

# Recherche avec différences

Thierry Lecroq  
Université de Rouen

On cherche toutes les occurrences avec au plus  $k$  différences d'un mot  $x$  de longueur  $m$  dans un texte  $y$  de longueur  $n$ .

# Programmation dynamique

On utilise une table à  $m+1$  lignes et  $n+1$  colonnes :

$$R[i,j] = \min \{ Lev(x[0..i], y[\ell..j]) \text{ avec } 0 \leq \ell \leq j+1 \}$$

$$R[-1,-1] = 0$$

$$R[i,-1] = R[i-1,-1] + \textit{Dél}(x[i])$$

$$R[-1,j] = 0$$

$$R[i,j] = \min \left\{ \begin{array}{l} R[i-1,j-1] + \textit{Sub}(x[i],y[j]) \\ R[i-1,j] + \textit{Dél}(x[i]) \\ R[i,j-1] + \textit{Ins}(y[j]) \end{array} \right.$$

<i>R</i>	<i>j</i>	-1	0	1	2	3	4	5	6	7	8	9	10	11
<i>i</i>		<i>y[j]</i>	<b>C</b>	<b>A</b>	<b>G</b>	<b>A</b>	<b>T</b>	<b>A</b>	<b>A</b>	<b>G</b>	<b>A</b>	<b>G</b>	<b>A</b>	<b>A</b>
-1	<i>x[i]</i>													
0	<b>G</b>													
1	<b>A</b>													
2	<b>T</b>													
3	<b>A</b>													
4	<b>A</b>													

<i>R</i>	<i>j</i>	-1	0	1	2	3	4	5	6	7	8	9	10	11
<i>i</i>		<i>y[j]</i>	<b>C</b>	<b>A</b>	<b>G</b>	<b>A</b>	<b>T</b>	<b>A</b>	<b>A</b>	<b>G</b>	<b>A</b>	<b>G</b>	<b>A</b>	<b>A</b>
-1	<i>x[i]</i>	0	0	0	0	0	0	0	0	0	0	0	0	0
0	<b>G</b>	1	1	1	0	1	1	1	1	0	1	0	1	1
1	<b>A</b>	2	2	1	1	0	1	1	1	1	0	1	0	1
2	<b>T</b>	3	3	2	2	1	0	1	2	2	1	1	1	1
3	<b>A</b>	4	4	3	3	2	1	0	1	2	2	2	1	1
4	<b>A</b>	5	5	4	4	3	2	1	0	1	2	3	2	1

**algo** L-DIFF-DYN( $x, m, y, n, k$ )

$R[-1, -1] \leftarrow 0$

**pour**  $i \leftarrow 0$  à  $m-1$  **faire**

$R[i, -1] \leftarrow R[i-1, -1] + \text{Dél}(x[i])$

**pour**  $j \leftarrow 0$  à  $n-1$  **faire**

$R[-1, j] \leftarrow 0$

**pour**  $i \leftarrow 0$  à  $m-1$  **faire**

$$R[i, j] \leftarrow \min \{ R[i-1, j-1] + \text{Sub}(x[i], y[j]), \\ R[i-1, j] + \text{Dél}(x[i]), \\ R[i, j-1] + \text{Ins}(y[j]) \}$$

**si**  $R[m-1, j] \leq k$  **alors**

signaler une occurrence de  $x$

En considérant les coûts  $Sub(a,b) = Dél(a) = Ins(b) = 1$   
on a :

$$R[-1,-1] = 0$$

$$R[i,-1] = i+1$$

$$R[-1,j] = 0$$

$$R[i,j] = \min \left\{ \begin{array}{ll} R[i-1,j-1] & \text{si } x[i] = y[j] \\ R[i-1,j-1] + 1 & \text{si } x[i] \neq y[j] \\ R[i-1,j] + 1 \\ R[i,j-1] + 1 \end{array} \right.$$

## Lemme

Pour chaque position  $j$  sur le mot  $y$ , on a

$$-1 \leq R[i,j] - R[i-1,j] \leq 1$$

pour  $0 \leq i \leq m-1$ .

## Lemme

Pour chaque position  $i$  sur le mot  $x$ , on a

$$-1 \leq R[i,j] - R[i,j-1] \leq 1$$

pour  $0 \leq j \leq n-1$ .

Proposition (monotonie sur les diagonales)

Pour chaque position  $j$  sur le mot  $y$ , on a

$$R[i-1, j-1] \leq R[i, j] \leq R[i-1, j-1] + 1$$

pour  $0 \leq i \leq m-1$  et  $0 \leq j \leq n-1$ .

- Lorsqu'une valeur égale à  $k+1$  est trouvée dans une colonne, il est inutile de calculer les valeurs suivantes dans la même diagonale.
- Pour élaguer le calcul, on garde trace, dans chaque colonne  $j$ , de la plus grande position à laquelle se trouve une valeur admissible.
- Si  $q_j$  est cette position, seules les valeurs des lignes  $-1$  à  $q_j+1$  sont calculées dans la colonne  $j+1$ .

```

algo L-DIFF-ELAG( $x, m, y, n, k$ )
  pour  $i \leftarrow -1$  à  $k-1$  faire
     $C_1[i] \leftarrow i + 1$ 
   $p \leftarrow k$ 
  pour  $j \leftarrow 0$  à  $n-1$  faire
     $C_2[-1] \leftarrow 0$ 
    pour  $i \leftarrow 0$  à  $p$  faire
      si  $x[i] = y[j]$  alors
         $C_2[i] \leftarrow C_1[i-1]$ 
      sinon
         $C_2[i] \leftarrow \min\{C_1[i-1], C_2[i-1], C_1[i]\} + 1$ 
       $C_1 \leftarrow C_2$ 
      tantque  $C_1[p] > k$  faire
         $p \leftarrow p - 1$ 
    si  $p = m+1$  alors
      signaler une occurrence de  $x$ 
     $p \leftarrow \min\{p + 1, m - 1\}$ 

```

$R$	$j$	-1	0	1	2	3	4	5	6	7	8	9	10	11
$i$		$y[j]$	<b>C</b>	<b>A</b>	<b>G</b>	<b>A</b>	<b>T</b>	<b>A</b>	<b>A</b>	<b>G</b>	<b>A</b>	<b>G</b>	<b>A</b>	<b>A</b>
-1	$x[i]$	0	0	0	0	0	0	0	0	0	0	0	0	0
0	<b>G</b>	1	1	1	0	1	1	1	1	0	1	0	1	1
1	<b>A</b>		2	1	1	0	1	1	1	1	0	1	0	1
2	<b>T</b>				2	1	0	1	2	2	1	1	1	1
3	<b>A</b>						1	0	1	2	2	2	1	1
4	<b>A</b>							1	0	1	2			1