Table des suffixes augmentées

Thierry Lecroq
Thierry.Lecroq@univ-rouen.fr

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

Plan

Plan

Table des suffixes

Exemple pour $y =$ acaaacatat$

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 ℓ 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 ℓ-intervalle (ou même ℓ-$[i;j]$) pour un lcp-intervalle de valeur lcp ℓ.

ℓindices

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

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

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

Exemple

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

$[0;5]$ 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 $\ell indices(0,5) = \{2,4\}$

Intervalle successeur

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

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

Le ℓ-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

Arbre des lcp-intervalles

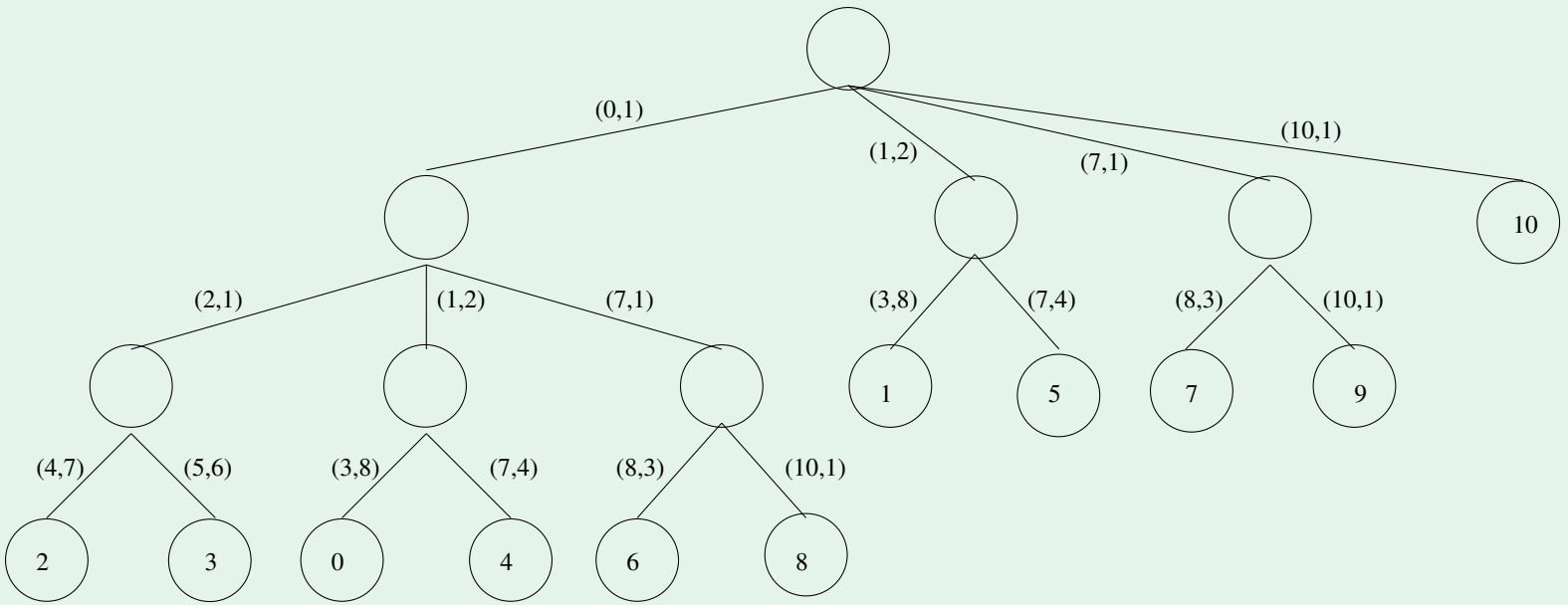
La relation prédécesseur-successeur sur les intervalles constitue l'arbre des lcp-intervalles de la table des suffixes.

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.

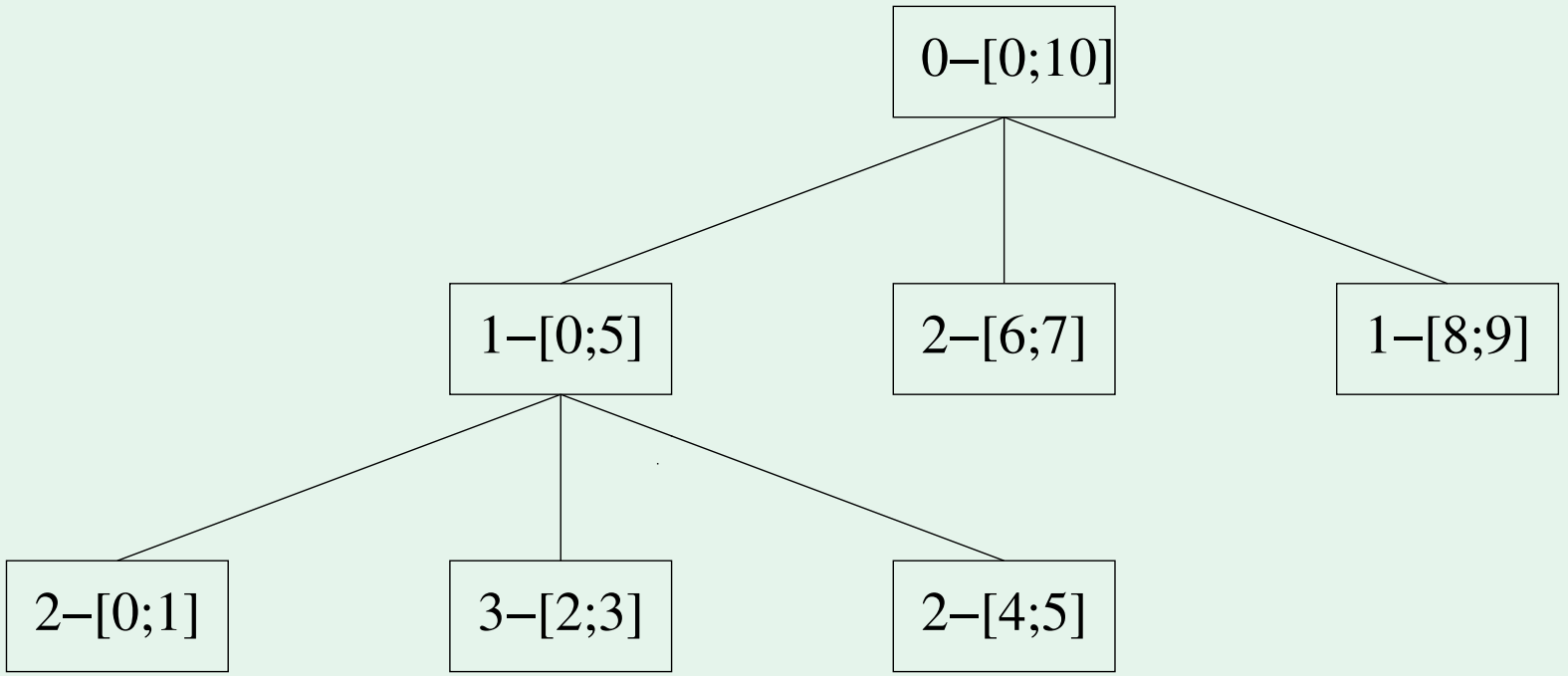
Arbre des suffixes

Exemple



Arbre des lcp-intervalles

Exemple



Plan

Tables additionnelles

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

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

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

Tables additionnelles

Pour un ℓ-intervalle $[i; j]$ dont les ℓ-indices sont $i_1 < i_2 < \cdots < i_k$, *succ*$[i]$.*haut* et *succ*$[i]$.*bas* sont utilisés pour déterminer le premier ℓ-indice i_1.
Les autres ℓ-indices sont
succ$[i_1]$.*proc*, *succ*$[i_2]$.*proc*, $\ldots$, *succ*$[i_{k-1}]$.*proc*.

Tables additionnelles

Lemme 3.1

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

Tables additionnelles

Lemme 3.2

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

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

Table des suffixes augmentées

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

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 $y = \texttt{acaaacatat\$}$

$[0;5]$ est un 1-intervalle et $\ell\mathit{indices}(0,5) = \{2,4\}$
Le premier 1-indice 2 est stocké dans $\mathit{succ}[0].\mathit{bas}$ et dans $\mathit{succ}[6].\mathit{haut}$.
Le second 1-indice 4 est stocké dans $\mathit{succ}[2].\mathit{proc}$.
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 Strring 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.