

MMIM Modèles mathématiques en informatique musicale

Marc Chemillier

Master M2 Atiam (Ircam), 2011-2012

Oracle des facteurs (logiciels d'improvisation OMax, Improtek)

- Construction de l'oracle
 - o Liens suffixiels
 - o Caractérisation des flèches additionnelles
- Langage reconnu par l'oracle

1. Construction de l'oracle

1.1 Liens suffixiels

L'ensemble des facteurs d'un mot x est fini, donc reconnaissable par automate. On obtient facilement un AFN reconnaissant tous les facteurs en prenant l'écorché de x et en rendant tous les états initiaux et terminaux.

Problème : **il n'existe pas de construction simple de l'AFD correspondant** (contrairement à l'automate des mots se terminant par x).

L'« oracle des facteurs » a été introduit pour pallier cette difficulté.

L'*oracle des facteurs* est un automate qui reconnaît tous les facteurs d'un mot x , mais avec quelques mots en plus. On peut donner directement l'AFD correspondant (sans détermination) par une technique analogue à l'algorithme de Morris & Pratt.

Exemple : écorché de la séquence $x = abcadbcd$



Construction de l'oracle : on part de l'écorché de x , et on ajoute des transitions à l'aide d'une fonction f dite de « lien suffixiel » entre états, en construisant simultanément f et les transitions.

Au départ, on place un lien suffixiel de 1 vers 0. Puis on suppose l'oracle construit avec ses liens suffixiels jusqu'à l'état p compris. La lettre suivante a donne un nouvel état $\delta(p, a) = p+1$.

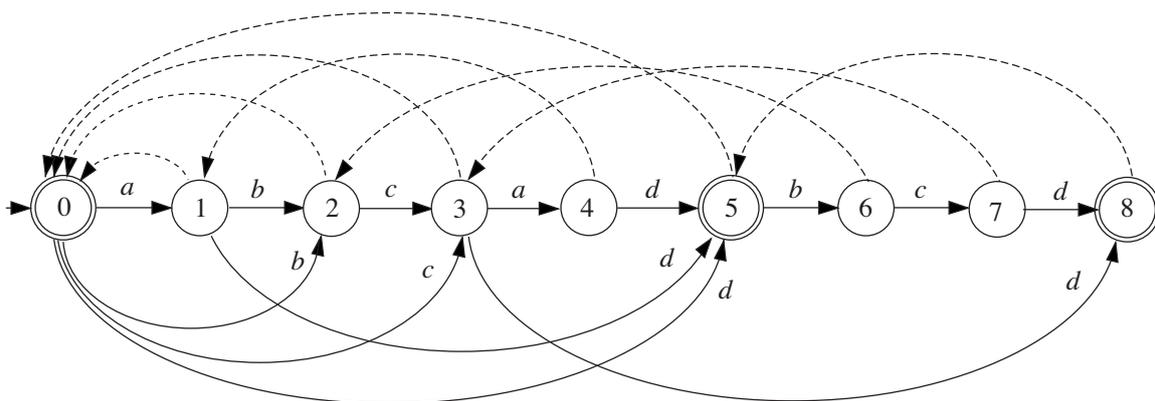
Pour ajouter les transitions, on suit les liens suffixiels déjà existants $p' = f(p)$, puis $p' = f(f(p))$, etc.

- si $\delta(p', a)$ est non défini, on ajoute une transition $\delta(p', a) = p+1$
 - si $p' \neq 0$, on continue à suivre les liens,
 - si $p' = 0$, on stoppe et on crée un lien suffixiel en posant $f(p+1) = 0$
- si $\delta(p', a)$ est défini, on stoppe (pas de nouvelle transition), et on crée un lien suffixiel en posant $f(p+1) = \delta(p', a)$ (= état d'arrivée de la transition existante)

Remarque : toutes les transitions arrivant dans un même état ont la même lettre.

Pour la séquence $abcdabcd$, l'oracle des suffixes donne l'AFD suivant (les liens suffixiels sont indiqués par des flèches en pointillé au-dessus).

- Oracle des facteurs : tous les états sont terminaux
- Oracle des suffixes : les états terminaux sont ceux du chemin suffixiel partant du dernier état, soit 8 (suffixe cd), 5 (suffixe d), 0 (suffixe vide)



Dans le logiciel d'improvisation OMax, on parcourt l'oracle en suivant les transitions, mais aussi en empruntant éventuellement les liens suffixiels :

d b c

1^{er} motif prélevé = $a b c a \underline{d b c} d$

5 6 7

On emprunte le lien suffixiel de 7 à 3 pour reprendre la lecture vers 4 :

$d \ b \ c \ \boxed{a \ d \ b}$
(3) 4 5 6

2^e motif prélevé = $a \ b \ c \ \underline{a \ d \ b} \ c \ d$

Le 1^{er} motif dbc se terminait par bc . Or le 2^{ème} motif adb est **précédé** par bc dans la séquence.

1^{er} motif prélevé = $a \ b \ c \ a \ \underline{d \ b \ c} \ d$

2^{ème} motif prélevé = $a \ (b \ c) \ \underline{a \ d \ b} \ c \ d$

Dans la succession des 2 motifs $dbc + adb$, le 2^{ème} prélèvement devient $(bc)adb$:

$d \ \boxed{b \ c \ a \ d \ b}$
(2 3) 4 5 6

2^e motif **réellement** prélevé = $a \ (\underline{b \ c}) \ \underline{a \ d \ b} \ c \ d$

Ainsi les deux motifs réellement prélevés se **chevauchent** : $d(bc)$ et $(bc)adb$.

C'est une propriété fondamentale de l'oracle :

Les dernières lettres lues avant de quitter un état par un lien suffixiel sont identiques à celles qui précèdent l'état d'arrivée du lien (ex : bc dans le parcours ci-dessus).

Ces lettres constituent une partie commune entre les motifs lus avant et après le lien suffixiel. Ainsi les liens suffixiels permettent d'enchaîner des motifs qui se chevauchent avec une partie commune (ex : $d(bc)$ et $(bc)adb$ dans le parcours ci-dessus).

Propriété fondamentale. *Si p est l'état d'arrivée d'un préfixe w de x , le lien suffixiel $f(p)$ correspond à l'état d'arrivée du plus long suffixe de w qui est répété à gauche, c'est-à-dire qui est facteur non suffixe de w .*

Attention : Il ne faut pas confondre

- fonction de lien suffixiel :

$f(p)$ = état d'arrivée du plus long suffixe propre de w qui est aussi **facteur** de w (donc de x)

- fonction de saut de Morris & Pratt :

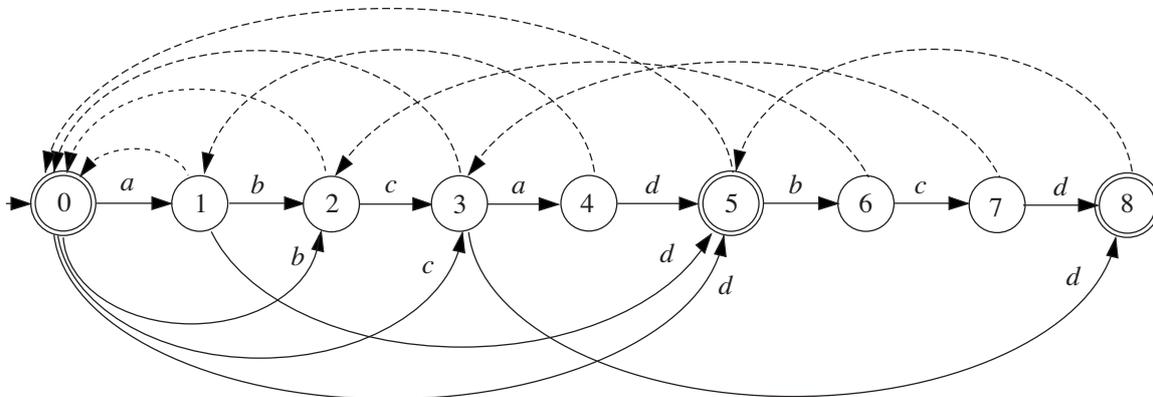
$f(p)$ = état d'arrivée du plus long suffixe propre de w qui est aussi **préfixe** de w (donc de x)

1.2 Caractérisation des flèches additionnelles

Les flèches additionnelles de l'oracle (non présentes dans l'écorché) vérifient la caractérisation suivante. On note :

$\min(p)$ = ensemble des mots de **longueur minimale** reconnus de 0 à p

Propriété caractéristique. Pour tout mot u de $\text{min}(p)$ et pour toute lettre $b \neq x_{p+1}$, il existe une flèche additionnelle $\delta(p, b) = q$ si et seulement si $q = \text{état d'arrivée de la première occurrence de } ub \text{ dans } x \text{ telle que } q > p + 1$.



état 0 : $\text{min}(0) = \{\varepsilon\}$, $x_1 = a$, flèches additionnelles : $\delta(0, b) = 2$, $\delta(0, c) = 3$, $\delta(0, d) = 5$

état 1 : $\text{min}(1) = \{a\}$, $x_2 = b$, flèche additionnelle : $\delta(1, d) = 5$ (1^{ère} occurrence de ad)

état 3 : $\text{min}(3) = \{c\}$, $x_4 = a$, flèche additionnelle : $\delta(3, d) = 8$ (1^{ère} occurrence de cd)

Remarque : Dans l'article de Allauzen, Crochemore & Raffinot 1999, l'oracle est défini par cette propriété caractéristique. La construction par liens suffixiels n'arrive qu'après une série de lemmes. Mais comme elle est plus efficace et donne le même automate, c'est elle qui est utilisée.

Lemme. Si w est facteur de x , alors w est reconnu dans l'oracle de 0 à $p \leq \text{état d'arrivée de sa première occurrence dans } x$.

Le Lemme montre que l'oracle reconnaît tous les facteurs de x .

En fait, l'oracle reconnaît **un peu plus** que les facteurs de x .

2. Langage reconnu par l'oracle

2.1 Facteurs canoniques et contractions

L'article Mancheron & Moan 2005 propose une caractérisation du langage reconnu par l'oracle (p. 142-143 de l'article).

Ensemble des facteurs canoniques d'un mot s par rapport à son oracle :

$F_s = \{\min(p) / \text{l'état } p \text{ a au moins deux flèches entrantes, ou deux flèches sortantes}\}$

Une contraction de s est un couple (p, q) où p et q sont les états d'arrivée de la dernière lettre de deux occurrences d'un facteur canonique

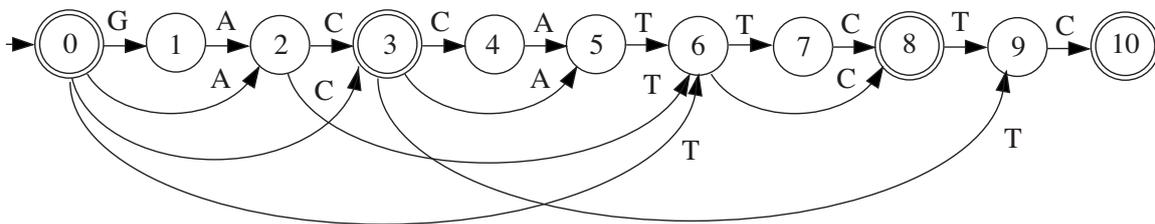
Appliquer une contraction (p, q) à s : on supprime les lettres de s comprises entre p inclus et q non inclus

Fermeture de s :

$\epsilon(s)$ = ensemble des mots obtenus en appliquant à s toutes les combinaisons de contractions possibles

Théorème (3.1, p. 150). *Le langage reconnu par l'oracle d'un mot s est l'ensemble des suffixes de tous les mots de sa fermeture $\epsilon(s)$.*

Exemple : oracle du mot $s = \text{GACCATTCTC}$



Etats = {0,	2,	3,	5,	6,	8,	9}
$F_s =$	{A,	C,	CA,	T,	TC,	CT}
Contractions	(2, 5)	(3, 4)		(6, 7)	(7, 9)	
		(3, 8)		(6, 9)		
		(3, 10)				

Noter qu'on limite l'énumération en fixant le premier état p = celui de la première occurrence du facteur.

Exemple d'application de la combinaison de contractions (2, 5) et (7, 9) :

	1	2	3	4	5	6	7	8	9	10
	G	A	C	C	A	T	T	C	T	C
(2, 5)	G	-	-	-	A	T	T	C	T	C
(7, 9)	G	-	-	-	A	T	-	-	T	C

d'où le mot GATTC

Petite remarque : pourquoi les lettres G, C, T, A ?

Le génome est fait d'ADN. Les gènes contenus dans le génome sont codés sous forme chimique le long des molécules d'ADN. Celles-ci sont constituées par l'enchaînement de "maillons" élémentaires nommés nucléotides. Les nucléotides ont une partie variable - une base, du point de vue chimique - qui peut exister sous 4 formes différentes ; ces formes sont symbolisées par les lettres A, T, G et C. Les instructions sont donc écrites dans un alphabet chimique à 4 lettres seulement.

Remarque : Dans les applications bioinformatiques, l'oracle est utilisé de façon négative. Si un mot n'est pas reconnu par l'oracle, on est sûr qu'il n'est pas suffixe.

2.2 Exemple de langage reconnu par l'oracle

Exemple : oracle du mot $s = \text{GCTCA}$

On peut vérifier que les suffixes de GCTCA sont reconnus (outre le mot vide) :

GCTCA

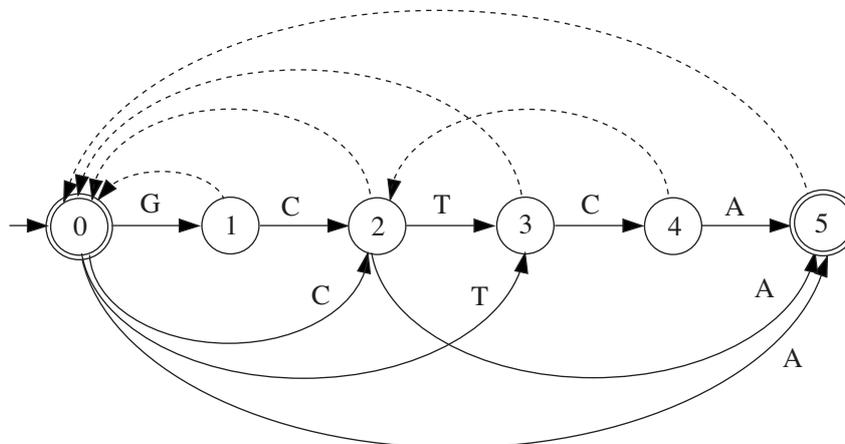
CTCA

TCA

CA

A

Existe-t-il d'autres mots reconnus par l'oracle qui ne sont pas suffixes de GCTCA ?



Oui, il y en a un : GCA

Etats = {0, 2, 3, 5}

F_s = {C, T, A}

Contractions (2, 4)

Si l'on applique la contraction (2, 4), on obtient précisément ce mot GCA.

La fermeture $\varepsilon(s)$ contient les suffixes de $s = \text{GCTCA}$ et de GCA. Comme les suffixes de GCA sont aussi suffixes de s excepté GCA, cela donne le langage reconnu par l'oracle :

$\varepsilon(s) = \{\text{GCTCA}, \text{CTCA}, \text{TCA}, \text{CA}, \text{A}, \varepsilon, \text{GCA}\}$

Références

- oracle des suffixes (simulation stylistique)

Allauzen, Cyril & Maxime Crochemore, Mathieu Raffinot, Factor oracle : A new structure for pattern matching, *SOFSEM '99: Proceedings of the 26th Conference on Current Trends in Theory and Practice of Informatics*, Lecture Notes in Computer Science, Springer-Verlag, 1999, p. 291-306 ([biblio](#)).

Mancheron Alban, Moan Christophe, Combinatorial Characterization of the Language Recognized by Factor and Suffix Oracles, *International Journal of Foundations of Computer Science*, 16 (6) (2005) 1179-1191 ([biblio](#)).

Assayag, Gérard, Shlomo Dubnov, Olivier Delerue, Guessing the Composer's Mind: Applying Universal Prediction to Musical Style, *Proceedings of the ICMC (Int. Computer Music Conf.)*, 1999, p. 496-499 ([biblio](#)).

Assayag, Gérard, Shlomo Dubnov, Using factor oracles for machine improvisation, *Soft Computing*, special issue on Formal Systems and Music, G. Assayag, V. Cafagna, M. Chemillier (eds.), 8 (9) (2004) 604–610 ([biblio](#)).