$  de longueur  $n$  (avec  $x, y \in A^*$  et  $m \ll n$ )

Deux types de problèmes

- " le mot est fixe
  - prétraitement en temps  $O(m)$
  - recherche en temps  $O(n)$
- " le texte est fixe
  - prétraitement en temps  $O(n)$
  - recherche en temps  $O(m)$

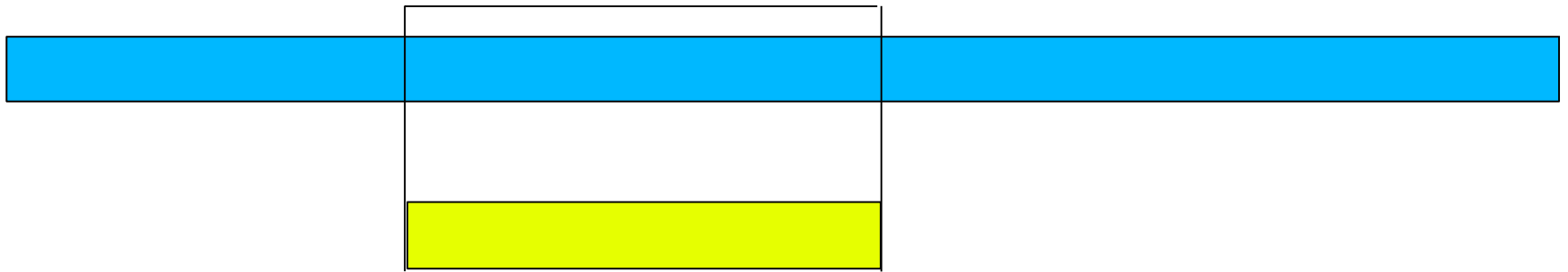
# Recherche exacte de mot

Mécanisme de fenêtre glissante



# Recherche exacte de mot

Mécanisme de fenêtre glissante



# Recherche exacte de mot

Mécanisme de fenêtre glissante



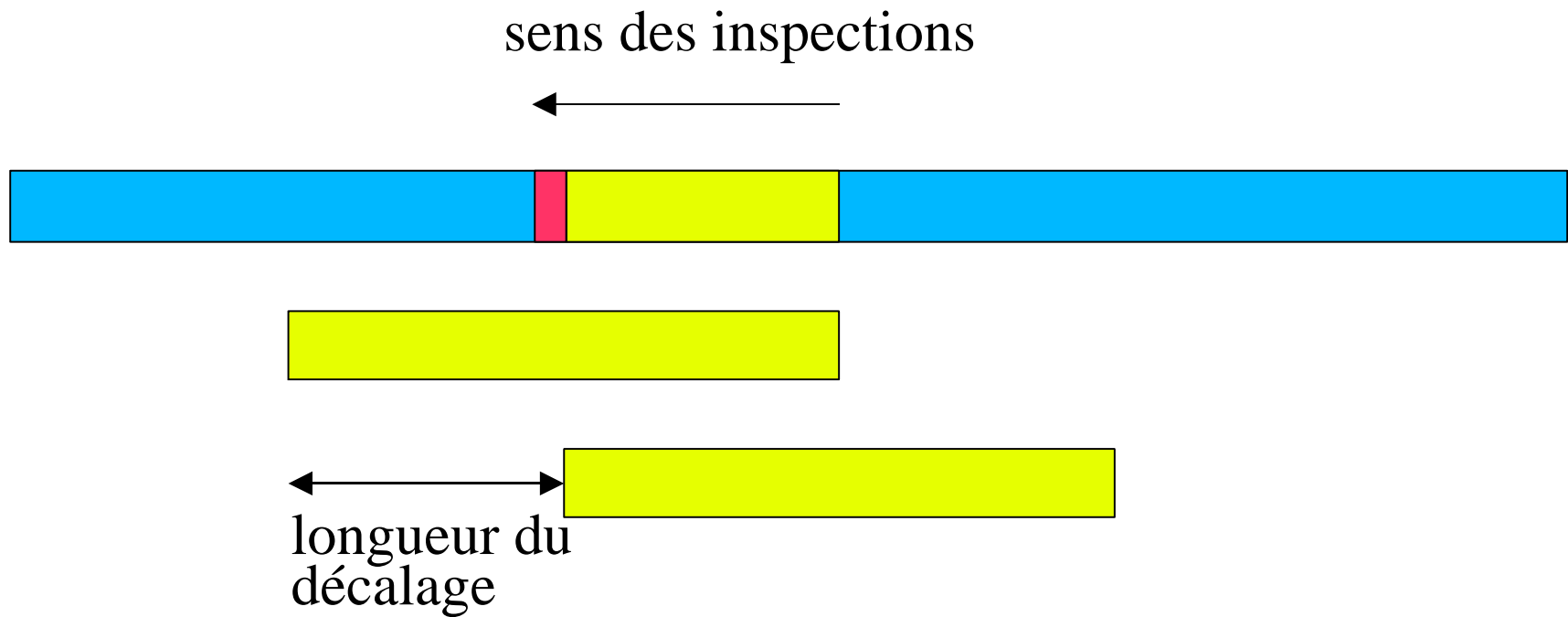
# Recherche exacte de mot

Sucession de

" tentatives

" décalages

# Recherche exacte de mot



# Recherche exacte de mot

L'oracle des facteurs d'un mot  $w$  de longueur  $m$  ne reconnaît qu'un seul mot de longueur  $m$

# Calcul de répétitions avec l'oracle des facteurs

Pour un mot  $w$  de longueur  $m$  on définit :

$$LRS[i] = \max \{ |v| \mid v \text{ est suffixe de } w[0..i] \text{ et } v \text{ est facteur de } w[0..i-1] \}$$

pour  $0 \leq i < m$

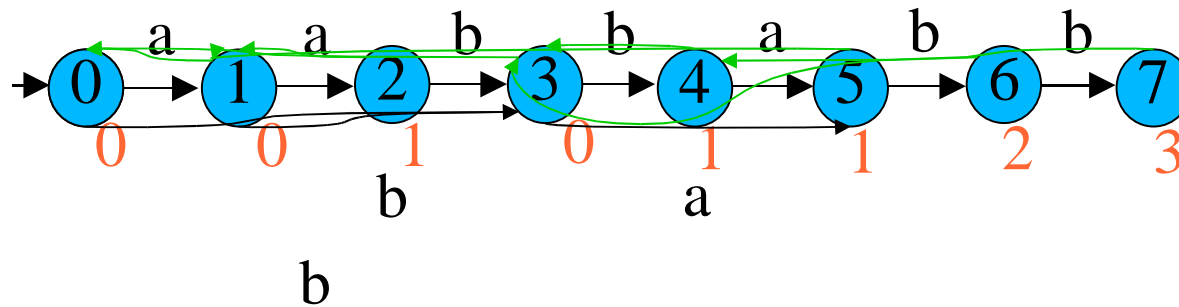
**Théorème [LL2000]** : il est possible de calculer en temps et espace  $O(m)$  pour  $0 \leq i < m$  :

$$0 \leq lrs[i] \leq LRS[i]$$

En pratique  $lrs[i]$  est très proche de  $LRS[i]$ .

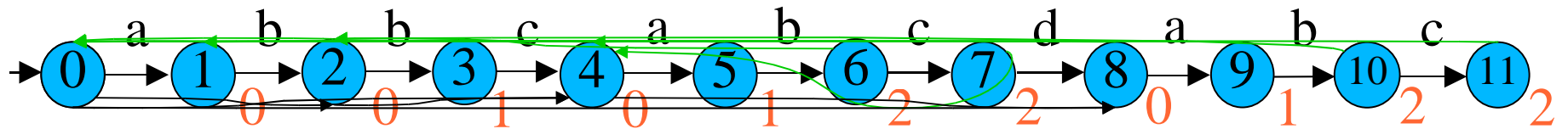
# Calcul de répétitions avec l'oracle des facteurs

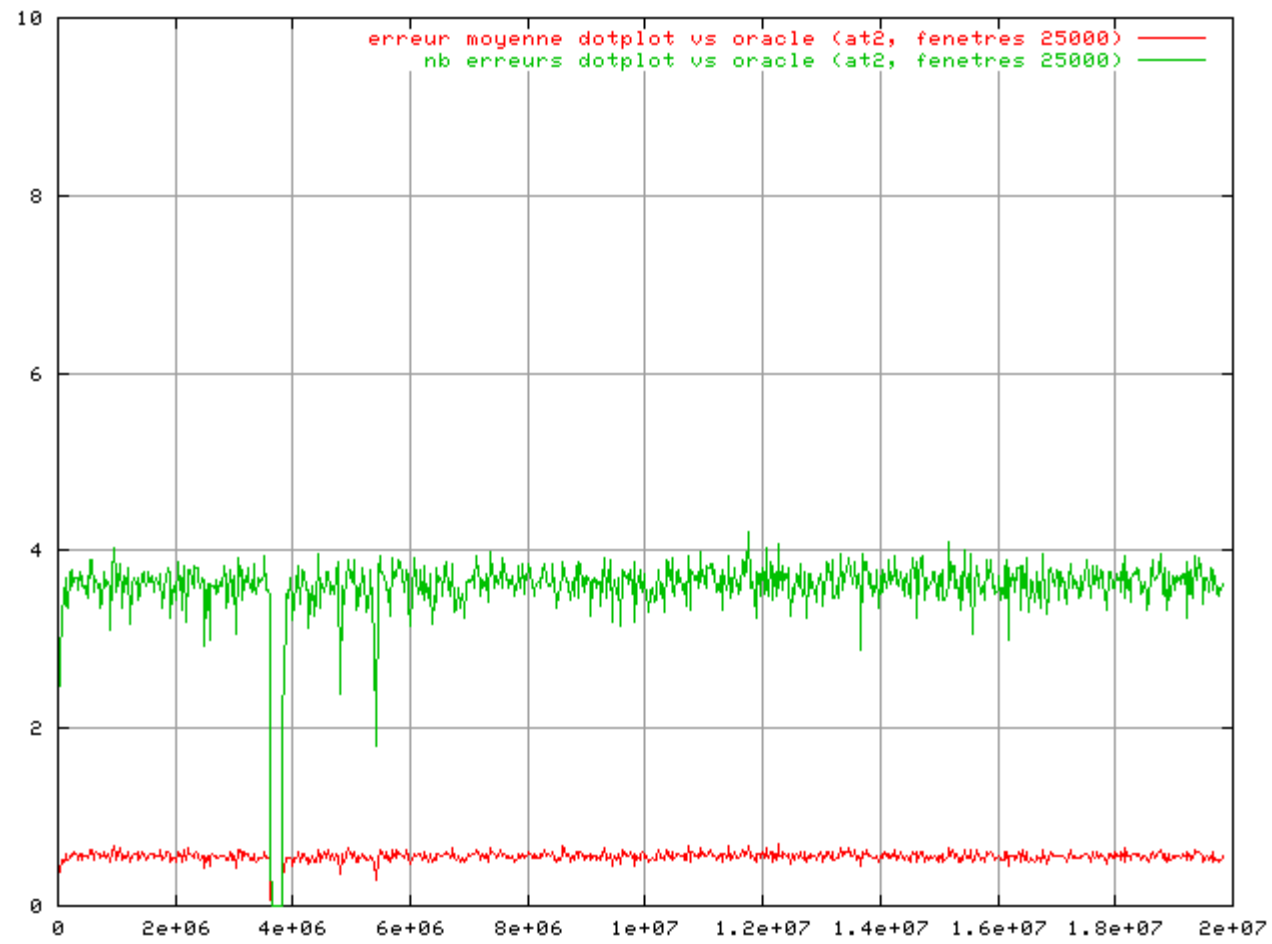
$w = \text{aabbabb}$

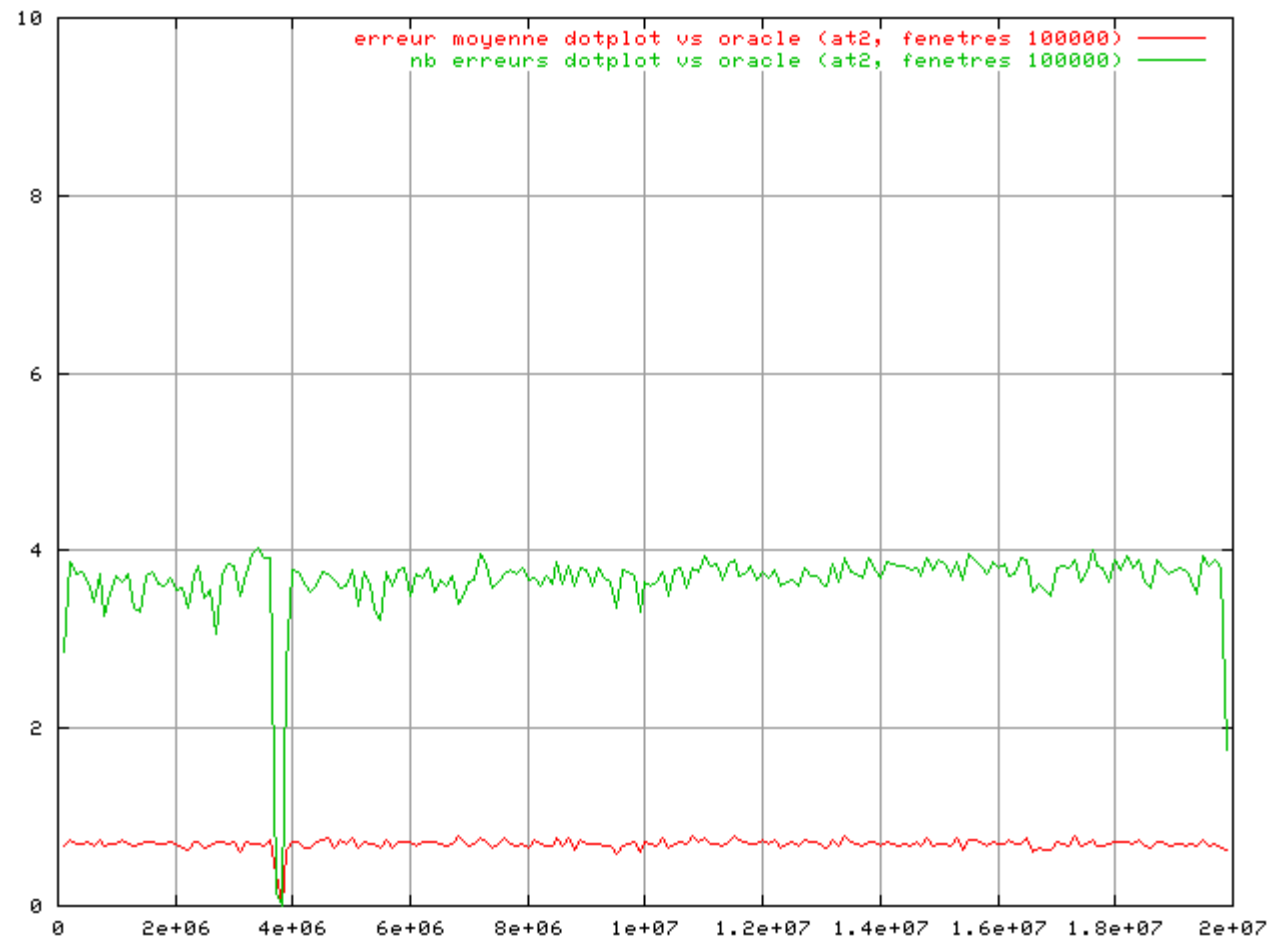


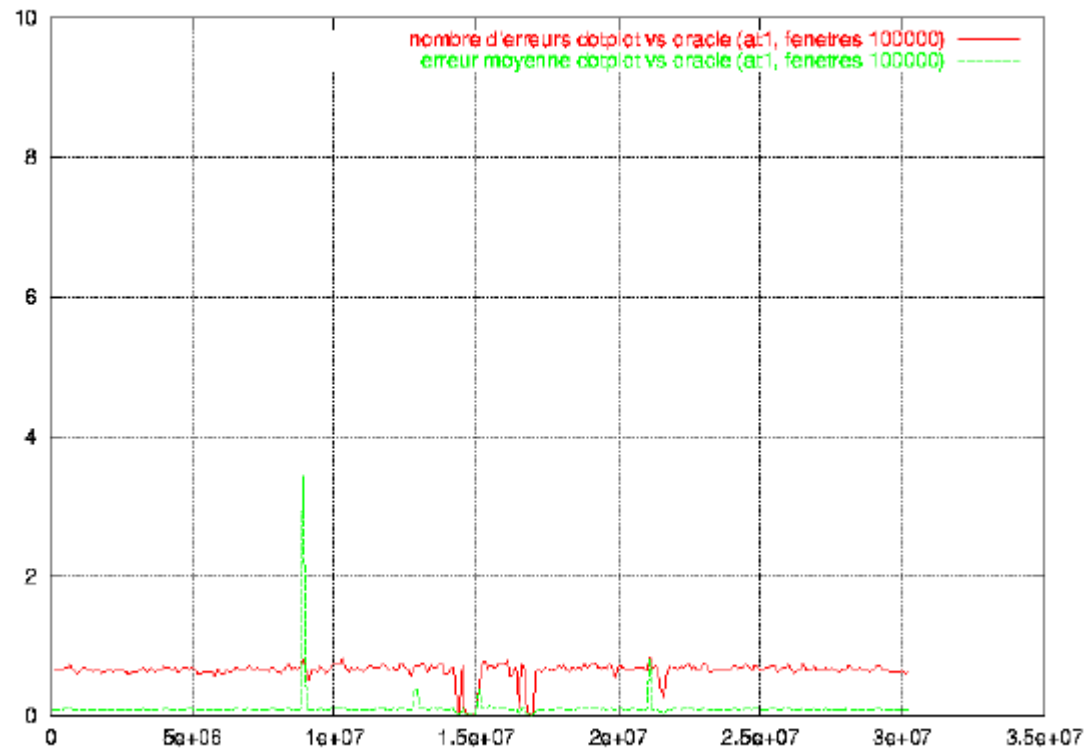
# Calcul de répétitions avec l'oracle des facteurs

∞  $w = \text{abbcabcdabc}$









## FORRepeats

Choisissez les répétitions à visualiser

Arabidopsis thaliana IV

Longueurs min et max des répétitions 200

Coordonnées min et max sur le premier chromosome

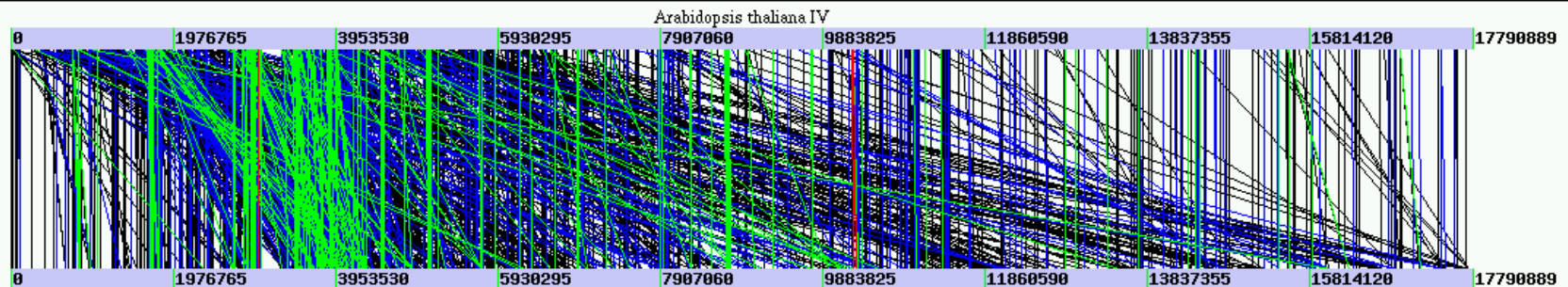
Coordonnées min et max sur le second chromosome

Distances min et max entre les répétitions

Zoom Map

jaune 100 noir 500 bleu 1000 vert 10000 rouge

Envoyer reset



[Visualiser et sauvegarder la fiche](#)

Organisme at4

Coordonnées sur la première séquence 10398282

Coordonnées sur la seconde séquence 13596263

Longueur de la répétition 274

```
> at4 10398282 274
aattgttaaaaaactgaagtttaacccctgttaaacagc atattaataatattttaaa
ttttataaatattgattcaataacatccacatataaccccaattccaaaataaaacccgt
tcgttaattcccttaaccgctccgtaataatttttaaaagtaataatttttaaaatttaa
gaattttattatataaaaagttttgcaaatgtatcattototaacacatttatatt
ttatagtttaattttatataatagtaacattatac
```

```
> at4 13596263 274
attgttaaaaaactgaagtttaaccttttataaaaacagc atactaataatattttaaa
ttttataaatattgattcaataacatccacatataaaccaattccaaaataaaacttt
tctttaattcttttaaccgctccgtaataatttttaaaagtaataatttttaaaatttaa
aaattttattatataaaaagttttgcaaatgtatcattototaacacatttatatt
ttatagtttaattttatataatagtaacattatac
```

# FORRepeats

Choisissez les répétitions à visualiser

Arabidopsis thaliana IV

Longueurs min et max des répétitions 2000

Coordonnées min et max sur le premier chromosome

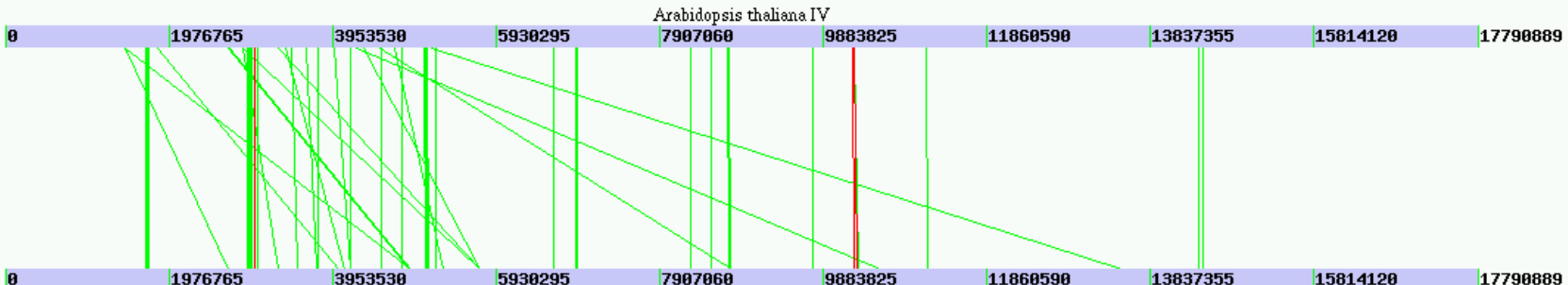
Coordonnées min et max sur le second chromosome

Distances min et max entre les répétitions

Zoom Map

jaune 100 noir 500 bleu 1000 vert 10000 rouge

Envoyer reset



[Visualiser et sauvegarder la fiche](#)

Organisme at4

Coordonnées sur la première séquence 10246374 Coordonnées sur la seconde séquence 10249633

Longueur de la répétition 50061

```
> at4 10246374 50061
aaggctgacattogagocgtgctggtcgcocactggagcaacttgagaggtttgctcoga
aaaacaaagctgacatggagacatgttcaggcttccatocccaaagctgatgacgttcaa
agcattgacttctcggagcgccttttggttaacccctaaagctgctgagttcggatgc
cgc aatcttcggcctctctgcacagaaatacctactatctctcagcacccttgccctg
attctcagcaccctcttgccocagcaaggtctctctgctcctttgagataagcctttac
ctataagacagaccttggtcttctttgtagtaagggtgaccttgaagaatctgacatt
ccgagctgcttgacctccgcctccctgctctcttcgagcaaccacgcgacttgaag
atctgctgactctttctgaagagacgtctctcagcattggagctgcttgacctctgctt
caacgagctcctctcgccttgctggtctccggagatgcctccgcctcacttaacatgattcc
```

```
> at4 10249633 50061
gaagatctcactcttgctgtagaagagaaaagcaaatgatgaaagctagtgatataagoga
aaatgtaactcaagcttaacaaatagtaagaaagtaattgaaattatctttatgtggg
aggtctctgagacggagacagaattggagatcgcacattgatgaaatgagtgaaagcctt
gagcattgctagtgatgacattgctgtaactctctttgctgatttgtagaagaact
gataactccgcttgaaacgttaaaattattgacccacccgtgaaactcttagctgagtg
cgcacgaagatttgctcgaatttttagtagatgggatattgctgaggtctcaccgctc
ttatgaagttgctgtagcagttttttctctcttgctgattcggagatgcattaat
aaccttattctgaacgcttgattctcagatccacccgttgaactattatctagtagctc
aagaactcctgtaatttatagcttgatggagctgtagtattctccatgactctctg
```

## FORRepeats

Choisissez les répétitions à visualiser

Arabidopsis thaliana IV

Longueurs min et max des répétitions

Coordonnées min et max sur le premier chromosome

1442000

1448000

Coordonnées min et max sur le second chromosome

1442000

1448000

Distances min et max entre les répétitions

Zoom Map

jaune

100

noir

500

bleu

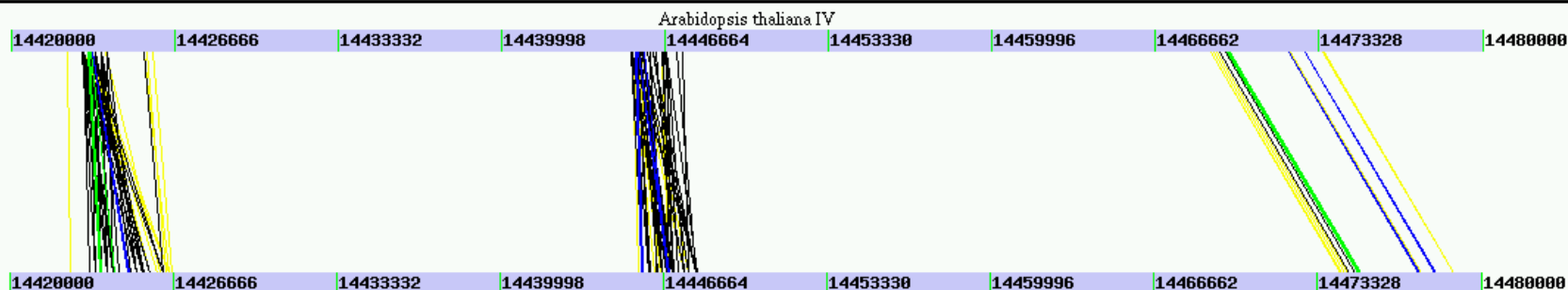
1000

vert

10000

rouge

Envoyer reset



[Visualiser et sauvegarder la fiche](#)

Organisme at4

Coordonnées sur la première séquence 14445563

Coordonnées sur la seconde séquence 14447014

Longueur de la répétition 371

```
> at4 14445563 371
tccatgtctagagatattcaagcogtgggtaccgaagcagagatcaatgtttgtcttaaga
cttagtctaaagactaaagtttgacogcgtgggtaccatttaacctccaccaaacataac
gcaatccaatctttgtctaaagagattggttacttcgaacctcaagctctgtctaaaga
cttttttaagcogtgggttaacttggcctttgtctaaagactaaaacagaaacaaaacat
aatccatagctctagagatattoaaagcogtgggtaccgaagcagagatcaatgtctagctc
agctctagctctaaagactaaagtttgaagctgtgggtgaccttaacctccaccaaacat
taacctaatcc
```

```
> at4 14447014 371
catatgtctagatattttcaagcogagggtaccgaagcagagatcaatgtctgtctcaagt
ctcaatctaaagactaaaatttgaagcogtgggtaccatttaacctccaccaaacataac
ccaatccaatctttgtctaaagaattggttatattgaacctcaagctctgtgtaaaga
cttttttaagcogtgggttaacttgaattttgtctaaagactgaagcagaaacccaacat
aatccatagctctagatattttcaagcogagggtaccgaagcagagatcaatgtctagctc
aagctctagctctaaagactaaagttagaagcogtgggtaccatctccctccaccaaacat
aacccaatcca
```

Précédent

Suivant

Recharger

Page d'accueil

Rechercher

Mozilla

Imprimer

Sécurité

Shopping

Arrêter

## FORRepeats

Choisissez les répétitions à visualiser

Ecoli

Longueurs min et max des répétitions 1100

Coordonnées min et max sur le premier chromosome

Coordonnées min et max sur le second chromosome

Distances min et max entre les répétitions

Zoom Map

jaune

100

noir

500

bleu

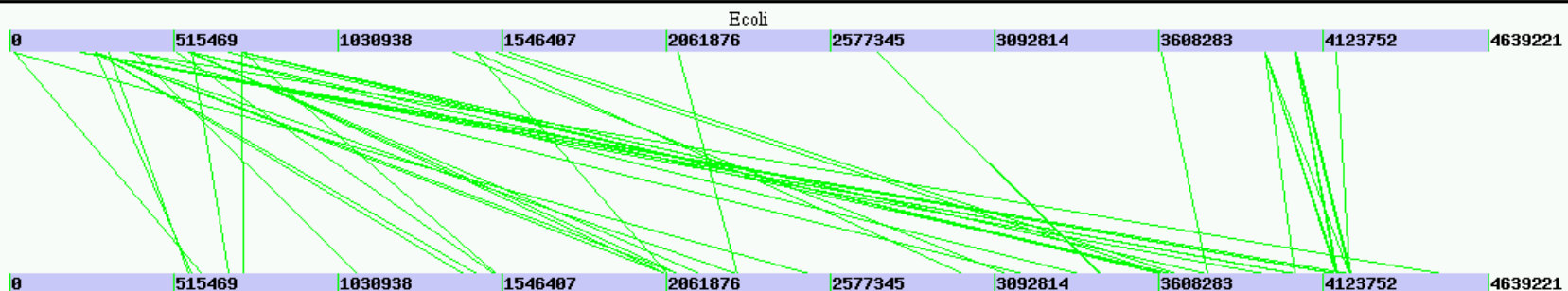
1000

vert

10000

rouge

Envoyer reset


[Visualiser et sauvegarder la fiche](#)

Organisme Ecoli

Coordonnées sur la première séquence 15168

Coordonnées sur la seconde séquence 607011

Longueur de la répétition 1773

```
> ecoli 15168 1773
cgaaagggcagaaacagctgctgcaagagctgcagaaagcttcgggtggcccaacgggoga
gcacaacagccggcgcacaagagcttcttggatggtgtgaagaagtttttgaagcact
gaaccgctaacctccccaaaagcctgcccgtgggagggcctgggttaaaaatagggctgct
tgaagatatgctgagcaccgttaaagtggcgggggacactcccataagcgtaacctaaagg
gttgggtattacgccgatattaaactggccgatgaattactctcagataaacctgg
toagcaattctggcccatttggtaagccgaagaactggatacttcggcagctaatgccc
ggggctctaacccggccgcgcaaatctgtgatgctgcaactctgctacgctggggctg
gcttacggccccgggggatgctattacgtgaagtaactgcatgggctcagctccatgac
gttgcaacattatctgactggctctcctgaagcggctgcccgaatgcccgcgactgggtt
```

```
> ecoli 607011 1773
gaacaattggatggtgctcctcttctgcatggaggcaatataaacatgctgaagaaat
atgccccttggggagctaatagctgtgttttaacggctggtgatttaagctctctggctg
gagactctgtgtgagtttaacgggtgaaggaaactaatattgagtttaagcctctctgctg
cttaagcaaccgaagaagttagccgttggcgggggagtaatcccataagcgtaacctaaagg
gttgggtattacgccgatattaaactggccgatgaattactctcagataaacctgg
toagcaattctggcccatttggtaagccgaagaactggatacttcggcagctaatgccc
ggggctctaacccggccgcgcaaatctgtgatgctgcaactctgctacgctggggctg
gcttacggccccgggggatgctattacgtgaagtaactgcatgggctcagctccatgac
gttgcaacattatctgactggctctcctgaagcggctgcccgaatgcccgcgactgggtt
```

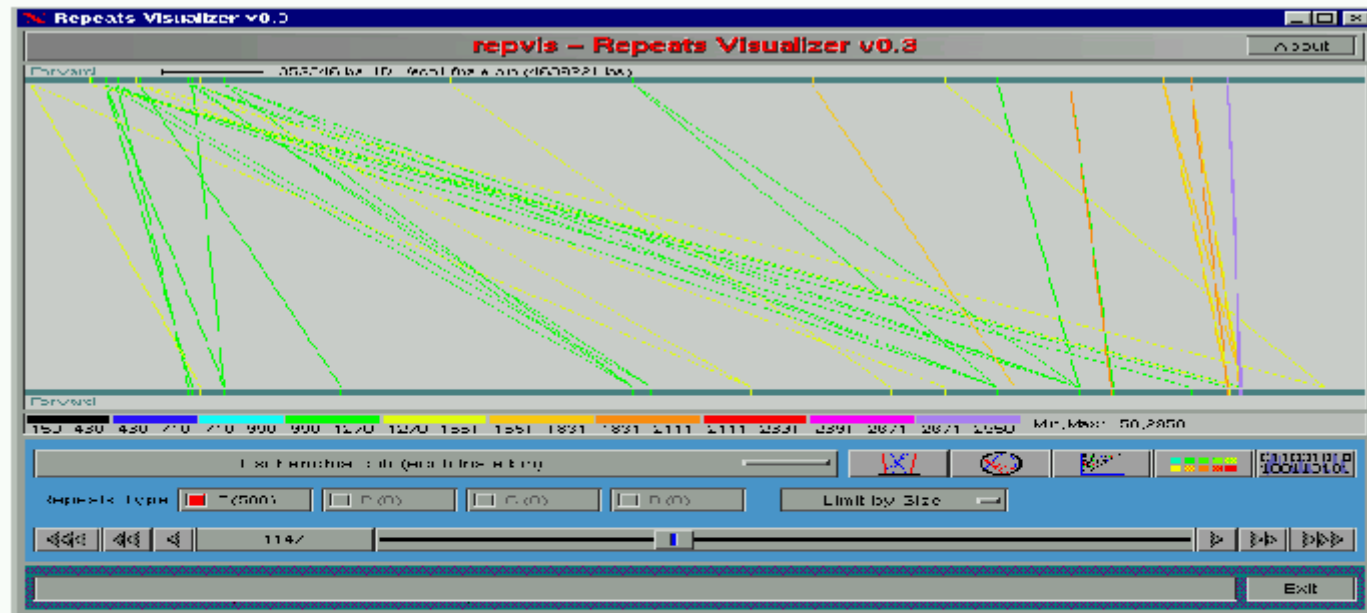


Figure 4: A typical application of *REPvis*, showing a view of the 50 most significant direct repeats in *E. coli* (4.6Mb), ranging from 1147 to 2950 bases in length. There are five repeats longer than the longest one found in *M. tuberculosis*; see Figure 5. In the main window graphics panel, two horizontal lines depict the input sequence and a copy of it. Diagonal lines stand for repeats by connecting their respective starting positions. Below the graphics panel, a choice box lists all calculated sequences in a user specified directory. Three further buttons switch the visualization mode to square graph, circle graph or dot plot. An additional button leads to the complete list of all repeats and their size distribution. Selector buttons specify which type of repeat to display. The symbols *F*, *P*, *C*, and *R* indicate direct (forward), palindromic (reverse complemented), complemented and reversed repeats; the number of repeats for each type is shown on the button.



# Résultats expérimentaux

	Mb	exact
<i>H. pylori</i>	1,59	2s
<i>B. subtilis</i>	4,02	6s
<i>E. coli</i>	4,42	6s
<i>S. cerevisiae</i>	11,50	14s
<i>A. thaliana II et IV</i>	35,47	57s
<i>C. elegans</i>	92,40	122s

500 MHz, 1 Go

# Compression séquentielle sans perte

On code les mots soit par une lettre lors de sa première occurrence soit par des couples (position, longueur) lors de répétitions de facteurs

On suppose que  $w[0..i]$  a déjà été codé

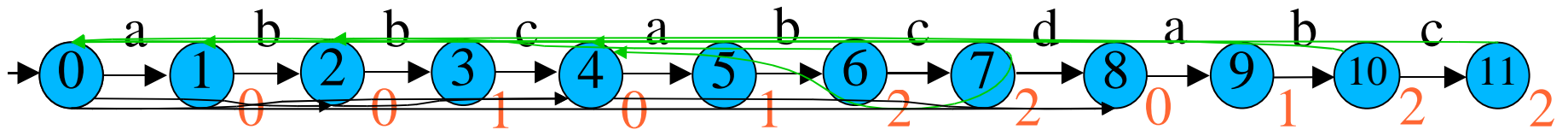
$$j = \min \{ k > i \mid k - lrs[k] > i + 1 \text{ et pour } i < h < k : \\ h - lrs[h] \leq i + 1 \}$$

alors on code :

- "  $(suf(j-1) - j + 1 + i, j - 1 - i)$
- "  $w[i + 1]$  si  $w[i + 1] \notin Alph(w[0..i])$

# Exemple de compression séquentielle

$w = \text{abbcabcdabc}$



$\text{ab}(1,1)\text{c}(0,2)(3,1)\text{d}(0,2)(3,1)$

# Compression séquentielle : résultats

Fichier	gzip			compror			bzip2		
	bpc	ct	dc	%	ct	dt	%	ct	dt
ChrIV	2,17	230,00	2,00	2,9	37,00	9,65	2,13	37,00	9,65
ChrII	2,18	261,00	2,00	2,9	42,00	10,90	2,14	40,00	12,00
bib	2,5	0,14	0,01	3,1	0,21	0,06	1,97	0,17	0,04
book1	3,25	1,18	0,08	3,8	2,00	0,50	2,42	1,47	0,47
book2	2,7	0,68	0,06	3,3	1,30	0,35	2,06	1,1	0,33
progc	2,67	0,04	0,01	3,8	0,09	0,02	2,53	0,07	0,02
trans	1,61	0,08	0,01	2,16	0,13	0,04	1,52	0,14	0,03
thesis	1,74	7,47	0,65	0,94	7,65	2,37	1,3	18,6	3,44
article1	1,65	5,14	2,5	1,2	37,5	2,5	1,65	5,14	2,5
article2	2,2	24,6	7,1	1,25	51,5	8,2	2,04	8,8	8,6
alpha	0,03	0,08	0,57	0,002	0,3	0,55	0,003	4,45	0,6

# Références

- " A. Lefebvre, T. Lecroq. Computing repeated factors with a factor oracle. *AWOCA 2000*.
- " A. Lefebvre, T. Lecroq. Compror: compression with a factor oracle. Poster à *DCC 2001*.
- " A. Lefebvre, T. Lecroq, J. Alexandre. Utilisations de l'oracle des facteurs. *JOBIM 2001*.
- " A. Lefebvre, T. Lecroq. Estimating topological entropy of biological sequences using a factor oracle. *SCI 2001*.
- " A. Lefebvre, T. Lecroq, H. Dauchel et J. Alexandre, FORRepeats, *Bioinformatics*, 2003.