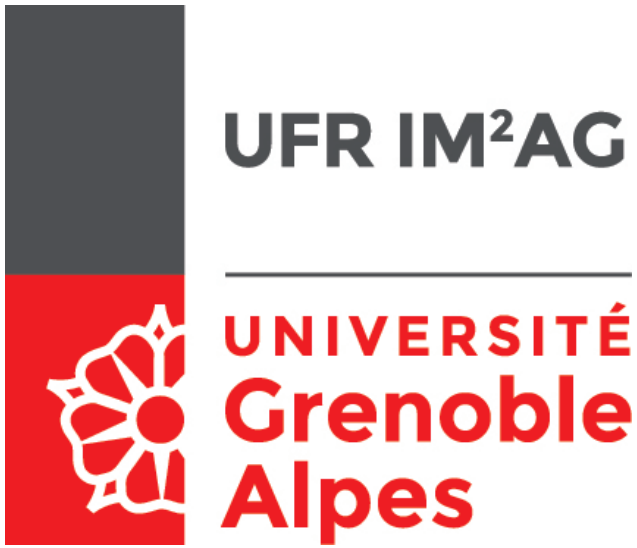
Chaînes de caractères et recherche de motifs

{Jean-Marc.Vincent,Benjamin.Wack,Vincent.Danjean}@imag.fr

DU Informatique et Sciences du Numérique : Information



5 juillet 2019

RECHERCHE DE MOTIF

UFR IM²AG
UNIVERSITÉ Grenoble Alpes

LE PROBLÈME

On recherche un motif dans un texte, mais en général le motif et le texte peuvent être :

- des suites de caractères (ASCII ou pas)
- des suites d'octets dans un fichier quelconque
- des séquences ADN (suites de bases ATGC)
- ...

LE PROBLÈME

On recherche un motif dans un texte, mais en général le motif et le texte peuvent être :

- des suites de caractères (ASCII ou pas)
- des suites d'octets dans un fichier quelconque
- des séquences ADN (suites de bases ATGC)
- ...

Formalisation

Plus généralement, on considère :

- un alphabet A fini
- deux suites ordonnées d'éléments de A (le "texte" et le "motif")

et on cherche à répondre à différentes questions plus ou moins équivalentes :

- le motif apparaît-il dans le texte ?
- si oui, déterminer la première position où il apparaît ;
- déterminer toutes les positions où ce motif apparaît.

STRUCTURES DE DONNÉES ET NOTATIONS

On considère que le texte et le motif sont stockés dans des tableaux, ce qui permet de se déplacer arbitrairement dans l'un ou dans l'autre.
(Concrètement ce n'est pas forcément le cas : flot/stream...)

On notera :

- T le tableau contenant le texte et n sa longueur.
- M le tableau contenant le motif et k sa longueur.

(indexés à partir de 0)

STRUCTURES DE DONNÉES ET NOTATIONS

On considère que le texte et le motif sont stockés dans des tableaux, ce qui permet de se déplacer arbitrairement dans l'un ou dans l'autre.
(Concrètement ce n'est pas forcément le cas : flot/stream...)

On notera :

- T le tableau contenant le texte et n sa longueur.
- M le tableau contenant le motif et k sa longueur.

(indexés à partir de 0)

Par exemple, soit l'alphabet $A = \{a, b, c\}$.
Si on recherche le motif *abaa* dans *aacabacabaabaaa* :

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
a	*a*	*c*	*a*	*b*	*a*	*c*	*a*	*b*	*a*	*a*	*b*	*a*	*a*	*a*

M présent aux positions 7 et 10 de T. Ces deux occurrences se chevauchent : la lettre $a = T[10]$ est commune.

RECHERCHE DE MOTIF

UFR IM²AG
UNIVERSITÉ Grenoble Alpes

UN ALGORITHME NAÏF

a	*a*	*c*	*a*	*b*	*a*	*c*	*a*	*b*	*a*	*a*	*b*	*a*	*a*	*a*
a	*b*	*a*	*a*											

UN ALGORITHME NAÏF

a	*a*	*c*	*a*	*b*	*a*	*c*	*a*	*b*	*a*	*a*	*b*	*a*	*a*	*a*
	a	*b*	*a*	*a*										

UN ALGORITHME NAÏF

a	*a*	*c*	*a*	*b*	*a*	*c*	*a*	*b*	*a*	*a*	*b*	*a*	*a*	*a*
		a	*b*	*a*	*a*									

UN ALGORITHME NAÏF

a	*a*	*c*	*a*	*b*	*a*	*c*	*a*	*b*	*a*	*a*	*b*	*a*	*a*	*a*
			a	*b*	*a*	*a*								

UN ALGORITHME NAÏF

a	*a*	*c*	*a*	*b*	*a*	*c*	*a*	*b*	*a*	*a*	*b*	*a*	*a*	*a*
							a	*b*	*a*	*a*				

UN ALGORITHME NAÏF

a	*a*	*c*	*a*	*b*	*a*	*c*	*a*	*b*	*a*	*a*	*b*	*a*	*a*	*a*
							a	*b*	*a*	*a*				

```
i := 0 ; j := 0
tant que j<k && i ⩽ n-k
    j := 0
    tant que j<k && T[i+j] = M[j]
        j := j + 1
    i := i + 1
return i -1     // "motif absent" si j < k
```

UN ALGORITHME NAÏF

a	a	c	a	b	a	c	a	b	a	a	b	a	a	a
							a	b	a	a				

```
i := 0 ; j := 0
tant que j<k && i ⩽ n-k
    j := 0
    tant que j<k && T[i+j] = M[j]
        j := j + 1
    i := i + 1
return i -1      // "motif absent" si j < k
```

Complexité

- Au pire k boucles internes et $n - k$ boucles externes
  D'où $\mathcal{O}(k \times (n - k))$, soit $\mathcal{O}(k \times n)$ si $k \ll n$

UN ALGORITHME NAÏF

a	a	c	a	b	a	c	a	b	a	a	b	a	a	a
							a	b	a	a				

```
i := 0 ; j := 0
tant que j<k && i ⩽ n-k
    j := 0
    tant que j<k && T[i+j] = M[j]
        j := j + 1
    i := i + 1
return i -1      // "motif absent" si j < k
```

Complexité

- Au pire k boucles internes et $n - k$ boucles externes
  D'où $\mathcal{O}(k \times (n - k))$, soit $\mathcal{O}(k \times n)$ si $k \ll n$
- Pire cas atteint pour $M = bbba$ et $T = bbbbbbbbbbbb$

RECHERCHE DE MOTIF

UFR IM²AG
UNIVERSITÉ Grenoble Alpes

DES ÉCONOMIES POSSIBLES

Tests redondants

Lorsqu'un préfixe de M est présent à une position i de T :

- on examine **déjà** les caractères $T[i+1], T[i+2], \ldots$ avant d'échouer
- mais on ne conserve pas cette information
- donc on les examine à nouveau pour tester la présence de M à la position $i+1$

a	a	c	a	b	a	c	a	b	a	a	b	a	a	a
			a	b	a	a								

DES ÉCONOMIES POSSIBLES

Tests redondants

Lorsqu'un préfixe de M est présent à une position i de T :

- on examine **déjà** les caractères $T[i+1], T[i+2], \ldots$ avant d'échouer
- mais on ne conserve pas cette information
- donc on les examine à nouveau pour tester la présence de M à la position $i+1$

a	a	c	a	b	a	c	a	b	a	a	b	a	a	a
			a	b	a	a								

Qu'a-t-on appris ici ?

DES ÉCONOMIES POSSIBLES

Tests redondants

Lorsqu'un préfixe de M est présent à une position i de T :

- on examine **déjà** les caractères $T[i+1], T[i+2], \ldots$ avant d'échouer
- mais on ne conserve pas cette information
- donc on les examine à nouveau pour tester la présence de M à la position $i+1$

a	a	c	a	b	a	c	a	b	a	a	b	a	a	a
			a	b	a	a								

Qu'a-t-on appris ici ?

- $T[4] = b$ donc inutile de chercher *abaa* pour $i = 4$;

DES ÉCONOMIES POSSIBLES

Tests redondants

Lorsqu'un préfixe de M est présent à une position i de T :

- on examine **déjà** les caractères $T[i+1], T[i+2], \ldots$ avant d'échouer
- mais on ne conserve pas cette information
- donc on les examine à nouveau pour tester la présence de M à la position $i+1$

a	a	c	a	b	a	c	a	b	a	a	b	a	a	a
			a	b	a	a								

Qu'a-t-on appris ici ?

- $T[4] = b$ donc inutile de chercher *abaa* pour $i = 4$;
- $T[5] = a$ donc on pourrait chercher *baa* à partir de $i = 6$;

DES ÉCONOMIES POSSIBLES

Tests redondants

Lorsqu'un préfixe de M est présent à une position i de T :

- on examine **déjà** les caractères $T[i+1], T[i+2], \ldots$ avant d'échouer
- mais on ne conserve pas cette information
- donc on les examine à nouveau pour tester la présence de M à la position $i+1$

a	a	c	a	b	a	c	a	b	a	a	b	a	a	a
			a	b	a	a								

Qu'a-t-on appris ici ?

- $T[4] = b$ donc inutile de chercher *abaa* pour $i = 4$;
- $T[5] = a$ donc on pourrait chercher *baa* à partir de $i = 6$;
- mais $T[6] \neq b$ et donc inutile de continuer la recherche à cet endroit.

L'IDÉE

Ce qu'il faut faire n'est pas toujours si clair

Si on recherche **maman** dans un texte commençant par **mamam...**, quelles sont les possibilités encore viables après l'échec ?

L'IDÉE

Ce qu'il faut faire n'est pas toujours si clair

Si on recherche **maman** dans un texte commençant par **mamam...**, quelles sont les possibilités encore viables après l'échec ?

On prend un parti pris :

- un caractère de T ne sera jamais lu plusieurs fois ;
- en cas d'échec on revient en arrière dans M autant que possible.

L'IDÉE

Ce qu'il faut faire n'est pas toujours si clair

Si on recherche **maman** dans un texte commençant par **mamam...**, quelles sont les possibilités encore viables après l'échec ?

On prend un parti pris :

- un caractère de T ne sera jamais lu plusieurs fois ;
- en cas d'échec on revient en arrière dans M autant que possible.

Pour cela, il ne suffit pas de connaître M sous forme d'un tableau :

- il faut analyser finement la structure de ce motif

L'IDÉE

Ce qu'il faut faire n'est pas toujours si clair

Si on recherche **maman** dans un texte commençant par **mamam...**, quelles sont les possibilités encore viables après l'échec ?

On prend un parti pris :

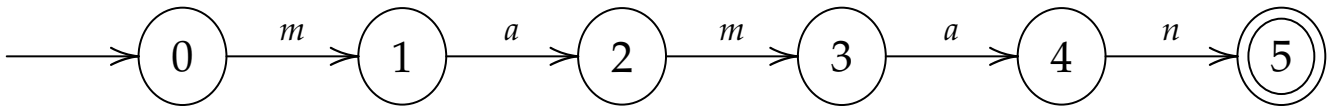
- un caractère de T ne sera jamais lu plusieurs fois ;
- en cas d'échec on revient en arrière dans M autant que possible.

Pour cela, il ne suffit pas de connaître M sous forme d'un tableau :

- il faut analyser finement la structure de ce motif
- cela ne dépend pas de T
- on peut donc le faire **une fois pour toutes** en amont de l'algorithme : phase de pré-traitement
- si possible peu coûteux, et en tous cas qui accélère la suite

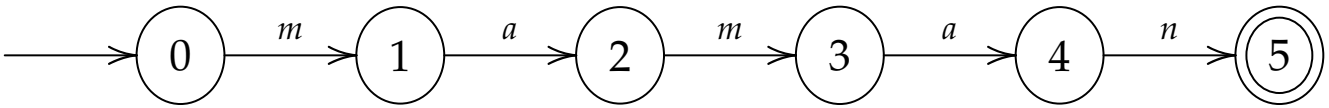
REPRÉSENTATION DE M COMME UN AUTOMATE

On cherche à mémoriser **toutes** les positions de M correspondant à la dernière lettre lue dans T.
Pour cela on construit l'automate qui reconnaît M :

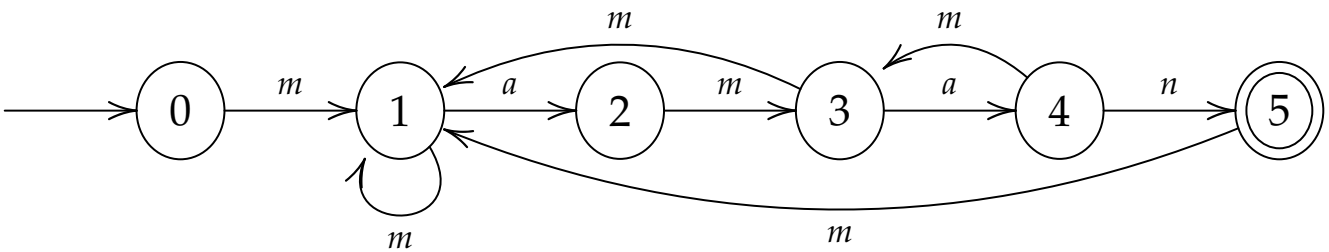


REPRÉSENTATION DE M COMME UN AUTOMATE

On cherche à mémoriser **toutes** les positions de M correspondant à la dernière lettre lue dans T.
Pour cela on construit l'automate qui reconnaît M :



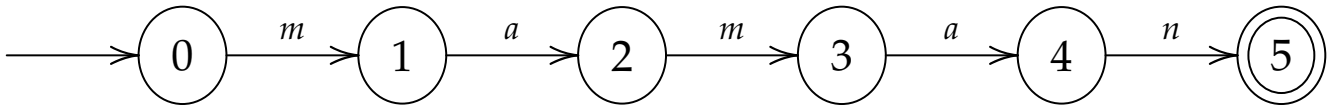
Puis on l'enrichit avec des retours vers les positions antérieures :



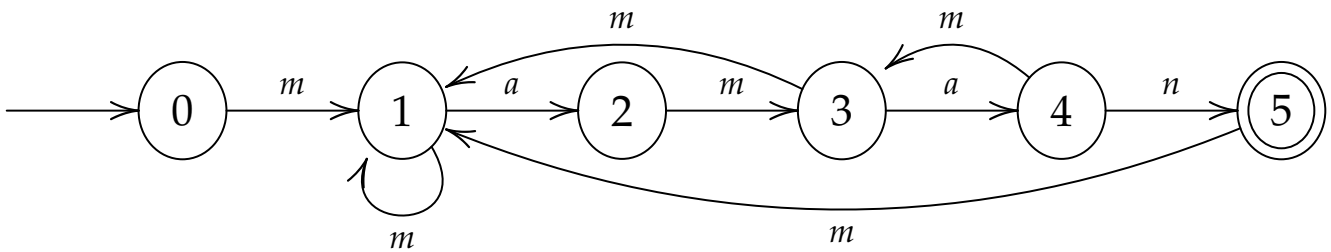
$j \xrightarrow{a}$ le plus long préfixe de M égal à un suffixe de $M[1..j]$

REPRÉSENTATION DE M COMME UN AUTOMATE

On cherche à mémoriser **toutes** les positions de M correspondant à la dernière lettre lue dans T.
Pour cela on construit l'automate qui reconnaît M :



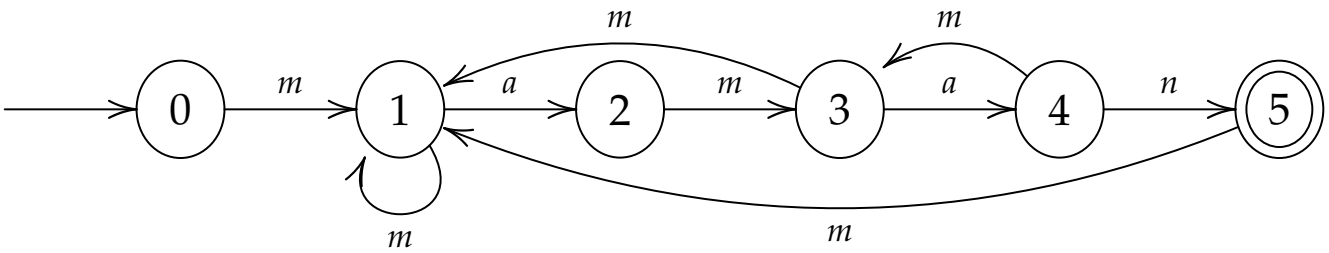
Puis on l'enrichit avec des retours vers les positions antérieures :



$j \xrightarrow{a}$ le plus long préfixe de M égal à un suffixe de $M[1..j]$

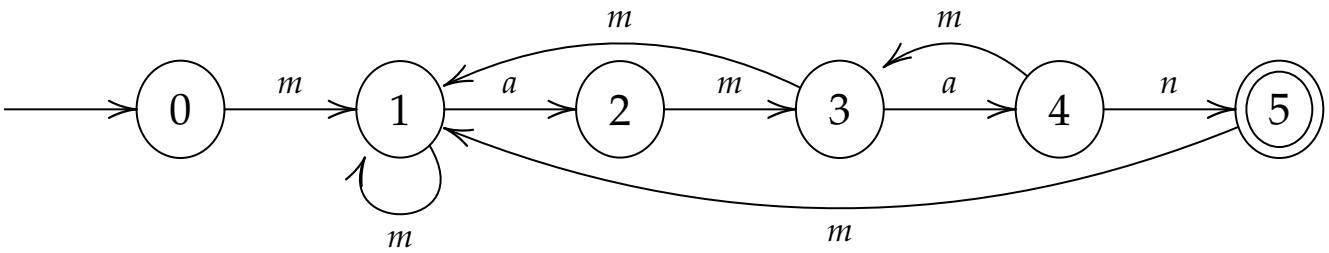
Implicitement, toutes les transitions manquantes reviennent à 0.

REPRÉSENTATION DE M COMME UN AUTOMATE : UTILISATION



Il suffit désormais de suivre les transitions de l'automate pour le texte T, l'état final marque la présence du motif.

REPRÉSENTATION DE M COMME UN AUTOMATE : UTILISATION

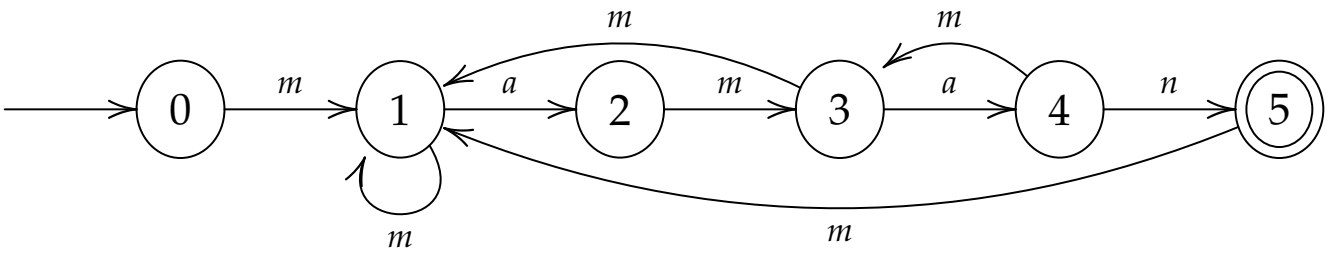


Il suffit désormais de suivre les transitions de l'automate pour le texte T, l'état final marque la présence du motif.

Complexité

Chaque caractère de T est lu une seule fois : recherche du motif en $O(n)$

REPRÉSENTATION DE M COMME UN AUTOMATE : UTILISATION



Il suffit désormais de suivre les transitions de l'automate pour le texte T, l'état final marque la présence du motif.

Complexité

Chaque caractère de T est lu une seule fois : recherche du motif en $O(n)$

Mais calcul de l'automate coûteux :

- Pour chaque position j dans M, pour chaque symbole de A, il faut tester l'égalité entre j préfixes de longueurs $1, 2 \ldots j$.
  Au total : $O(k^3 \times |A|)$
- En fait on peut se ramener à $O(k \times |A|)$
- Cependant il n'est exécuté qu'une seule fois en pré-traitement

RECHERCHE DE MOTIF

UFR IM²AG
UNIVERSITÉ Grenoble Alpes

LA RECHERCHE DU MOTIF

En cas d'échec on recule dans M indépendemment du prochain symbole :

LA RECHERCHE DU MOTIF

En cas d'échec on recule dans M indépendemment du prochain symbole :

- $D[j]$ donne la position **la plus à droite** dans M d'où reprendre

LA RECHERCHE DU MOTIF

En cas d'échec on recule dans M indépendemment du prochain symbole :

- $D[j]$ donne la position **la plus à droite** dans M d'où reprendre
- $D[j] = 0$ si rien ne convient et qu'il faut avancer dans T.

LA RECHERCHE DU MOTIF

En cas d'échec on recule dans M indépendemment du prochain symbole :

- $D[j]$ donne la position **la plus à droite** dans M d'où reprendre
- $D[j] = 0$ si rien ne convient et qu'il faut avancer dans T.

$i := 0$; $j := 0$
tant que $i < n$
  si $T[i] = M[j]$
    $i := i + 1$
    $j := j + 1$
  sinon si $D[j] > 0$
    $j := D[j]$
  sinon
    $i := i + 1$
    $j := 0$
  si $j = k$
    renvoyer $i - k$
renvoyer "motif absent"

LA RECHERCHE DU MOTIF

En cas d'échec on recule dans M indépendemment du prochain symbole :

- $D[j]$ donne la position **la plus à droite** dans M d'où reprendre
- $D[j] = 0$ si rien ne convient et qu'il faut avancer dans T.

$i := 0$; $j := 0$
tant que $i < n$
  si $T[i] = M[j]$
    $i := i + 1$
    $j := j + 1$
  sinon si $D[j] > 0$
    $j := D[j]$
  sinon
    $i := i + 1$
    $j := 0$
  si $j = k$
    renvoyer $i - k$
renvoyer "motif absent"

LA RECHERCHE DU MOTIF

En cas d'échec on recule dans M indépendemment du prochain symbole :

- $D[j]$ donne la position **la plus à droite** dans M d'où reprendre
- $D[j] = 0$ si rien ne convient et qu'il faut avancer dans T.

$i := 0$; $j := 0$
tant que $i < n$
  si $T[i] = M[j]$
    $i := i + 1$
    $j := j + 1$
  sinon si $D[j] > 0$
    $j := D[j]$
  sinon
    $i := i + 1$
    $j := 0$
  si $j = k$
    renvoyer $i - k$
renvoyer "motif absent"

LA RECHERCHE DU MOTIF

En cas d'échec on recule dans M indépendemment du prochain symbole :

- $D[j]$ donne la position **la plus à droite** dans M d'où reprendre
- $D[j] = 0$ si rien ne convient et qu'il faut avancer dans T.

$i := 0$; $j := 0$
tant que $i < n$
  si $T[i] = M[j]$
    $i := i + 1$
    $j := j + 1$
  sinon si $D[j] > 0$
    $j := D[j]$
  sinon
    $i := i + 1$
    $j := 0$
  si $j = k$
    renvoyer $i - k$
renvoyer "motif absent"

PRÉ-TRAITEMENT : CALCUL DE D

Objectif

Pour toute position j dans M, on cherche la longueur d du plus long préfixe de M de la forme $M[j-d \ldots j-1]$.
C'est cette longueur que l'on appellera $D[j]$.

PRÉ-TRAITEMENT : CALCUL DE D

Objectif

Pour toute position j dans M, on cherche la longueur d du plus long préfixe de M de la forme $M[j-d \ldots j-1]$.
C'est cette longueur que l'on appellera $D[j]$.

C'est plus simple que ça n'en a l'air :

Si $M[j-d \ldots j-1]$ est un préfixe de M, alors $M[j-d \ldots j-2]$ aussi.

PRÉ-TRAITEMENT : CALCUL DE D

Objectif

Pour toute position j dans M, on cherche la longueur d du plus long préfixe de M de la forme $M[j-d \ldots j-1]$.
C'est cette longueur que l'on appellera $D[j]$.

C'est plus simple que ça n'en a l'air :

$$\text{Si } M[j-d \ldots j-1] \text{ est un préfixe de } M, \text{ alors } M[j-d \ldots j-2] \text{ aussi.}$$

Par conséquent :

- Si $M[j-1] = M[d]$ alors alors on prolonge le préfixe (éventuellement vide) déjà reconnu
- Sinon on essaye de se rabattre sur un préfixe plus court,
  le meilleur est donné par $D[d]$
- S'il n'y a plus de préfixe plus court utilisable ($D[d] = 0$) on repart du début de M

PRÉ-TRAITEMENT : CALCUL DE D (SUITE)

calcul_D(M)

Données: Un motif de longueur k

Résultat: La table D utilisée dans l'algorithme KMP

$D[0] := 0$

$D[1] := 0$

$d := 0$ est la longueur du plus long préfixe déjà identifié

pour $j := 2$ **jusqu'à** k

  tant que $d > 0$ **et** $M[j-1] \neq M[d]$

    $d := D[d]$

  si $M[j-1] = M[d]$

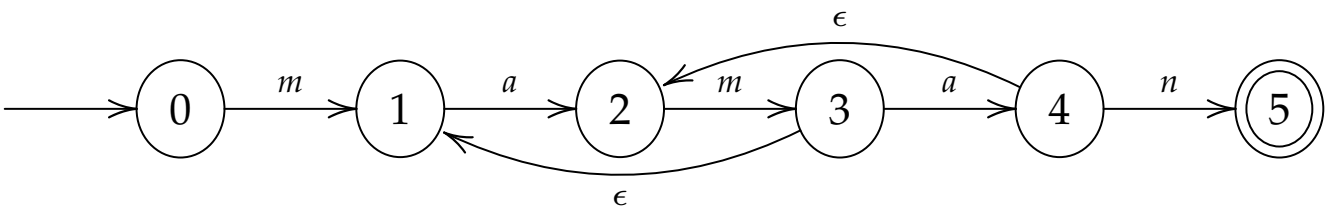
    $d := d + 1$

  $D[j] := d$

renvoyer D

TRADUCTION EN TERMES D'AUTOMATE

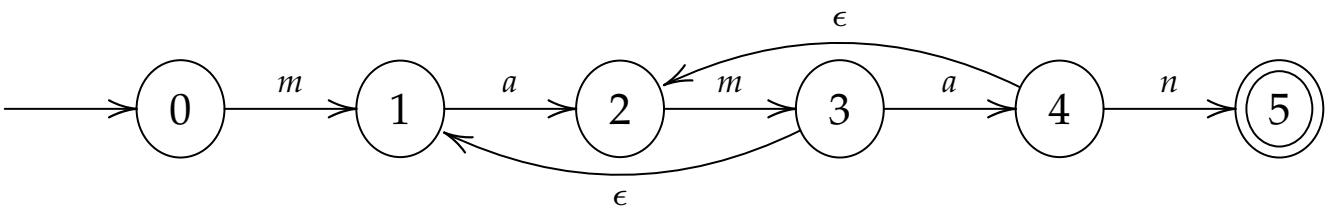
Cela revient à construire



où ϵ représente une **transition vide**

TRADUCTION EN TERMES D'AUTOMATE

Cela revient à construire



où ϵ représente une **transition vide**

On lève le non-déterminisme en imposant de ne suivre les transitions vides **que** si les transitions "vraies" ont échoué.

Comme il y a au maximum une transition vide à partir de chaque état, il y en a au maximum k au lieu des $k \times |A|$ précédentes.

COMPLEXITÉ : RECHERCHE DU MOTIF

On considère la quantité $2i - j$. À chaque itération du Tantque :

tant que $i < n$
- **si** $T[i] = M[j]$
  - $i := i + 1$
  - $j := j + 1$
- **sinon si** $D[j] > 0$
  - $j := D[j]$
- **sinon**
  - $i := i + 1$
  - $j := 0$
- **si** $j = k$
  - **return** $i - k$

COMPLEXITÉ : RECHERCHE DU MOTIF

On considère la quantité $2i - j$. À chaque itération du Tantque :

tant que $i < n$
- **si** $T[i] = M[j]$
  - $i := i + 1$
  - $j := j + 1$
- **sinon si** $D[j] > 0$
  - $j := D[j]$
- **sinon**
  - $i := i + 1$
  - $j := 0$
- **si** $j = k$
  - **return** $i - k$

} $2i - j$ augmente de 1

COMPLEXITÉ : RECHERCHE DU MOTIF

On considère la quantité $2i - j$. À chaque itération du Tantque :

tant que $i < n$
- **si** $T[i] = M[j]$
  - $i := i + 1$
  - $j := j + 1$

  } $2i - j$ augmente de 1
- **sinon si** $D[j] > 0$
  - $j := D[j]$

  } $2i - j$ augmente car j diminue
- **sinon**
  - $i := i + 1$
  - $j := 0$
- **si** $j = k$
  - **return** $i - k$

COMPLEXITÉ : RECHERCHE DU MOTIF

On considère la quantité $2i - j$. À chaque itération du Tantque :

tant que $i < n$
- **si** $T[i] = M[j]$
  - $i := i + 1$
  - $j := j + 1$

  } $2i - j$ augmente de 1
- **sinon si** $D[j] > 0$
  - $j := D[j]$

  } $2i - j$ augmente car j diminue
- **sinon**
  - $i := i + 1$
  - $j := 0$

  } $2i - j$ augmente d'au moins 2
- **si** $j = k$
  - **return** $i - k$

UFR IM²AG
UNIVERSITÉ Grenoble Alpes

COMPLEXITÉ : RECHERCHE DU MOTIF

On considère la quantité $2i - j$. À chaque itération du Tantque :

tant que $i < n$
- **si** $T[i] = M[j]$
  - $i := i + 1$
  - $j := j + 1$

  } $2i - j$ augmente de 1
- **sinon si** $D[j] > 0$
  - $j := D[j]$

  } $2i - j$ augmente car j diminue
- **sinon**
  - $i := i + 1$
  - $j := 0$

  } $2i - j$ augmente d'au moins 2
- **si** $j = k$
  - **return** $i - k$

Comme $2i - j$ est positif et majoré par $2n$, il y a au pire $2n$ itérations.

COMPLEXITÉ : PRÉTRAITEMENT

On considère cette fois $2j - d$:

pour $j := 2$ **jusqu'à** k
  tant que $d > 0$ **et** $M[j-1] \neq M[d]$
    $d := D[d]$
  si $M[j-1] = M[d]$
    $d := d + 1$
  $D[j] := d$

COMPLEXITÉ : PRÉTRAITEMENT

On considère cette fois $2j - d$:

pour $j := 2$ **jusqu'à** k
  tant que $d > 0$ **et** $M[j-1] \neq M[d]$
    $d := D[d]$
  si $M[j-1] = M[d]$
    $d := d + 1$
  $D[j] := d$

} $2j - d$ augmente car d diminue

COMPLEXITÉ : PRÉTRAITEMENT

On considère cette fois $2j - d$:

pour $j := 2$ **jusqu'à** k
  tant que $d > 0$ **et** $M[j-1] \neq M[d]$
    $d := D[d]$
  si $M[j-1] = M[d]$
    $d := d + 1$
  $D[j] := d$

} $2j - d$ augmente car d diminue

} $2j - d$ diminue de 1

COMPLEXITÉ : PRÉTRAITEMENT

On considère cette fois $2j - d$:

pour $j := 2$ **jusqu'à** k } mais $2j - d$ augmente de 2

  tant que $d > 0$ **et** $M[j-1] \neq M[d]$

    $d := D[d]$ } $2j - d$ augmente car d diminue

  si $M[j-1] = M[d]$ } $2j - d$ diminue de 1

    $d := d + 1$

  $D[j] := d$

COMPLEXITÉ : PRÉTRAITEMENT

On considère cette fois $2j - d$:

pour $j := 2$ **jusqu'à** k } mais $2j - d$ augmente de 2

  tant que $d > 0$ **et** $M[j-1] \neq M[d]$

    $d := D[d]$ } $2j - d$ augmente car d diminue

  si $M[j-1] = M[d]$ } $2j - d$ diminue de 1

    $d := d + 1$

  $D[j] := d$

Cette quantité étant majorée par $2k$, le pré-traitement est en $\mathcal{O}(k)$.

RECHERCHE DE MOTIF

UFR IM²AG
UNIVERSITÉ Grenoble Alpes

SYNTHÈSE

- La recherche des occurrences d'un motif de longueur k dans un texte de longueur n est de l'ordre de $n \times k$
- Un prétraitement (calcul de l'automate des préfixes) permet de rendre le coût de l'ordre de n (mais avec un surcoût au début)
- En parcourant le motif par suffixes croissant on peut encore diminuer le nombre de lectures (mais on change de sens dans le parcours du texte) Algorithme de Boyer Moore
- On peut également restreindre le test de motif par une fonction adéquate Algorithme de Karp-Rabin