

# Table des suffixes augmentées

Thierry Lecroq

Thierry.Lecroq@univ-rouen.fr

Laboratoire d'Informatique, Traitement de l'Information, Systèmes.



# Plan

- 1 Introduction
- 2 Arbre des lcp-intervalles
- 3 Table des suffixes augmentées

# Plan

- 1 Introduction
- 2 Arbre des lcp-intervalles
- 3 Table des suffixes augmentées

# Table des suffixes

Exemple pour  $y = \text{acaacatat}\$$

$i$	$p[i]$	$LCP[i]$											
0	2	0	a	a	a	c	a	t	a	t	\$		
1	3	2	a	a	c	a	t	a	t	\$			
2	0	1	a	c	a	a	a	c	a	t	a	t	\$
3	4	3	a	c	a	t	a	t	\$				
4	6	1	a	t	a	t	\$						
5	8	2	a	t	\$								
6	1	0	c	a	a	a	c	a	t	a	t	\$	
7	5	2	c	a	t	a	t	\$					
8	7	0	t	a	t	\$							
9	9	1	t	\$									
10	10	0	\$										

# lcp-intervalle

Un intervalle  $[i; j]$ ,  $0 \leq i < j \leq n$ , est un **lcp-intervalle** de valeur **lcp**  $\ell$  si

- 1  $LCP[i] < \ell$ ,
- 2  $LCP[k] \geq \ell$  pour  $i + 1 \leq k \leq j$ ,
- 3  $LCP[k] = \ell$  pour au moins un  $k$  avec  $i + 1 \leq k \leq j$ ,
- 4  $LCP[j + 1] < \ell$ .

On parlera de  **$\ell$ -intervalle** (ou même  **$\ell$ - $[i; j]$** ) pour un lcp-intervalle de valeur lcp  $\ell$ .

# $\ell$ indices

Chaque indice  $k$ ,  $i + 1 \leq k \leq j$ , avec  $LCP[k] = \ell$  est appelé un  $\ell$ -indice.

L'ensemble des  $\ell$ -indices d'un  $\ell$ -intervalle  $[i; j]$  est noté  $\ell\text{indices}(i, j)$ .

Si  $[i; j]$  est un  $\ell$ -intervalle tel que  
 $w = y[p[i]..p[i] + \ell - 1]$   
est le plus long préfixe commun aux suffixes  
 $y[p[i]..n], y[p[i + 1]..n], \dots, y[p[j]..n]$   
alors  $[i; j]$  est appelé un  $w$ -intervalle.

## Exemple

Pour  $y = \text{acaaacatat}\$$

$[0; 5]$  est un 1-intervalle puisque

- 1  $p[0] = 0 < 1$ ,
- 2  $LCP[k] \geq 1$  pour  $1 \leq k \leq 5$ ,
- 3  $LCP[2] = 1$ ,
- 4  $LCP[5 + 1] = 0 < 1$ .

$[0; 5]$  est un a-intervalle et  $\ell_{\text{indices}}(0, 5) = \{2, 4\}$

# Intervalle successeur

Un  $m$ -intervalle  $[g; d]$  est **inclus** dans un  $\ell$ -intervalle  $[i; j]$  si

- $[g; d]$  est un sous-intervalle de  $[i; j]$  ( $i \leq g < d \leq j$ ) et
- $m > \ell$ .

Le  $\ell$ -intervalle  $[i; j]$  est appelé l'intervalle **englobant** l'intervalle  $[g; d]$ .

Si  $[i; j]$  englobe  $[g; d]$  et qu'il n'y a pas d'intervalle inclus dans  $[i; j]$  qui englobe  $[g; d]$  alors  $[g; d]$  est un **intervalle successeur** de  $[i; j]$ .

# Plan

- 1 Introduction
- 2 Arbre des lcp-intervalles**
- 3 Table des suffixes augmentées

# Arbre des lcp-intervalles

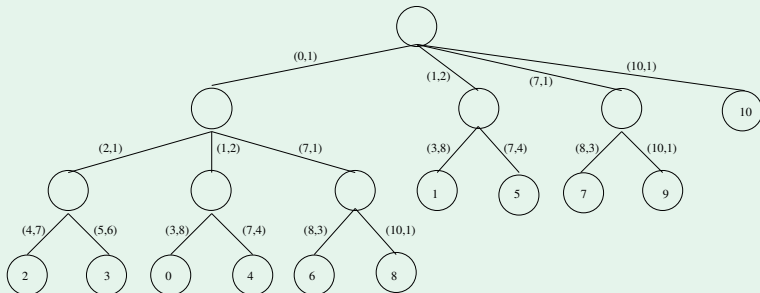
La relation prédécesseur-successeur sur les intervalles constitue l'**arbre des lcp-intervalles** de la table des suffixes.

La racine de l'arbre est le 0-intervalle  $[0; n]$ .

L'arbre des lcp-intervalles est essentiellement l'arbre des suffixes sans les feuilles.

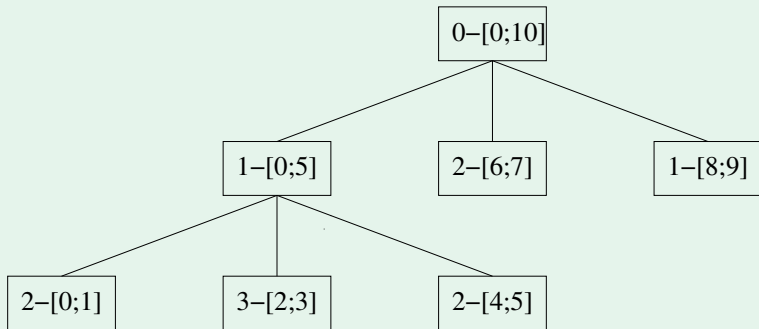
# Arbre des suffixes

## Exemple



# Arbre des lcp-intervalles

## Exemple



# Plan

- 1 Introduction
- 2 Arbre des lcp-intervalles
- 3 Table des suffixes augmentées**

## Tables additionnelles

$$\text{succ}[i].\text{haut} = \min\{q \in [0; i - 1] \mid \text{LCP}[q] > \text{LCP}[i] \text{ et} \\ \forall k \in [q + 1; i - 1], \text{LCP}[k] \geq \text{LCP}[q]\}$$

$$\text{succ}[i].\text{bas} = \max\{q \in [i + 1; n] \mid \text{LCP}[q] > \text{LCP}[i] \text{ et} \\ \forall k \in [i + 1; q - 1], \text{LCP}[k] > \text{LCP}[q]\}$$

$$\text{succ}[i].\text{proc} = \min\{q \in [i + 1; n] \mid \text{LCP}[q] = \text{LCP}[i] \text{ et} \\ \forall k \in [i + 1; q - 1], \text{LCP}[k] > \text{LCP}[i]\}$$

## Tables additionnelles

Pour un  $l$ -intervalle  $[i; j]$  dont les  $l$ -indices sont  $i_1 < i_2 < \dots < i_k$ ,  $\text{succ}[i].\text{haut}$  et  $\text{succ}[i].\text{bas}$  sont utilisés pour déterminer le premier  $l$ -indice  $i_1$ .

Les autres  $l$ -indices sont

$\text{succ}[i_1].\text{proc}, \text{succ}[i_2].\text{proc}, \dots, \text{succ}[i_{k-1}].\text{proc}$ .

## Tables additionnelles

### Lemme 3.1

Soit  $[i; j]$  un  $l$ -intervalle. Si  $i_1 < i_2 < \dots < i_k$  sont les  $l$ -indices de  $[i; j]$  alors les intervalles successeurs de  $[i; j]$  sont  $[i; i_1 - 1], [i_1; i_2 - 1], \dots, [i_k; j]$ .

## Tables additionnelles

### Lemme 3.2

Pour chaque  $\ell$ -intervalle  $[i; j]$  on a :

- ①  $i < \text{succ}[j + 1].\text{haut} \leq j$  ou  $i < \text{succ}[i].\text{bas} \leq j$ .
- ②  $\text{succ}[j + 1].\text{haut}$  est le premier  $\ell$ -indice de  $[i; j]$  si  $i < \text{succ}[j + 1].\text{haut} \leq j$ .
- ③  $\text{succ}[i].\text{bas}$  est le premier  $\ell$ -indice de  $[i; j]$  si  $i < \text{succ}[i].\text{bas} \leq j$ .

# Table des suffixes augmentées

Exemple pour  $y = \text{acaaacatat}\$$

$i$	$p[i]$	$LCP[i]$	$haut$	$bas$	$proc$	
0	2	0		2	6	a a a c a t a t \$
1	3	2				a a c a t a t \$
2	0	1	1	3	4	a c a a a c a t a t \$
3	4	3				a c a t a t \$
4	6	1	3	5		a t a t \$
5	8	2				a t \$
6	1	0	2	7	8	c a a a c a t a t \$
7	5	2				c a t a t \$
8	7	0	7	9	10	t a t \$
9	9	1				t \$
10	10	0	9			\$

# Applications

- Recherche exacte de mot.
- Calcul des répétitions supermaximales et de facteurs communs uniques maximaux (MUM).
- Calcul de répétitions maximales et de facteurs communs maximaux (MEM).
- Calcul de répétitions en tandem.

## Exemple

Exemple pour  $y = \text{acaacatat}\$$ 

$[0; 5]$  est un 1-intervalle et  $\text{lindices}(0, 5) = \{2, 4\}$

Le premier 1-indice 2 est stocké dans  $\text{succ}[0].\text{bas}$  et dans  $\text{succ}[6].\text{haut}$ .

Le second 1-indice 4 est stocké dans  $\text{succ}[2].\text{proc}$ .

Donc les intervalles successeurs de  $[0; 5]$  sont  $[0; 1]$ ,  $[2; 3]$  et  $[4; 5]$ .

## References



M.I. Abouelhoda, S. Kurtz, E. Ohlebusch.  
Replacing suffix trees with enhanced suffix arrays.  
*J. Discrete Algorithms* 2(1) : 53-86 (2004).



M.I. Abouelhoda, E. Ohlebusch, S. Kurtz.  
Optimal Exact String Matching Based on Suffix Arrays.  
*SPIRE 2002* 31-43 (2002).



M.I. Abouelhoda, S. Kurtz, E. Ohlebusch.  
Enhanced suffix arrays and applications.  
In S. Aluru, editor, *Handbook on Computational Molecular Biology*, pp  
7-1-7-27. Chapman and Hall/CRC, 2006.