Combinatoire des mots

Alphabet et mots

alphabet A : ensemble fini non vide de lettres (symboles, caractères)

Exemples

ADN : acides nucléiques, paires de bases, bases

ARN : acides nucléiques

protéines : acides aminés

Définitions et notations

mot : suite finie de lettres

mot vide : suite de zéro lettre notée ε

Exemple

$A = \{$ `a`, `c` , `g`, `t` $\}$

ε, `a`, `t`, `tata` sont des mots sur l'alphabet A

Définitions et notations

A^* : ensemble de tous les mots finis sur l'alphabet A (monoïde libre)

A^+ : ensemble de tous les mots finis non vides sur l'alphabet A

Définitions et notations

La longueur d'un mot $x \in A^*$ est définie comme étant le nombre de lettres du mot x.

Elle est notée $|x|$.

Définitions et notations

On note $x[i]$ pour $i = 0, 1, \ldots, |x|\text{-}1$, la lettre du mot x à l'indice i (la numérotation commence en 0).

Pour $x \neq \varepsilon$, chaque indice $i = 0, 1, \ldots, |x|\text{-}1$ est une position sur x.

$$x = x[0]x[1]\ldots x[|x|\text{-}1]$$

Définition

D'où une définition élémentaire d'identité entre deux mots quelconques x et y :

$$x = y$$

$$\Leftrightarrow$$

$$|x| = |y|$$

et

$$x[i] = y[i] \text{ pour } i = 0, 1, \ldots, |x|\text{-}1$$

Définitions et notations

L'ensemble de lettres sur lequel est formé le mot x est noté $alph(x)$.

Exemple

$$alph(\texttt{tata}) = \{\, \texttt{a}, \texttt{t} \,\}$$

Définitions et notations

Le produit ou concaténation de deux mots x et y est le mot composé des lettres de x suivies de lettres de y.

On le note $x \cdot y$ ou plus simplement xy.

le mot vide ε est l'élément neutre pour la concaténation.

Définitions et notations

Pour un mot $x \in A^*$ et un entier naturel $n \in \mathbf{N}$ on définit la n-ième puissance de x notée x^n par

$x^0 = \varepsilon$ et $x^k = x^{k-1}x$ pour $k = 1, 2, \ldots, n$.

$$x^n = \underbrace{xx\ldots x}_{n \text{ fois}}$$

Exemple

$(\texttt{ta})^4 = \texttt{tatatata}$ et $\texttt{ta}^4 = \texttt{taaaa}$

Définitions et notations

Si $z = xy$ alors

$$x = zy^{-1}$$

et

$$y = x^{-1}z$$

Définitions et notations

Le renversé ou image miroir ou miroir d'un mot $x \in A^*$ est le mot $x^\sim$ (ou x^R) défini par

$$x^\sim = x[|x|-1]x[|x|-2]\ldots x[0]$$

Définitions et notations

Un mot x est un facteur d'un mot y s'il existe deux mots u et v tels que $y = uxv$.

Un mot x est un préfixe d'un mot y s'il existe un mot v tel que $y = xv$.

Un mot x est un suffixe d'un mot y s'il existe un mot u tel que $y = ux$.

Un mot x est un sous-mot d'un mot y s'il existe $|x|+1$ mots $w_0, w_1, \ldots, w_{|x|}$ tels que

$$y = w_0x[0]w_1x[1]w_2\ldots w_{|x|-1}x[|x|-1]w_{|x|}.$$

Exemple

$y =$ `acgat`

`ga` est un facteur de y

`acg` est un préfixe de y

`gat` est un suffixe de y

`ca` est un sous-mot de y

Définitions et notations

Un facteur, un préfixe, un suffixe ou un sous-mot x de y est qualifié de propre si $x \neq y$.

Définitions et notations

On note respectivement $x \preccurlyeq_{fact} y$, $x \prec_{fact} y$, $x \preccurlyeq_{préf} y$, $x \prec_{préf} y$, $x \preccurlyeq_{suff} y$, $x \prec_{suff} y$, $x \preccurlyeq_{smot} y$, $x \prec_{smot} y$ lorsque x est respectivement un facteur, un facteur propre, un préfixe, un préfixe propre, un suffixe, un suffixe propre, un sous-mot, un sous-mot propre de y.

$\preccurlyeq_{fact}$, $\preccurlyeq_{préf}$, $\preccurlyeq_{suff}$, $\preccurlyeq_{smot}$ sont des relations d'ordre.

Définitions et notations

On note respectivement $\mathit{Fact}(x)$, $\mathit{Préf}(x)$, $\mathit{Suff}(x)$ et $\mathit{SMot}(x)$ l'ensemble des facteurs, préfixes, suffixes et sous-mots de x.

Définitions et notations

L'ordre lexicographique, noté $\leq$, est un ordre sur les mots induits par un ordre sur les lettres notés de la même façon.

Pour $x, y \in A^*$

$$x \leq y \Rightarrow \begin{cases} x \preccurlyeq_{préf} y \\ \text{ou} \\ x = uav \text{ et } y = ubw \text{ avec } u,v,w \in A^*,\ a,b \in A \text{ et } a < b \end{cases}$$

Exemple

$A = \{\, \texttt{a}, \texttt{c}, \texttt{g}, \texttt{t} \,\}$ $\quad$ $\texttt{a} < \texttt{c} < \texttt{g} < \texttt{t}$

$\texttt{acgat} < \texttt{agact} < \texttt{agcat}$

Définitions et notations

Si x est un facteur de y on dit que x apparaît dans y ou qu'il y a une occurrence de x dans y.

Toute occurrence (de x dans y) peut être caractérisée par une position sur y.

Définitions et notations

La notation entre crochets définie sur les lettres est étendue au facteur. On définit le facteur de x de la position i à la position j par

$$x[i..j] = x[i]x[i+1]\ldots x[j]$$

pour $0 \leq i \leq j \leq |x|-1$.

$$x[i..j] = \varepsilon \text{ si } i > j.$$

Lemme de Lévi

Lemme de Lévi, 1944 (extrait)

Pour tout mot $x, y, z, t \in A^*$

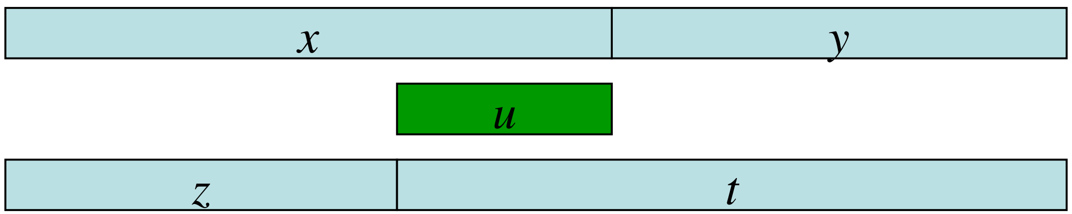
$xy = zt$ implique qu'il existe un mot $u \in A^*$ tel que

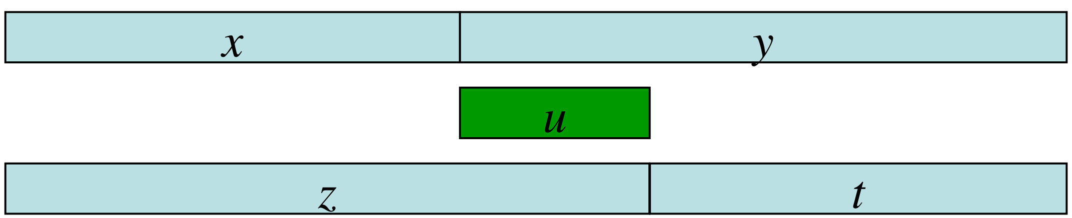
- soit $x = zu$ et $t = uy$;
- soit $z = xu$ et $y = ut$.

Lemme de Lévi

Preuve

graphique

- 

- 

Lemme de Lévi

Conséquence : on peut se fier à ce que l'on voit !

Les mots de Fibonacci

Les nombres de Fibonacci sont définis par

$F_0 = 0$, $F_1 = 1$ et $F_n = F_{n\text{-}1} + F_{n\text{-}2}$ pour $n \geq 2$.

Les mots de Fibonacci sont définis par

$f_0 = \varepsilon$, $f_1 = b$, $f_2 = a$ et $f_n = f_{n\text{-}1}f_{n\text{-}2}$ pour $n \geq 3$.

On a $|f_n| = F_n$.

Périodicités et bords

Soit x un mot non vide.

Un entier p tel que $0 < p \leq |x|$ est une période de x si

$x[i] = x[i+p]$ pour $0 \leq i \leq |x|-p-1$.

La période d'un mot x non vide est la plus petite de ses périodes.

Elle est notée $pér(x)$.

Exemple

$x = \texttt{aataataa}$

3, 6, 7 et 8 sont des périodes de x.

$pér(x) = 3$

Proposition 1

Soient x un mot non vide et p un entier tel que $0 < p \leq |x|$. Alors les cinq propriétés suivantes sont équivalentes :

- L'entier p est une période de x.
- Il existe deux mots uniques $u,v \in A^*$ et un entier $k > 0$ tels que $x = (uv)^k u$ et $|uv| = p$.
- Il existe un mot t et un entier $m > 0$ tels que $x \preccurlyeq_{préf} t^m$ et $|t| = p$.
- Il existe trois mots u', v' et w' tel que $x = u'w' = w'v'$ et $|u'| = |v'| = p$.
- Il existe un mot t' tel que $x \preccurlyeq_{préf} t'x$ et $|t'| = p$.

Proposition 1

Preuve

- $1 \Rightarrow 2$: si $v \neq \varepsilon$ et $k > 0$ alors k est le dividende de la division entière de $|x|$ par p.

Si (u',v',k') satisfait les mêmes conditions que (u,v,k), on a $k' = k$ et donc $|u'| = |u|$ d'où $u' = u$ et $v' = v$. Ce qui montre l'unicité de la décomposition si elle existe.

Soient k et r le dividende et le reste de la division euclidienne de $|x|$ par p.

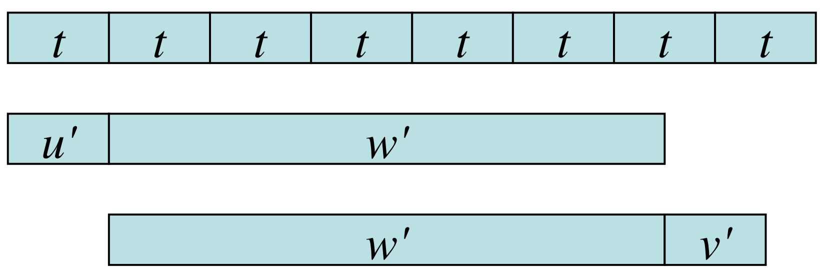
Alors si $u = x[0..r\text{-}1]$ et $v = x[r..p\text{-}1]$ on a $x = (uv)^k u$ et $|uv| = p$.

Proposition 1

$2 \Rightarrow 3$: $t = uv$ et $m > |x|/p$.

$3 \Rightarrow 4$: $w' = t^{-1}x$ donc $u' = t$.

Comme $x \preccurlyeq_{préf} t^k$, w' est aussi préfixe de x :



Donc il existe bien trois mots u',v' et w' tels que $x = u'w' = w'v'$ et $|u'| = |t| = p$.

Proposition 1

$4 \Rightarrow 5$: puisque $u'w' \preccurlyeq_{préf} u'w'v'$ on a $x \preccurlyeq_{préf} tx$ avec $|t| = p$ en posant $t = u'$.

$5 \Rightarrow 1$: soit i un entier tel que $0 \leq i \leq |x|$-p-1

$x[i+p] = (t'x)[i+p]$ car $x \preccurlyeq_{préf} t'x$

$= x[i]$ car $|t'| = p$

ce qui montre que p est une période de x.

$\square$

Définitions et notations

Un mot w est un bord d'un mot x s'il est à la fois préfixe et suffixe propre de x.

Autrement dit, il existe deux mots $u, v \in A^*$ tels que $x = uw = wv$.

Exemple

$x =$ `aataataa`

ε, `a`, `aa` et `aataa` sont des bords de x

Définitions et notations

Le bord de x est le plus long de ses bords.

Il est noté $Bord(x)$.

Exemple

$Bord(\texttt{aataataa}) = \texttt{aataa}$

Les notions de bords et de périodes sont duales comme le montre la propriété 4 de la proposition 1.

Proposition 2

Soient x un mot non vide et n le plus grand des entiers k pour lequel $Bord^k(x)$ est défini (soit $Bord^n(x) = \varepsilon$).

Alors

$B = (Bord(x), Bord^2(x), \ldots, Bord^n(x))$

est la suite de tous les bords de x classés par ordre décroissant de longueur, et

$P = (|x|-|Bord(x)|, |x|-|Bord^2(x)|, \ldots, |x|-|Bord^n(x)|)$

est la suite des périodes de x classées en ordre croissant.

□

Lemme de périodicité

Lemme 3 (Lemme de périodicité, Fine et Wilf 1965)

Si p et q sont des périodes d'un mot non vide x telles que

$$p+q-\mathrm{pgcd}(p,q) \leq |x|$$

alors $\mathrm{pgcd}(p,q)$ est aussi une période de x.

Primitivité

Définition

Un mot non vide est primitif s'il n'est puissance d'aucun autre mot que lui-même :

$x \in A^+$ est primitif $\Leftrightarrow$ si $\exists\, u \in \mathrm{A}^*$ et $n \in \mathbf{N}$ tel que

$$x = u^n \Rightarrow u = x \text{ et } n = 1$$

Exemple

`tatta` est primitif

ε, `tatata` = (`ta`)^3 ne sont pas primitifs

Lemme de primitivité

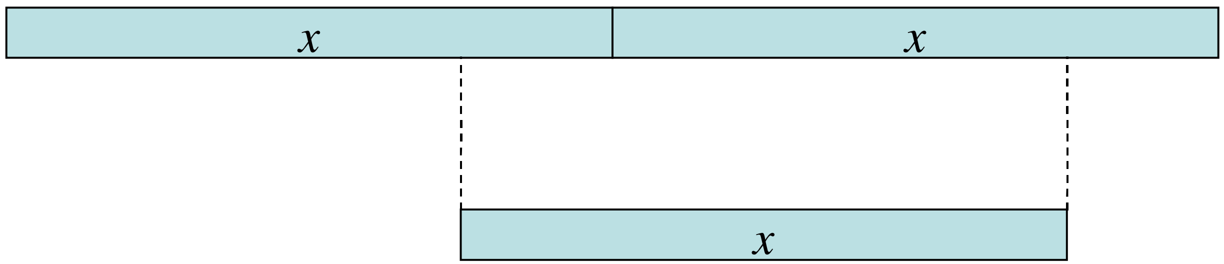
Lemme 4 (lemme de primitivité)

Un mot non vide est primitif si et seulement si il n'est facteur de son carré qu'en tant que préfixe et suffixe. Autrement dit, pour tout mot non vide x :

$$x \text{ est primitif} \Leftrightarrow yx \preccurlyeq_{préf} x^2 \Rightarrow y = \varepsilon \text{ ou } y = x.$$

Lemme de primitivité

Preuve



impossible

Définitions et notations

Si x est un mot non vide, on dit du mot primitif z dont x est la puissance qu'il est la racine de x, et du naturel n tel que $x = z^n$ qu'il est l'exposant de x.

Conjugaison

Deux mots x et y sont conjugués s'il existe deux mots u et v tels que $x = uv$ et $y = vu$.

Exemple

`taata` et `atata` sont conjugués

$u =$ `ta`

$v =$ `ata`