

Une nouvelle heuristique pour l'alignement de motifs 2D par programmation dynamique

Émilie Chanoni, Thierry Lecroq et Alexandre Pauchet
Emilie.Chanoni@univ-rouen.fr, Thierry.Lecroq@univ-rouen.fr,
Alexandre.Pauchet@insa-rouen.fr

Psychologie et Neurosciences de la Cognition et de l'Affectivité (EA 4306)
Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes (EA 4108)
Université de Rouen & INSA-Rouen

Journées Francophones de Planification, Décision et Apprentissage pour la
conduite des systèmes
19 – 20 juin 2008 – Metz



Plan

- 1 Alignement de séquences
- 2 Alignements de motifs 2D
- 3 Expérimentation
- 4 Conclusion et perspectives

Plan

- 1 **Alignement de séquences**
- 2 Alignements de motifs 2D
- 3 Expérimentation
- 4 Conclusion et perspectives

Alignement de séquences

Alignements deux à deux

- utilisés pour comparer 2 séquences x (de longueur m) et y (de longueur n)
- comment transformer x en y ?
- largement utilisés en bioinformatique
- constituent un moyen pour visualiser les ressemblances entre 2 séquences
- basés sur des notions de distance ou de similarité
- calculés par programmation dynamique en $O(mn)$

2 types

- globaux
- locaux (algorithme de Smith et Waterman, 1981)

Alignement de séquences

Exemple

A C G — — A
A T G C T A est un alignement de ACGA et ATGCTA.

Une solution peut également être donnée sous forme de script d'édition comme suit :

Opération	Séquence résultat
substitution de A par A	A
substitution de C par T	AT
substitution de G par G	ATG
insertion de C	ATGC
insertion de T	ATGCT
substitution de A par A	ATGCTA

Alignements locaux

3 opérations d'édition

- substitution d'un symbole de x à une position donnée par un symbole de y
- suppression d'un symbole de x à une position donnée
- insertion d'un symbole de y dans x à une position donnée

Scores

- $Sub(a, b)$: score de la substitution du symbole a par le symbole b
- $Del(a)$: score de la suppression du symbole a
- $Ins(a)$: score d'insertion du symbole a

Mesure de similarité

Mesure de similarité globale

$$d(x, y) = \max\{\text{score de } \gamma \mid \gamma \in \Gamma_{x,y}\}$$

où :

- $\Gamma_{x,y}$: ensemble de toutes les suites d'opérations d'édition qui transforment x en y
- le score d'un élément $\gamma \in \Gamma_{x,y}$ est la somme des scores de ses opérations d'édition élémentaires

Score d'édition

$s(x, y) =$ similarité maximale entre un segment de x et un segment de y

Programmation dynamique

$$t[i, j] = s(x[0..i], y[0..j]) \text{ pour } i = 0, \dots, m - 1 \text{ et } j = 0, \dots, n - 1$$

$$s(x, y) = \max\{t[i, j]\}$$

Formules de récurrence

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

$$t[i, -1] = 0,$$

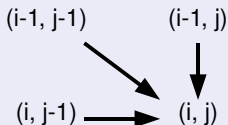
$$t[-1, j] = 0,$$

$$t[i, j] = \max \begin{cases} t[i - 1, j - 1] + \mathit{Sub}(x[i], y[j]), \\ t[i - 1, j] + \mathit{Del}(x[i]), \\ t[i, j - 1] + \mathit{Ins}(y[j]), \\ 0 \end{cases}$$

$$\text{pour } i = 0, 1, \dots, m - 1 \text{ et } j = 0, 1, \dots, n - 1$$

Programmation dynamique

La valeur à la position (i, j) de la table t ne dépend que des valeurs aux 3 positions voisines :



Un alignement optimal (de score maximal) peut être produit en effectuant un **tracé arrière** des calculs des valeurs de la table t à partir de la position maximale jusqu'à une position de valeur 0.

Alignements locaux

Exemple

$Sub(a, a) = 2$, $Sub(a, b) = -1$ et $Del(a) = Ins(a) = -2$

T	j	-1	0	1	2	3	4	5	6	7
i		$y[j]$	Z	A	T	G	C	T	A	W
-1	$x[i]$	0	0	0	0	0	0	0	0	0
0	X	0	0	0	0	0	0	0	0	0
1	A	0	0	2	0	0	0	0	2	0
2	C	0	0	0	1	0	2	2	0	1
3	G	0	0	0	0	3	1	1	0	0
4	A	0	0	2	0	1	2	0	3	1
5	Y	0	0	0	1	0	0	1	2	2

Alignements locaux

Exemple

$Sub(a, a) = 2$, $Sub(a, b) = -1$ et $Del(a) = Ins(a) = -2$

T	j	-1	0	1	2	3	4	5	6	7
i		$y[j]$	Z	A	T	G	C	T	A	W
-1	$x[i]$	0	0	0	0	0	0	0	0	0
0	X	0	0	0	0	0	0	0	0	0
1	A	0	0	2	0	0	0	0	2	0
2	C	0	0	0	1	0	2	2	0	1
3	G	0	0	0	0	3	1	1	0	0
4	A	0	0	2	0	1	2	0	3	1
5	Y	0	0	0	1	0	0	1	2	2

Alignements locaux

Exemple

$Sub(a, a) = 2$, $Sub(a, b) = -1$ et $Del(a) = Ins(a) = -2$

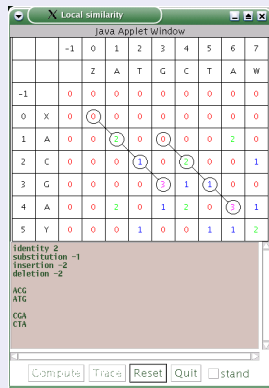
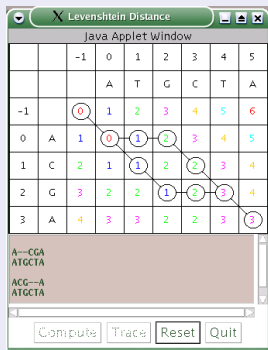
T	j	-1	0	1	2	3	4	5	6	7
i		$y[j]$	Z	A	T	G	C	T	A	W
-1	$x[i]$	0	0	0	0	0	0	0	0	0
0	X	0	0	0	0	0	0	0	0	0
1	A	0	0	2	0	0	0	0	2	0
2	C	0	0	0	1	0	2	2	0	1
3	G	0	0	0	0	3	1	1	0	0
4	A	0	0	2	0	1	2	0	3	1
5	Y	0	0	0	1	0	0	1	2	2

C G A
C T A

Comparaison de séquences

Sur le web

<http://monge.univ-mlv.fr/~lecroq/seqcomp>



Plan

- 1 Alignement de séquences
- 2 Alignements de motifs 2D**
- 3 Expérimentation
- 4 Conclusion et perspectives

2D

X	0	1	2
0	A	B	C
1	D	E	F
2	G	H	I
3	J	K	L

$$|X| = m_1 \times n_1$$

Y	0	1
0	E	C
1	H	I
2	K	L

$$|Y| = m_2 \times n_2$$

Travaux précédents



K. Krithivasan and R. Sitalakshmi

Efficient two-dimensional pattern matching in the presence of errors
Information Sciences, 43(3), 169–184, 1987



R. Baeza-Yates

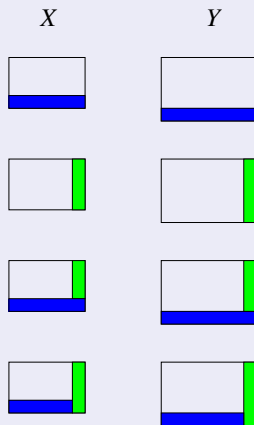
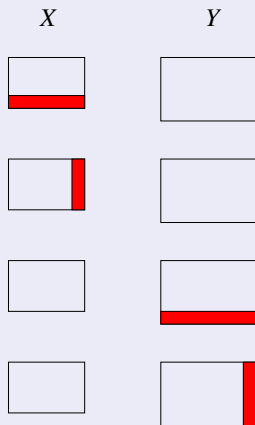
Similarity in Two-Dimensional Strings
COCOON, LNCS 1449, 319–328, 1998



A. Arslan

A Largest Common d -Dimensional Subsequence of Two d -Dimensional
Strings
FCT, LNCS 4639, 40–51, 2007

2D : Alignements globaux



2D : Alignements globaux

4 tables de dimension 2

- $D_R[i, j]$ est le score de la suppression du préfixe de longueur $j + 1$ de la ligne i de X ,
- $D_C[i, j]$ est le score de la suppression du préfixe de longueur $i + 1$ de la colonne j de X ,
- $I_R[i, j]$ est le score de l'insertion du préfixe de longueur $j + 1$ de la ligne i de Y et
- $I_C[i, j]$ est le score de l'insertion du préfixe de longueur $i + 1$ de la colonne j de Y .

2D : Alignements globaux

2 tables de dimension 4

- $R[i, j, k, \ell] = d(X[i, 0..j], Y[k, 0.. \ell])$
- $C[i, j, k, \ell] = d(X[0..i, j], Y[0..k, \ell])$

- $R[i, j, k, \ell]$: score entre le préfixe de longueur $j + 1$ de la ligne i de X et le préfixe de longueur $\ell + 1$ de la ligne k de Y .
- $C[i, j, k, \ell]$: score entre le préfixe de longueur $i + 1$ de la colonne j de X et le préfixe de longueur $k + 1$ de la colonne ℓ de Y .

2D : Alignements globaux

$$T[i, j, k, \ell] = \max \begin{cases} T[i-1, j, k, \ell] + D_R[X[i, 0..j]] \\ T[i, j-1, k, \ell] + D_C[X[0..i, j]] \\ T[i, j, k-1, \ell] + I_R[Y[k, 0.. \ell]] \\ T[i, j, k, \ell-1] + I_C[Y[0..k, \ell]] \\ T[i-1, j, k-1, \ell] + R[i, j, k, \ell] \\ T[i, j-1, k, \ell-1] + C[i, j, k, \ell] \\ T[i-1, j-1, k-1, \ell-1] + C[i, j-1, k, \ell-1] + R[i, j, k, \ell] \\ T[i-1, j-1, k-1, \ell-1] + C[i, j, k, \ell] + R[i-1, j, k-1, \ell] \end{cases}$$

2D : Alignements globaux

Exemple

<i>X</i>			<i>Y</i>	
A	B	C	E	C
D	E	F	H	I
G	H	I	K	L
J	K	L		

2D : Alignements globaux

Un alignement global entre 2 motifs rectangulaires de taille respective $M = m_1 \times n_1$ et $N = m_2 \times n_2$ peut être calculé en temps et espace $O(M \times N)$.

2D : Alignements locaux

X	0	1	2	3	4
0	U	U	U	U	U
1	U	A	B	C	U
2	U	D	E	F	U
3	U	G	H	I	U
4	U	J	K	L	U
5	U	U	U	U	U

$$|X| = m_1 \times n_1$$

Y	0	1	2	3
0	V	V	V	V
1	V	E	C	V
2	V	H	I	V
3	V	K	L	V
4	V	V	V	V

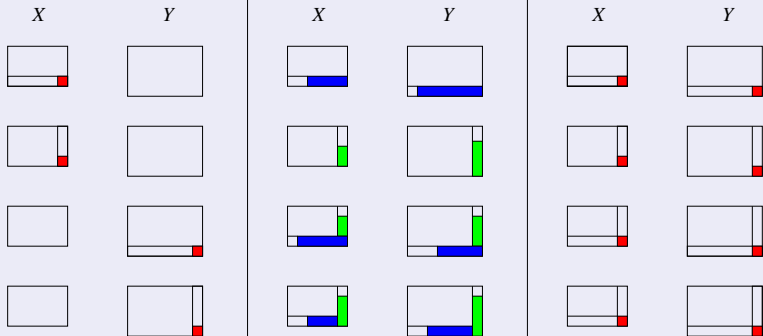
$$|Y| = m_2 \times n_2$$

2D : Alignements locaux

2 tables de dimension 4

- $R_S[i, j, k, \ell] = s(X[i, 0..j], Y[k, 0.. \ell])$
- $C_S[i, j, k, \ell] = s(X[0..i, j], Y[0..k, \ell])$
- $R_S[i, j, k, \ell]$: similarité maximale entre un suffixe du préfixe de longueur $j + 1$ de la ligne i de X et un suffixe du préfixe de longueur $\ell + 1$ de la ligne k de Y
- $C_S[i, j, k, \ell]$: similarité maximale entre un suffixe du préfixe de longueur $i + 1$ de la colonne j de X et un suffixe du préfixe de longueur $k + 1$ de la colonne ℓ de Y

2D : Alignements locaux



2D : Alignements locaux

$$\begin{aligned}r &= R_S[i, j, k, \ell], & c &= C_S[i, j, k, \ell] \\r' &= R_S[i - 1, j, k - 1, \ell], & c' &= C_S[i, j - 1, k, \ell - 1]\end{aligned}$$

Formule de récurrence

$$T[i, j, k, \ell] = \max \begin{cases} T[i - 1, j, k, \ell] + Del[X[i, j]] \\ T[i, j - 1, k, \ell] + Del[X[i, j]] \\ T[i, j, k - 1, \ell] + Ins[Y[k, \ell]] \\ T[i, j, k, \ell - 1] + Ins[Y[k, \ell]] \\ T[i - 1, j, k - 1, \ell] + (r \text{ si } r \neq 0 \text{ sinon } Del[X[i, j]] + Ins[Y[k, \ell]]) \\ T[i, j - 1, k, \ell - 1] + (c \text{ si } c \neq 0 \text{ sinon } Del[X[i, j]] + Ins[Y[k, \ell]]) \\ T[i - 1, j - 1, k - 1, \ell - 1] + (c' + r \text{ si } c', r \neq 0 \text{ sinon } Del[X[i, j]] + Ins[Y[k, \ell]]) \\ T[i - 1, j - 1, k - 1, \ell - 1] + (c + r' \text{ si } c, r' \neq 0 \text{ sinon } Del[X[i, j]] + Ins[Y[k, \ell]]) \\ 0 \end{cases}$$

2D : Alignements locaux

Exemple

<i>X</i>					<i>Y</i>			
U	U	U	U	U	V	V	V	V
U	A	B	C	U	V	E	C	V
U	D	E	F	U	V	H	I	V
U	G	H	I	U	V	K	L	V
U	J	K	L	U	V	V	V	V
U	U	U	U	U				

2D : Alignements locaux

Un alignement local entre 2 motifs rectangulaires de taille respective $M = m_1 \times n_1$ et $N = m_2 \times n_2$ peut être calculé en temps et espace $O(M \times N)$.

Plan

- 1 Alignement de séquences
- 2 Alignements de motifs 2D
- 3 Expérimentation**
- 4 Conclusion et perspectives

Expérimentation

- Dialogues entre parents et enfants de 4 ans lors de la narration de 2 histoires enfantines (A et B)
- Chaque énoncé est retranscrit et codé selon une grille [Chanoni 2004]
- Un énoncé correspond à une ligne
- Une matrice de substitution (*Sub*) a été spécifiquement construite

:
 :
 25 Pb7 t'inquiète pas
 26 Pb7 on va la retrouver ta couronne
 27 Pb7 t'inquiète pas
 28 Pb7 donc là ils se cachent
 29 Pb7 ils cherchent
 30 Pb7 qui pourrait avoir pris la couronne
 31 b7 elle dedans, elle est dedans la
 couronne
 32 Pb7 donc ils suspectent plein de monde,
 Cornélius, Céleste, la vieille dame
 33 Pb7 qui a bien pu prendre la couronne ?
 34 b7 la couronne elle est dedans
 35 Pb7 tu crois ? !
 36 b7 oui
 37 Pb7 mais Babar il ne sait pas qu'elle
 est dedans
 38 Pb7 donc il se dit que c'est une bombe,
 la couronne
 39 Pb7 ou je ne sais quoi ?
 :
 :

:
 :
 7 Pb9 même son ami Zéphir la cherche
 partout avec sa loupe
 8 Pb9 mais où est elle donc passée
 9 Pb9 Babar remarque Cornélius donner
 un paquet à l'intrus
 10 Pb9 Qu'est ce qui peut y avoir dedans ?
 11 Pb9 mais qui est donc cet individu
 masqué
 12 Pb9 qui a volé la couronne
 13 Pb9 Babar la démasque !
 14 Pb9 c'est la reine céleste !
 15 Pb9 il se pose bien des questions
 16 Pb9 pourquoi donc la reine Céleste
 s'est déguisée.
 17 Pb9 Babar va donc chez la vieille dame
 lui demander
 18 b9 oui
 19 Pb9 la vieille dame ne veut pas qu'il
 rentre !
 20 Pb9 mais derrière se cachait
 21 Pb9 pour lui une surprise en réalité
 22 Pb9 puis Babar rentra chez lui
 :
 :

b7-BABAR

24	a	[{)]
25	A	P	E)]
26	A	P	B)]
27	a	[{)]
28	q	[{)]
29	a	[{)]
30	A	P	Y	C	J
31	q	[{)]
32	a	[{)]
33	Q	H	K)]
34	a	[{)]
35	A	P	N	O	J
36	A	P	N	C	J
37	A	R	N)]
38	a	[{)]

b9-BABAR

8	q	[{)]
9	A	P	B)]
10	q	[{)]
11	q	[{)]
12	q	[{)]
13	A	P	B)]
14	a	[{)]
15	A	P	N)]
16	q	[{)]
17	a	[{)]
18	a	[{)]
19	A	P	V)]
20	A	P	B	O	J
21	A	P	S	O	J
22	a	[{)]
23	a	[{)]

b7-BABAR

24	a	[{)]
25	A	P	E)]
26	A	P	B)]
27	a	[{)]
28	q	[{)]
29	a	[{)]
30	A	P	Y	C	J
31	q	[{)]
32	a	[{)]
33	Q	H	K)]
34	a	[{)]
35	A	P	N	O	J
36	A	P	N	C	J
37	A	R	N)]
38	a	[{)]

b9-BABAR

8	q	[{)]
9	A	P	B)]
10	q	[{)]
11	q	[{)]
12	q	[{)]
13	A	P	B)]
14	a	[{)]
15	A	P	N)]
16	q	[{)]
17	a	[{)]
18	a	[{)]
19	A	P	V)]
20	A	P	B	O	J
21	A	P	S	O	J
22	a	[{)]
23	a	[{)]

b7-BABAR

24	a	[{)]
25	A	P	E)]
26	A	P	B)]
27	a	[{)]
28	q	[{)]
29	a	[{)]
30	A	P	Y	C	J
31	q	[{)]
32	a	[{)]
33	Q	H	K)]
34	a	[{)]
35	A	P	N	O	J
36	A	P	N	C	J
37	A	R	N)]
38	a	[{)]

b9-BABAR

8	q	[{)]
9	A	P	B)]
10	q	[{)]
11	q	[{)]
12	q	[{)]
13	A	P	B)]
14	a	[{)]
15	A	P	N)]
16	q	[{)]
17	a	[{)]
18	a	[{)]
19	A	P	V)]
20	A	P	B	O	J
21	A	P	S	O	J
22	a	[{)]
23	a	[{)]

b7-BABAR

- 25 Pb7 t'inquiète pas
 26 Pb7 on va la retrouver ta couronne
 27 Pb7 t'inquiète pas
 28 Pb7 donc là ils se cachent
 29 Pb7 ils cherchent
 30 Pb7 qui pourrait avoir pris la couronne
 31 b7 elle dedans, elle est dedans la couronne
 32 Pb7 donc ils suspectent plein de monde,
 Cornélius, Céleste, la vieille dame
 33 Pb7 qui a bien pu prendre la couronne ?
 34 b7 la couronne elle est dedans
 35 Pb7 tu crois ? !
 36 b7 oui
 37 Pb7 mais Babar il ne sait pas qu'elle
 est dedans
 38 Pb7 donc il se dit que c'est une
 bombe, la couronne
 39 Pb7 ou je ne sais quoi ?

b9-BABAR

- 7 Pb9 même son ami Zéphir la cherche
 partout avec sa loupe
 8 Pb9 mais où est elle donc passée
 9 Pb9 Babar remarque Cornélius donner
 un paquet à l'intrus
 10 Pb9 Qu'est ce qui peut y avoir dedans ?
 11 Pb9 mais qui est donc cet individu
 masqué
 12 Pb9 qui a volé la couronne
 13 Pb9 Babar la démasque !
 14 Pb9 c'est la reine céleste !
 15 Pb9 il se pose bien des questions
 16 Pb9 pourquoi donc la reine Céleste
 s'est déguisée.
 17 Pb9 Babar va donc chez la vieille dame
 lui demander
 18 b9 oui
 19 Pb9 la vieille dame ne veut pas qu'il
 rentre !
 20 Pb9 mais derrière se cachait
 21 Pb9 pour lui une surprise en réalité
 22 Pb9 puis Babar rentra chez lui

Plan

- 1 Alignement de séquences
- 2 Alignements de motifs 2D
- 3 Expérimentation
- 4 Conclusion et perspectives**

Conclusions

- Alignements globaux de motifs 2D
- Alignements locaux de motifs 2D

Limitations

- Limitation de la taille des fichiers/dialogues
- L'algorithme implique des motifs en forme de 'L'
- Difficulté à définir un motif 'consensuel'

Questionnements

- Est-il possible d'aligner des motifs 2D en temps sous-quadratique comme pour les séquences [CLZ 2003] ?
- Est-il possible de construire une heuristique telle que BLAST [AGMML 1990] ou FASTA [PL 1988] pour aligner rapidement un motif 2D avec une banque de motifs 2D ?
- Peut-on utiliser des techniques de vecteurs de bits dans le cas de « petits » motifs [Myers 1998] ?

Perspectives

- Recherche des alignements dans les 4 dimensions (rotations à 90°)
- Validation statistique des motifs trouvés
- Recherche des k meilleurs alignements
- Constitution d'une banque de motifs similaires
- Visualisation des alignements en 2D
- Autres applications nécessitant la recherche de motifs en 2D