

Coût de la mise à jour d'une table des suffixes

Martine Léonard, Laurent Mouchard, Mikaël Salson

LITIS, Université de Rouen

25 janvier 2010



Table des suffixes

- ▶ Structure d'indexation introduite en 1990.
- ▶ Recherche de motif de longueur m dans un texte T de longueur n en temps $O(m \log n)$.
- ▶ Tables compressées des suffixes (2000).

Table des suffixes

- ▶ Structure d'indexation introduite en 1990.
- ▶ Recherche de motif de longueur m dans un texte T de longueur n en temps $O(m \log n)$.
- ▶ Tables compressées des suffixes (2000).

Mise à jour

- ▶ Modifier la table des suffixes après l'insertion ou la suppression d'un facteur dans le texte.

Table des suffixes

- ▶ Structure d'indexation introduite en 1990.
- ▶ Recherche de motif de longueur m dans un texte T de longueur n en temps $O(m \log n)$.
- ▶ Tables compressées des suffixes (2000).

Mise à jour

- ▶ Modifier la table des suffixes après l'insertion ou la suppression d'un facteur dans le texte.
- ▶ Est-ce avantageux par rapport à la reconstruction ?

$$T = \begin{array}{c} 0 \ 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \\ \text{C G A G A C G A A \$} \end{array}$$

$$T = \begin{array}{c} 0 \ 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \\ \text{C G A G A C G A A \$} \end{array}$$

Suffixes triés

$T =$

	0	1	2	3	4	5	6	7	8	9
	C	G	A	G	A	C	G	A	A	\$



Suffixes triés

9 \$

$T =$

	0	1	2	3	4	5	6	7	8	9
	C	G	A	G	A	C	G	A	A	\$



Suffixes triés

9 \$

8 A \$

$T =$

	0	1	2	3	4	5	6	7	8	9
	C	G	A	G	A	C	G	A	A	\$



Suffixes triés

9 \$

8 A \$

7 A A \$

$T =$ 0 1 2 3 4 5 6 7 8 9
 C G A G A C G A A \$



Suffixes triés

9 \$

8 A \$

7 A A \$

4 A C G A A \$

$T =$ 0 1 2 3 4 5 6 7 8 9
 C G A G A C G A A \$



Suffixes triés

9 \$

8 A \$

7 A A \$

4 A C G A A \$

2 A G A C G A A \$

$T =$ 0 1 2 3 4 5 6 7 8 9
 C G A G A C G A A \$



Suffixes triés

9 \$

8 A \$

7 A A \$

4 A C G A A \$

2 A G A C G A A \$

5 C G A A \$

$T =$ 0 1 2 3 4 5 6 7 8 9
 C G A G A C G A A \$



Suffixes triés

9 \$

8 A \$

7 A A \$

4 A C G A A \$

2 A G A C G A A \$

5 C G A A \$

0 C G A G A C G A A \$

$T =$ 0 1 2 3 4 5 6 7 8 9
 C G A G A C G A A \$



Suffixes triés

9 \$

8 A \$

7 A A \$

4 A C G A A \$

2 A G A C G A A \$

5 C G A A \$

0 C G A G A C G A A \$

6 G A A \$

$T =$ 0 1 2 3 4 5 6 7 8 9
 C G A G A C G A A \$



Suffixes triés

9 \$

8 A \$

7 A A \$

4 A C G A A \$

2 A G A C G A A \$

5 C G A A \$

0 C G A G A C G A A \$

6 G A A \$

3 G A C G A A \$

$T =$

0	1	2	3	4	5	6	7	8	9
C	G	A	G	A	C	G	A	A	\$



Suffixes triés

9	\$
8	A \$
7	A A \$
4	A C G A A \$
2	A G A C G A A \$
5	C G A A \$
0	C G A G A C G A A \$
6	G A A \$
3	G A C G A A \$
1	G A G A C G A A \$

$T =$

0	1	2	3	4	5	6	7	8	9
C	G	A	G	A	C	G	A	A	\$

LCP	Suffixes triés
9	\$
8	A \$
7	A A \$
4	A C G A A \$
2	A G A C G A A \$
5	C G A A \$
0	C G A G A C G A A \$
6	G A A \$
3	G A C G A A \$
1	G A G A C G A A \$

$T =$

	0	1	2	3	4	5	6	7	8	9
	C	G	A	G	A	C	G	A	A	\$

LCP		Suffixes triés
	9	\$
0	8	A \$
	7	A A \$
	4	A C G A A \$
	2	A G A C G A A \$
	5	C G A A \$
	0	C G A G A C G A A \$
	6	G A A \$
	3	G A C G A A \$
	1	G A G A C G A A \$

$T =$ 0 1 2 3 4 5 6 7 8 9
 C G A G A C G A A \$

LCP		Suffixes triés
	9	\$
0	8	A \$
1	7	A A \$
	4	A C G A A \$
	2	A G A C G A A \$
	5	C G A A \$
	0	C G A G A C G A A \$
	6	G A A \$
	3	G A C G A A \$
	1	G A G A C G A A \$

$T =$ 0 1 2 3 4 5 6 7 8 9
 C G A G A C G A A \$

LCP		Suffixes triés
	9	\$
0	8	A \$
1	7	A A \$
1	4	A C G A A \$
	2	A G A C G A A \$
	5	C G A A \$
	0	C G A G A C G A A \$
	6	G A A \$
	3	G A C G A A \$
	1	G A G A C G A A \$

$T =$

	0	1	2	3	4	5	6	7	8	9
	C	G	A	G	A	C	G	A	A	\$

LCP		Suffixes triés
	9	\$
0	8	A \$
1	7	A A \$
1	4	A C G A A \$
1	2	A G A C G A A \$
	5	C G A A \$
	0	C G A G A C G A A \$
	6	G A A \$
	3	G A C G A A \$
	1	G A G A C G A A \$

$T =$ 0 1 2 3 4 5 6 7 8 9
 C G A G A C G A A \$

LCP		Suffixes triés
	9	\$
0	8	A \$
1	7	A A \$
1	4	A C G A A \$
1	2	A G A C G A A \$
0	5	C G A A \$
	0	C G A G A C G A A \$
	6	G A A \$
	3	G A C G A A \$
	1	G A G A C G A A \$

$T =$ 0 1 2 3 4 5 6 7 8 9
 C G A G A C G A A \$

LCP		Suffixes triés
	9	\$
0	8	A \$
1	7	A A \$
1	4	A C G A A \$
1	2	A G A C G A A \$
0	5	C G A A \$
3	0	C G A G A C G A A \$
	6	G A A \$
	3	G A C G A A \$
	1	G A G A C G A A \$

$T =$ 0 1 2 3 4 5 6 7 8 9
 C G A G A C G A A \$

LCP		Suffixes triés
	9	\$
0	8	A \$
1	7	A A \$
1	4	A C G A A \$
1	2	A G A C G A A \$
0	5	C G A A \$
3	0	C G A G A C G A A \$
0	6	G A A \$
	3	G A C G A A \$
	1	G A G A C G A A \$

$T =$ 0 1 2 3 4 5 6 7 8 9
 C G A G A C G A A \$

LCP		Suffixes triés
	9	\$
0	8	A \$
1	7	A A \$
1	4	A C G A A \$
1	2	A G A C G A A \$
0	5	C G A A \$
3	0	C G A G A C G A A \$
0	6	G A A \$
2	3	G A C G A A \$
	1	G A G A C G A A \$

$T =$ 0 1 2 3 4 5 6 7 8 9
 C G A G A C G A A \$

LCP		Suffixes triés
	9	\$
0	8	A \$
1	7	A A \$
1	4	A C G A A \$
1	2	A G A C G A A \$
0	5	C G A A \$
3	0	C G A G A C G A A \$
0	6	G A A \$
2	3	G A C G A A \$
2	1	G A G A C G A A \$

$T =$

	0	1	2	3	4	5	6	7	8	9
	C	G	A	G	A	C	G	A	A	\$

LCP		Suffixes triés
	9	\$
0	8	A \$
1	7	A A \$
1	4	A C G A A \$
1	2	A G A C G A A \$
0	5	C G A A \$
3	0	C G A G A C G A A \$
0	6	G A A \$
2	3	G A C G A A \$
2	1	G A G A C G A A \$

Méthodes existantes

- ▶ Gallé, Peterlongo et Coste, 2008 : mise à jour d'une table des suffixes pour l'inférence grammaticale ;
- ▶ Salson, Lecroq, Léonard et Mouchard, 2009 : mise à jour d'une table compressée des suffixes.

Méthodes existantes

- ▶ Gallé, Peterlongo et Coste, 2008 : mise à jour d'une table des suffixes pour l'inférence grammaticale ;
- ▶ Salson, Lecroq, Léonard et Mouchard, 2009 : mise à jour d'une table compressée des suffixes.

Principe commun

Certains suffixes n'ont plus le même rang dans l'ordre alphabétique et doivent être déplacés.

$$T = \begin{array}{c} 0 \ 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \\ C \ G \ A \ G \ A \ C \ G \ A \ A \ \$ \end{array}$$

$$T = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ & C & G & A & G & A & C & G & A & A & \$ \end{matrix}$$

Suffixes triés

9 \$

8 A\$

7 AA\$

4 ACGAA\$

2 AGACGAA\$

5 CGAA\$

0 CGAGACGAA\$

6 GAA\$

3 GACGAA\$

1 GAGACGAA\$

$T' = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ & C & G & A & G & A & C & G & \color{red}{C} & A & A & \$ \end{matrix}$

Suffixes triés

9 \$

8 A\$

7 AA\$

4 ACGAA\$

2 AGACGAA\$

5 CGAA\$

0 CGAGACGAA\$

6 GAA\$

3 GACGAA\$

1 GAGACGAA\$

$T' =$

	0	1	2	3	4	5	6	7	8	9	10
	C	G	A	G	A	C	G	C	A	A	\$

déplacés vers la droite

Suffixes triés

9 \$

8 A\$

7 AA\$

4 ACGAA\$

2 AGACGAA\$

5 CGAA\$

0 CGAGACGAA\$

6 GAA\$

3 GACGAA\$

1 GAGACGAA\$

$T' =$

0	1	2	3	4	5	6	7	8	9	10
C	G	A	G	A	C	G	C	A	A	\$

déplacés vers la droite

Suffixes triés

10 $\xleftarrow{+1}$ 9 \$

9 $\xleftarrow{+1}$ 8 A\$

8 $\xleftarrow{+1}$ 7 AA\$

4 ACGAA\$

2 AGACGAA\$

5 CGAA\$

0 CGAGACGAA\$

6 GAA\$

3 GACGAA\$

1 GAGACGAA\$

$T' = \overset{0}{C} \overset{1}{G} \overset{2}{A} \overset{3}{G} \overset{4}{A} \overset{5}{C} \overset{6}{G} \overset{7}{C} \overset{8}{A} \overset{9}{A} \overset{10}{\$}$

Suffixes triés

10 \$

9 A\$

8 AA\$

4 ACGAA\$

2 AGACGAA\$

5 CGAA\$

0 CGAGACGAA\$

6 GAA\$

3 GACGAA\$

1 GAGACGAA\$

$T' = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ & C & G & A & G & A & C & G & C & A & A & \$ \end{matrix}$

Suffixes triés

10 \$

9 A\$

8 AA\$

4 ACGAA\$

2 AGACGAAS
 ← CAA\$

5 CGAA\$

0 CGAGACGAAS

6 GAA\$

3 GACGAAS

1 GAGACGAAS

$T' =$

0	1	2	3	4	5	6	7	8	9	10
C	G	A	G	A	C	G	C	A	A	\$

Suffixes triés

10 \$

9 A\$

8 AA\$

4 ACGAA\$

2 AGACGAA\$
← CAA\$

5 CGAA\$

0 CGAGACGAA\$

→ 6 GAA\$

3 GACGAA\$

1 GAGACGAA\$

$T' =$

0	1	2	3	4	5	6	7	8	9	10
C	G	A	G	A	C	G	C	A	A	\$

Suffixes triés

10 \$

9 A\$

8 AA\$

4 ACGAA\$

2 AGACGAAS

5 CGAA\$

0 CGAGACGAAS

→ 6 ~~GAAS~~

3 GACGAAS

1 GAGACGAAS

$T' =$

0	1	2	3	4	5	6	7	8	9	10
C	G	A	G	A	C	G	C	A	A	\$

Suffixes triés

10 \$

9 A\$

8 AA\$

4 ACGAA\$

2 AGACGAAS

5 CGAA\$

0 CGAGACGAAS

→ 6 ~~GAAS~~ GCAAS

3 GACGAAS

1 GAGACGAAS

$T' =$

0	1	2	3	4	5	6	7	8	9	10
C	G	A	G	A	C	G	C	A	A	\$

Suffixes triés

10 \$

9 A\$

8 AA\$

4 ACGAA\$

2 AGACGAAS

5 CGAA\$

0 CGAGACGAAS

→ 6 ~~GAAS~~ GCAAS

3 GACGAAS

1 GAGACGAAS



$T' =$

	0	1	2	3	4	5	6	7	8	9	10
	C	G	A	G	A	C	G	C	A	A	\$

Suffixes triés

10 \$

9 A\$

8 AA\$

4 ACGAA\$

2 AGACGAAS

→ 5 CGAA\$

0 CGAGACGAAS

6 ~~GAAS~~ GCAAS

3 GACGAAS

1 GAGACGAAS



$T' =$

0	1	2	3	4	5	6	7	8	9	10
C	G	A	G	A	C	G	C	A	A	\$

Suffixes triés

10 \$

9 A\$

8 AA\$

4 ACGAA\$

2 AGACGAAS CAA\$

→ 5 ~~CGAA\$~~

0 CGAGACGAAS

6 ~~GAAS~~ GCAAS

3 GACGAAS

1 GAGACGAAS



$T' =$

0	1	2	3	4	5	6	7	8	9	10
C	G	A	G	A	C	G	C	A	A	\$

Suffixes triés

10 \$

9 A\$

8 AA\$

4 ACGAA\$

2 AGACGAAS

→ 5 ~~CGAA\$~~ CGCAAS

0 CGAGACGAAS

6 ~~GAAS~~ GCAAS

3 GACGAAS

1 GAGACGAAS

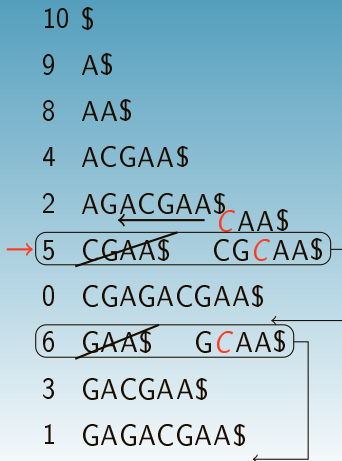


$T' =$

0	1	2	3	4	5	6	7	8	9	10
C	G	A	G	A	C	G	C	A	A	\$

Suffixes triés

10 \$
 9 A\$
 8 AA\$
 4 ACGAA\$
 2 AGACGAAS\$
 5 ~~CGAAS\$~~ CGCAAS\$
 0 CGAGACGAAS\$
 6 ~~GAAS\$~~ GCAAS\$
 3 GACGAAS\$
 1 GAGACGAAS\$



$T' =$

0	1	2	3	4	5	6	7	8	9	10
C	G	A	G	A	C	G	C	A	A	\$

Suffixes triés

10 \$

9 A\$

8 AA\$

→ 4 ACGAA\$

2 AGACGAAS

5 ~~CGAA\$~~ CGCAAS

0 CGAGACGAAS

6 ~~GAAS~~ GCAAS

3 GACGAAS

1 GAGACGAAS



$T' =$

0	1	2	3	4	5	6	7	8	9	10
C	G	A	G	A	C	G	C	A	A	\$

Suffixes triés

10 \$

9 A\$

8 AA\$

→ 4 ~~ACGAAS~~

2 AGACGAAS

5 ~~CGAAS~~ CGAAS

0 CGAGACGAAS

6 ~~GAA\$~~ GAA\$

3 GACGAAS

1 GAGACGAAS

$T' =$

	0	1	2	3	4	5	6	7	8	9	10
	C	G	A	G	A	C	G	C	A	A	\$

Suffixes triés

10 \$

9 A\$

8 AA\$

→ 4 ~~ACGAAS~~ ACGC~~AA~~\$

2 AGACGAAS

5 ~~CGAAS~~ CGC~~AA~~\$

0 CGAGACGAAS

6 ~~GAA\$~~ GC~~AA~~\$

3 GACGAAS

1 GAGACGAAS



$T' =$

0	1	2	3	4	5	6	7	8	9	10
C	G	A	G	A	C	G	C	A	A	\$

Suffixes triés

10 \$

9 A\$

8 AA\$

→ 4 ~~ACGAAS~~ ACGC~~AA~~\$

2 AGACGAAS

5 ~~CGAAS~~ CGC~~AA~~\$

0 CGAGACGAAS

6 ~~GAA\$~~ GC~~AA~~\$

3 GACGAAS

1 GAGACGAAS



$T' = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ & C & G & A & G & A & C & G & C & A & A & \$ \end{matrix}$

Suffixes triés

10 \$

9 A\$

8 AA\$

4 ~~ACGAAS~~ ACGCAAS\$

2 AGACGAAS\$
← CAA\$

5 ~~CGAAS~~ CGCAAS\$

0 CGAGACGAAS\$

6 ~~GAAS~~ GCAAS\$

3 GACGAAS\$

1 GAGACGAAS\$



$T' = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ & C & G & A & G & A & C & G & C & A & A & \$ \end{matrix}$

Suffixes triés

10 10 \$

9 A\$

8 AA\$

4 ~~ACGAAS~~ ACGCAAS\$

2 AGACGAAS\$

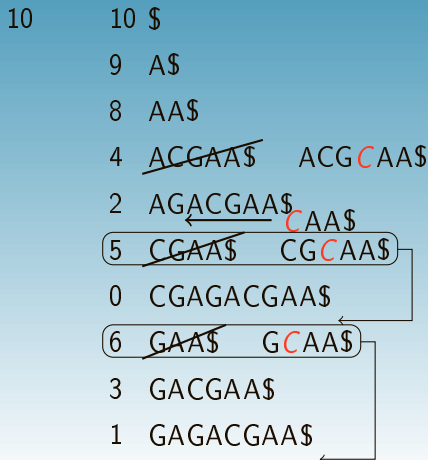
5 ~~CGAAS~~ CGCAAS\$

0 CGAGACGAAS\$

6 ~~GAA\$~~ GCAAS\$

3 GACGAAS\$

1 GAGACGAAS\$



$T' = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \text{C} & \text{G} & \text{A} & \text{G} & \text{A} & \text{C} & \text{G} & \text{C} & \text{A} & \text{A} & \$ \end{matrix}$

Suffixes triés

10	10	\$
9	9	A\$
	8	AA\$
	4	ACGAAS ACGC AAS\$
	2	AGACGAAS\$
	5	CGAAS CGC AAS\$
	0	CGAGACGAAS\$
	6	GAAS GC AAS\$
	3	GACGAAS\$
	1	GAGACGAAS\$

$T' = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ & C & G & A & G & A & C & G & C & A & A & \$ \end{matrix}$

Suffixes triés

10	10	\$
9	9	A\$
8	8	AA\$
4	ACGAAS	ACGCAAS\$
2	AGACGAAS\$	← CAA\$
5	CGAAS	CGCAAS\$
0	CGAGACGAAS\$	
6	GAA\$	GCAAS\$
3	GACGAAS\$	
1	GAGACGAAS\$	



$T' = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ & C & G & A & G & A & C & G & C & A & A & \$ \end{matrix}$

Suffixes triés

10	10	\$
9	9	A\$
8	8	AA\$
4	4	ACGAAS ACGCAAS\$
2	2	AGACGAAS\$
5	5	CGAAS CGCAAS\$
0	0	CGAGACGAAS\$
6	6	GAAS GCAAS\$
3	3	GACGAAS\$
1	1	GAGACGAAS\$

Diagram illustrating the update of a suffix table. The table lists suffixes for the string T' = CGAGACGCAAS\$. The suffixes are sorted by their starting index. The diagram shows the process of inserting a new suffix (index 4) and updating the existing ones. The suffixes are: 10 \$, 9 A\$, 8 AA\$, 4 ~~ACGAAS~~ ACGCAAS\$, 2 AGACGAAS\$, 5 ~~CGAAS~~ CGCAAS\$, 0 CGAGACGAAS\$, 6 ~~GAAS~~ GCAAS\$, 3 GACGAAS\$, 1 GAGACGAAS\$. Arrows indicate the insertion and linking of the new suffix (index 4) into the sorted order.

$T' = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \text{C} & \text{G} & \text{A} & \text{G} & \text{A} & \text{C} & \text{G} & \text{C} & \text{A} & \text{A} & \$ \end{matrix}$

Suffixes triés

10	10	\$
9	9	A\$
8	8	AA\$
4	4	ACGAAS ACGC AAS\$
2	2	AGACGAAS\$
		← CAA\$
	5	CGAAS CGC AAS\$
	0	CGAGACGAAS\$
	6	GAAS GC AAS\$
	3	GACGAAS\$
	1	GAGACGAAS\$



$T' = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ & C & G & A & G & A & C & G & C & A & A & \$ \end{matrix}$

Suffixes triés

10	10	\$
9	9	A\$
8	8	AA\$
4	4	ACGAAS ACGCAAS\$
2	2	AGACGAAS\$
7		← CAA\$
	5	CGAAS CGCAAS\$
	0	CGAGACGAAS\$
	6	GAAS GCAAS\$
	3	GACGAAS\$
	1	GAGACGAAS\$

$T' = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ & C & G & A & G & A & C & G & C & A & A & \$ \end{matrix}$

Suffixes triés

10	10	\$
9	9	A\$
8	8	AA\$
4	4	ACGAAS ACGCAAS\$
2	2	AGACGAAS\$
7		← CAA\$
0	5	CGAAS CGCAAS\$
	0	CGAGACGAAS\$
	6	GAAS GCAAS\$
	3	GACGAAS\$
	1	GAGACGAAS\$

$T' = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ & C & G & A & G & A & C & G & C & A & A & \$ \end{matrix}$

Suffixes triés

10	10	\$
9	9	A\$
8	8	AA\$
4	4	ACGAAS ACGCAAS\$
2	2	AGACGAAS\$
7		← CAA\$
0	5	CGAAS CGCAAS\$
5	0	CGAGACGAAS\$
	6	GAAS GCAAS\$
	3	GACGAAS\$
	1	GAGACGAAS\$

$T' = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ & C & G & A & G & A & C & G & C & A & A & \$ \end{matrix}$

Suffixes triés

10	10	\$
9	9	A\$
8	8	AA\$
4	4	ACGAAS ACGCAAS\$
2	2	AGACGAAS\$
7		← CAA\$
0	5	CGAAS CGCAAS\$
5	0	CGAGACGAAS\$
3	6	GAA\$ GCAAS\$
	3	GACGAAS\$
	1	GAGACGAAS\$

$T' = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ & C & G & A & G & A & C & G & C & A & A & \$ \end{matrix}$

Suffixes triés

10	10	\$
9	9	A\$
8	8	AA\$
4	4	ACGAAS ACGCAAS\$
2	2	AGACGAAS\$
7		← CAA\$
0	5	CGAAS CGCAAS\$
5	0	CGAGACGAAS\$
3	6	GAA\$ GCAAS\$
1	3	GACGAAS\$
	1	GAGACGAAS\$

$T' = \begin{matrix} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \text{C} & \text{G} & \text{A} & \text{G} & \text{A} & \text{C} & \text{G} & \text{C} & \text{A} & \text{A} & \$ \end{matrix}$

Suffixes triés

10	10	\$
9	9	A\$
8	8	AA\$
4	4	ACGAAS ACGC AAS\$
2	2	AGACGAAS\$
7		← CAA\$
0	5	CGAAS CGC AAS\$
5	0	CGAGACGAAS\$
3	6	GAA\$ GC AAS\$
1	3	GACGAAS\$
6	1	GAGACGAAS\$

Question

Combien a-t-on de suffixes mal ordonnés, dans la table des suffixes, après une modification du texte ?

Question

Combien a-t-on de suffixes mal ordonnés, dans la table des suffixes, après une modification du texte ?

Réponse

Au pire, de l'ordre de n (par exemple $A^n\$$, modifié en $A^nB\$$).

Question

Combien a-t-on de suffixes mal ordonnés, dans la table des suffixes, après une modification du texte ?

Réponse

Au pire, de l'ordre de n (par exemple $A^n\$$, modifié en $A^nB\$$).

Remarque

Les expériences montrent que les approches de mise à jour sont plus rapides que les algorithmes de reconstruction

Question

Combien a-t-on de suffixes mal ordonnés, dans la table des suffixes, après une modification du texte ?

Réponse

Au pire, de l'ordre de n (par exemple $A^n\$$, modifié en $A^nB\$$).

Remarque

Les expériences montrent que les approches de mise à jour sont plus rapides que les algorithmes de reconstruction

Déduction

Le nombre de réordonnements en moyenne, doit donc être très inférieur à n .

Identifions ce qui caractérise un réordonnement.

Retour sur l'exemple

GAA\$

GACGAA\$

GAGACGAA\$

\$

A\$

AA\$

ACGAA\$

AGACGAA\$

CGAA\$

CGAGACGAA\$

GAA\$

GACGAA\$

GAGACGAA\$

Identifions ce qui caractérise un réordonnement.

Retour sur l'exemple

GAAS
GACGAAS
GAGACGAAS
LCP

\$
A\$
AA\$
ACGAAS
AGACGAAS
CGAAS
CGAGACGAAS
GAAS
GACGAAS
GAGACGAAS

Identifions ce qui caractérise un réordonnement.

Retour sur l'exemple

G C A A \$
 G A C G C A A \$
 G A G A C G C A A \$
 LCP

\$
 A \$
 A A \$
 A C G A A \$
 A G A C G A A \$
 C G A A \$
 C G A G A C G A A \$
 G A A \$
 G A C G A A \$
 G A G A C G A A \$

Identifions ce qui caractérise un réordonnement.

Retour sur l'exemple

GCAAS\$
 GACGCAA\$
 GAGACGCAA\$
 LCP

\$
 A\$
 AA\$
 ACGAA\$
 AGACGAA\$
 CGAA\$
 CGAGACGAA\$
 GAA\$
 GACGAA\$
 GAGACGAA\$

Identifions ce qui caractérise un réordonnement.

Retour sur l'exemple

$\begin{array}{l} \boxed{G}CAAS \\ > \\ GACGCAA\$ \\ \boxed{G}AGACGCAA\$ \end{array}$
 LCP

\$
 A\$
 AA\$
 ACGAA\$
 AGACGAA\$
 CGAA\$
 CGAGACGAA\$
 GAA\$
 GACGAA\$
 GAGACGAA\$

Identifions ce qui caractérise un réordonnement.

Retour sur l'exemple

$\begin{array}{l} \boxed{G}CAAS \\ > \\ GACGCAAS \\ \boxed{GAGACG}CAAS \end{array}$
 LCP ←

\$
 A\$
 AA\$
 ACGAA\$
 AGACGAA\$
 CGAA\$
 CGAGACGAA\$
 GAA\$
 GACGAA\$
 GAGACGAA\$

Identifions ce qui caractérise un réordonnement.

Retour sur l'exemple

CGAAS\$
CGAGACGAAS\$
LCP

\$
A\$
AA\$
ACGAAS\$
AGACGAAS\$
CGAAS\$
CGAGACGAAS\$
GAA\$
GACGAAS\$
GAGACGAAS\$

Identifions ce qui caractérise un réordonnement.

Retour sur l'exemple

CGCAA\$
CGAGACGCAA\$
LCP

\$
A\$
AA\$
ACGAA\$
AGACGAA\$
CGAA\$
CGAGACGAA\$
GAA\$
GACGAA\$
GAGACGAA\$

Identifions ce qui caractérise un réordonnement.

Retour sur l'exemple

$\begin{array}{l} \boxed{\text{CG}} \text{CAA\$} \\ \text{CGAGACG} \text{CAA\$} \end{array}$
 LCP ←

\$
 A\$
 AA\$
 ACGAA\$
 AGACGAA\$
 CGAA\$
 CGAGACGAA\$
 GAA\$
 GACGAA\$
 GAGACGAA\$

Identifions ce qui caractérise un réordonnement.

Retour sur l'exemple

ACGAAS
 AGACGAAS
 LCP

\$
 A\$
 AA\$
 ACGAA\$
 AGACGAAS\$
 CGAAS\$
 CGAGACGAAS\$
 GAA\$
 GACGAAS\$
 GAGACGAAS\$

Identifions ce qui caractérise un réordonnement.

Retour sur l'exemple

ACGCAA\$
 AGACGCAA\$
 LCP

\$
 A\$
 AA\$
 ACGAA\$
 AGACGAA\$
 CGAA\$
 CGAGACGAA\$
 GAA\$
 GACGAA\$
 GAGACGAA\$

Identifions ce qui caractérise un réordonnement.

Retour sur l'exemple

$\begin{array}{c} \boxed{\text{ACG}}\text{CAA\$} \\ < \\ \boxed{\text{AGACG}}\text{CAA\$} \end{array}$
 LCP

\$
 A\$
 AA\$
 ACGAA\$
 AGACGAA\$
 CGAA\$
 CGAGACGAA\$
 GAA\$
 GACGAA\$
 GAGACGAA\$

Identifions ce qui caractérise un réordonnement.

Retour sur l'exemple

	\$
	A\$
	AA\$
	ACGAA\$
	AGACGAA\$
	CGAA\$
	CGAGACGAA\$
	GAA\$
	GACGAA\$
	GAGACGAA\$

ACG	CAA\$
<	
AGACG	CAA\$

LCP

Conclusion

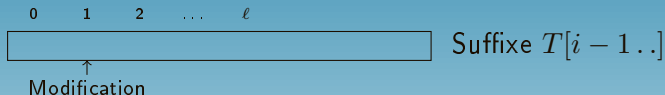
	LCP	
Suffixe		
	Modification ici $\Rightarrow$ ré- ordonnement possible	Modification ici $\Rightarrow$ réor- donnement impossible

Soit $r[i]$ le nombre maximal de suffixes à réordonner après une modification dans le texte en position i .

Soit $r[i]$ le nombre maximal de suffixes à réordonner après une modification dans le texte en position i .

Exemple

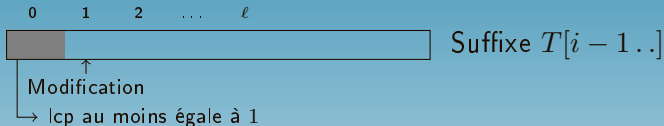
$r[i] = \ell$ si et seulement si :



Soit $r[i]$ le nombre maximal de suffixes à réordonner après une modification dans le texte en position i .

Exemple

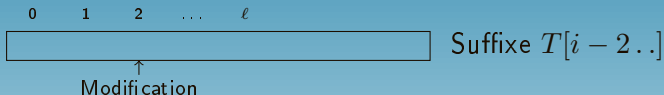
$r[i] = \ell$ si et seulement si :



Soit $r[i]$ le nombre maximal de suffixes à réordonner après une modification dans le texte en position i .

Exemple

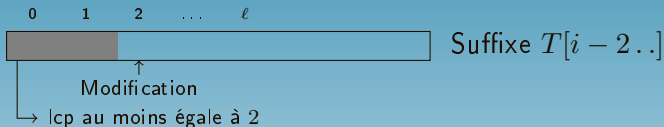
$r[i] = \ell$ si et seulement si :



Soit $r[i]$ le nombre maximal de suffixes à réordonner après une modification dans le texte en position i .

Exemple

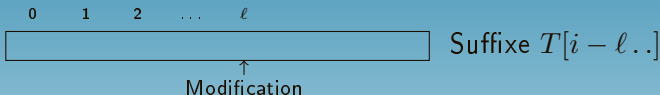
$r[i] = \ell$ si et seulement si :



Soit $r[i]$ le nombre maximal de suffixes à réordonner après une modification dans le texte en position i .

Exemple

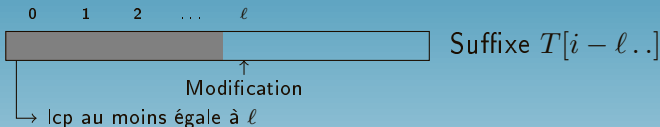
$r[i] = \ell$ si et seulement si :



Soit $r[i]$ le nombre maximal de suffixes à réordonner après une modification dans le texte en position i .

Exemple

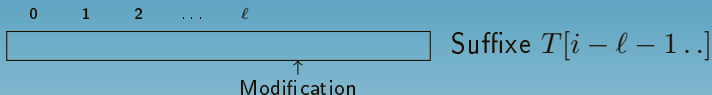
$r[i] = \ell$ si et seulement si :



Soit $r[i]$ le nombre maximal de suffixes à réordonner après une modification dans le texte en position i .

Exemple

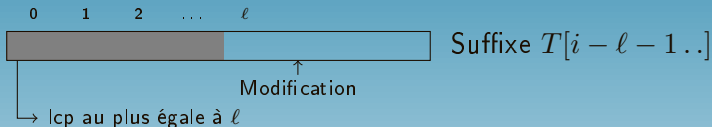
$r[i] = \ell$ si et seulement si :



Soit $r[i]$ le nombre maximal de suffixes à réordonner après une modification dans le texte en position i .

Exemple

$r[i] = \ell$ si et seulement si :



Soit $r[i]$ le nombre maximal de suffixes à réordonner après une modification dans le texte en position i .

Théorème

La table r est une permutation de la table LCP.

Soit $r[i]$ le nombre maximal de suffixes à réordonner après une modification dans le texte en position i .

Théorème

La table r est une permutation de la table LCP.

Corollaire

Pour un texte donné, on a :

- ▶ L_{ave} , la valeur moyenne des LCP, est le nombre moyen de réordonnements à réaliser, au maximum ;
- ▶ L_{max} , la valeur maximale des LCP, est le nombre maximal de réordonnements qui seront réalisés ;

On connaît maintenant le nombre moyen de réordonnements pour un texte : L_{ave} .

On connaît maintenant le nombre moyen de réordonnements pour un texte : L_{ave} .

Que sait-on sur L_{ave} ?

On connaît maintenant le nombre moyen de réordonnements pour un texte : L_{ave} .

Que sait-on sur L_{ave} ?

- ▶ varie de 0 à $n/2$ en fonction des textes;

On connaît maintenant le nombre moyen de réordonnements pour un texte : L_{ave} .

Que sait-on sur L_{ave} ?

- ▶ varie de 0 à $n/2$ en fonction des textes ;
- ▶ logarithmique pour des textes générés aléatoirement ;

On connaît maintenant le nombre moyen de réordonnements pour un texte : L_{ave} .

Que sait-on sur L_{ave} ?

- ▶ varie de 0 à $n/2$ en fonction des textes ;
- ▶ logarithmique pour des textes générés aléatoirement ;

En pratique

On indexe principalement des :

On connaît maintenant le nombre moyen de réordonnements pour un texte : L_{ave} .

Que sait-on sur L_{ave} ?

- ▶ varie de 0 à $n/2$ en fonction des textes ;
- ▶ logarithmique pour des textes générés aléatoirement ;

En pratique

On indexe principalement des :

- ▶ séquences génomiques ;

On connaît maintenant le nombre moyen de réordonnements pour un texte : L_{ave} .

Que sait-on sur L_{ave} ?

- ▶ varie de 0 à $n/2$ en fonction des textes ;
- ▶ logarithmique pour des textes générés aléatoirement ;

En pratique

On indexe principalement des :

- ▶ séquences génomiques ;
- ▶ textes en langages naturels (formatés ou non).

Séquences les plus répétées

Haubold et Wiehe (2006) ont classé les séquences génomiques de la plus répétée à la moins répétée (au sens informatique).

Résultat : *Methylobacillus Flagellatus* est la plus répétée.

Séquences les plus répétées

Haubold et Wiehe (2006) ont classé les séquences génomiques de la plus répétée à la moins répétée (au sens informatique).

Résultat : *Methylobacillus Flagellatus* est la plus répétée.

Résultats

Nom	Taille	L_{max}	L_{ave}
Les deux séquences les plus répétées			
<i>M. flagellatus</i>	2 971 519	143 034	3 452
<i>S. agalactiae</i>	2 211 485	47 068	546
Séquences génomiques d'intérêt			
<i>D. melanogaster</i>	120 290 946	30 892	66
<i>C. elegans</i>	100 269 917	38 987	45

Choix des textes

- textes bruts : concaténation d'ouvrages du projet Gutenberg ;

Choix des textes

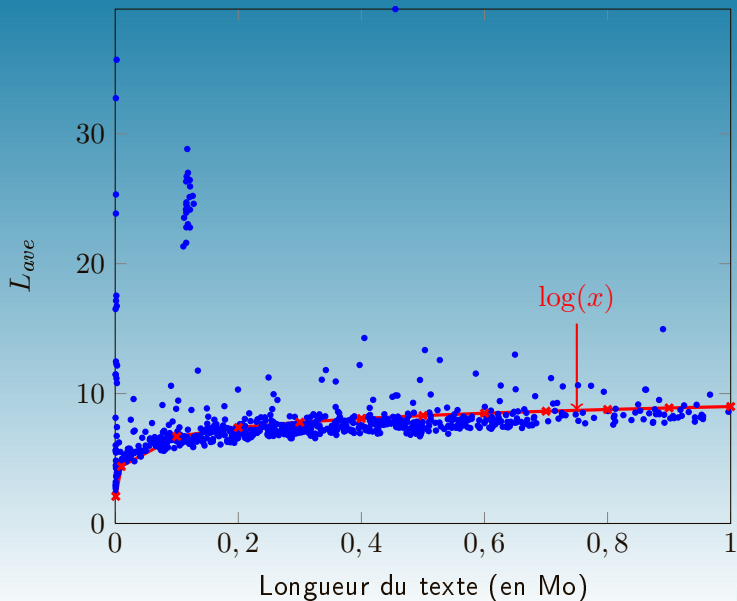
- ▶ textes bruts : concaténation d'ouvrages du projet Gutenberg ;
- ▶ textes formatés : corpus Wikipedia.

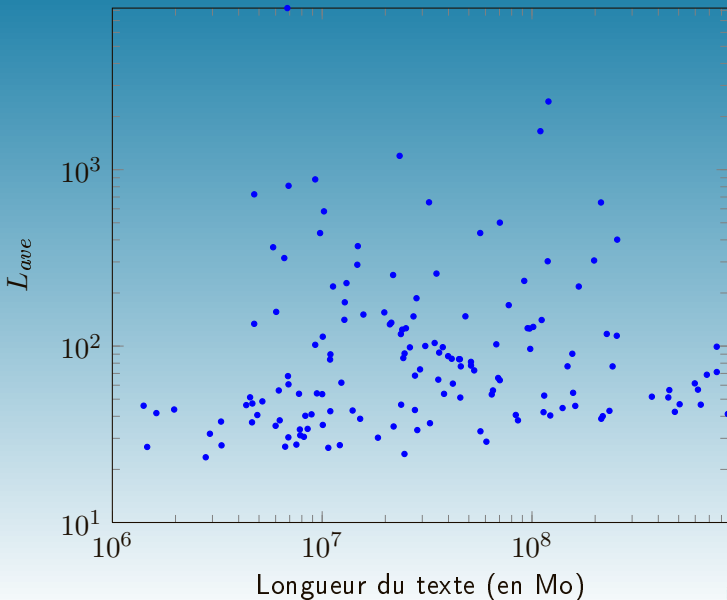
Choix des textes

- ▶ textes bruts : concaténation d'ouvrages du projet Gutenberg ;
- ▶ textes formatés : corpus Wikipedia.

Résultats

Nom	Taille	L_{max}	L_{ave}
Gutenberg	91 070 340	1 727	12
Wikip. afrikaans	68 989 658	34 205	66
Wikip. latin	85 735 379	8 099	38





En théorie

Le nombre de réordonnements est logarithmique en moyenne.

En théorie

Le nombre de réordonnements est logarithmique en moyenne.

Sur des textes d'intérêts

En théorie

Le nombre de réordonnements est logarithmique en moyenne.

Sur des textes d'intérêts

- ▶ Séquences génomiques : au pire $n/1000$ éléments réordonnés en moyenne ;

En théorie

Le nombre de réordonnements est logarithmique en moyenne.

Sur des textes d'intérêts

- ▶ Séquences génomiques : au pire $n/1000$ éléments réordonnés en moyenne ;
- ▶ textes bruts : comportement logarithmique ;

En théorie

Le nombre de réordonnements est logarithmique en moyenne.

Sur des textes d'intérêts

- ▶ Séquences génomiques : au pire $n/1000$ éléments réordonnés en moyenne ;
- ▶ textes bruts : comportement logarithmique ;
- ▶ textes formatés (Wikipedia) : comportement imprévisible (mais $< n/1000$).