

Alignement multiple

Hélène Touzet

Équipe Bioinfo — LIFL — USTL

Master recherche informatique

www.lifl.fr/~touzet/masterrecherche.html

Alignement 2 à 2

2 séquences quelconques



détection d'une similarité
syntaxique



? fonction commune ?

Alignement multiple

Famille de séquences
partageant une même fonction



A quelle conservation
syntaxique
cela correspond-il ?

Motifs protéiques

▶ Exemple 1: hormone pancréatique

PMY_PETMA/1-36	PEE..LSKYMLAVRN Y INLIT RQRY
PPY_LOPAM/1-36	PED..WASYQA AVRH Y VNLIT RQRY
PAHO_BOVIN/30-65	PEQ..MAQYAAELRR Y INMLT RPRY
PAHO_CHICK/26-61	VED..LIRFYNDLQQ Y LNVT RHRY
PAHO_ANSAN/1-36	VED..LRFYDNLQQ Y RLNVF RHRY
NPF_HELAS/4-39	PNE..LRQYLKELNE Y YAIMG RTRF
NPF_MONEX/1-39	DNKAALRDYLRQINE Y FAIIG RPRF

▶ Expression Prosite

[FY]-x(3)-[LIVM]-x(2)-**Y**-x(3)-[LIVMFY]-x-**R**-x-**R**-[YF]

▶ Syntaxe

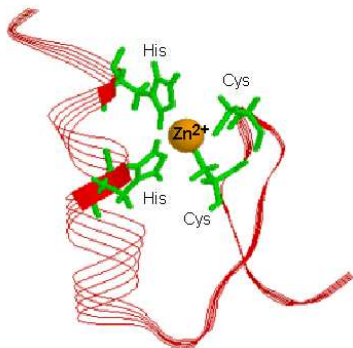
- : séparation des éléments
- x : n'importe quel acide aminé
- (3,5) : nombre d'occurrences (entre 3 et 5)
- [FY] : alternative (F ou Y)

► Exemple 2 : doigt de zinc

YVCPFDGCN---KKFAQSTNLKSHILT---H
YKCT--VCR---KDISSESRLRTHMFKQ--HH
FQCD--ICK---KTFKNACSVKIHHKN--MH
LKCSVPGCK---RSFRKKRALRIHVSE---H
FECN--MCG---YHSQDRYEFSSHITRG--EH
YTCG--YCTEDSPSFPRPSLLESHISL--MH
YKCEFADCE---KAFSNASDRAKHQNR--TH
YKCN--QCG---IIFSQNSPFIVHQIA---H
FVCHWQDCSRELRPFKAQYMLVVHMRR---H
FRCS--ECS---RSFTHNSDLTAHMRK---H
CKCETENCN---LAFTTASNMRLLHFKR--AH
YRCSYEDCQ---TVSPTWTALQTHLKK---H
FRCV--WCK---QSFPTLEALTTHMKDS--KH
FRCGYKCG---RLYTTAHLKVVHERA---H
YRCPRENCN---RTYTTKFNLSHILT--FH
YTCPEPHCG---RGFTSATNYKNHVRI---H

C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

► Exemple 2 : doigt de zinc



C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

Comment scorer un alignement multiple

- ▶ **Score SP - sums of pairs**: somme des scores de ses colonnes
- ▶ Comment scorer une colonne ?
 - ▶ adaptable à un nombre quelconque de lignes
 - ▶ indépendant de l'ordre
 - ▶ reflète la similarité

$$\text{scoreSP} \left(\begin{pmatrix} c_1 \\ \vdots \\ c_k \end{pmatrix} \right) = \sum_{1 \leq i < j \leq k} \text{score}(c_i, c_j)$$

$c_1, \dots, c_k \in \mathcal{A} \cup \{-\}$ et $\text{score}(-, -) = 0$

A A C G T A C G A T A
A - C G T A - A A T G
G T C G T A - - T T A

Identité : +1,
Substitution: -2,
Indel : -3

Définition alternative (équivalente)

- ▶ α : alignement multiple pour les séquences s_1, \dots, s_k
- ▶ α_{ij} : projection de l'alignement pour s_i et s_j

$$\text{scoreSP}(\alpha) = \sum_{1 \leq i < j \leq k} \text{score}(\alpha_{ij})$$

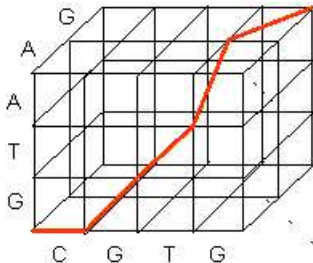
- ▶ Retour à l'exemple

```
A A C G T A C G A T A
A - C G T A - A A T G
G T C G T A - - T T A
```

Identité : +1,
Substitution: -2,
Indel : -3

Algorithme exact

- ▶ Trouver l'alignement multiple de score SP maximal
- ▶ Programmation dynamique
 - ▶ Alignement deux à deux : chemin dans une matrice de dimension 2
 - ▶ Alignement multiple de ℓ séquences: chemin dans une matrice de dimension ℓ



C	G	T	-	G
-	G	T	A	-
-	-	-	A	G

Exemple pour trois séquences (U , V et W)

- ▶ Matrice en dimension trois
- ▶ $\text{Sim}(i, j, k)$: score optimal entre $U(1..i)$, $V(1..j)$ et $W(1..k)$.
- ▶ Formule de récurrence :

$$\text{Sim}(0, 0, 0) = 0$$

$$\text{Sim}(0, 0, k) = \text{Sim}(0, 0, k - 1) + SP(-, -, W(k))$$

$$\text{Sim}(0, j, 0) = \text{Sim}(0, j - 1, 0) + SP(-, V(j), -)$$

$$\text{Sim}(i, 0, 0) = \text{Sim}(i - 1, 0, 0) + SP(U(i), -, -)$$

$$\text{Sim}(0, j, k) = \max \begin{cases} \text{Sim}(0, j - 1, k - 1) + SP(-, V(j), W(k)) \\ \text{Sim}(0, j - 1, k) + SP(-, V(j), -) \\ \text{Sim}(0, j, k - 1) + SP(-, -, W(k)) \end{cases}$$

$$\text{Sim}(i, 0, k) = \max \begin{cases} \text{Sim}(i - 1, 0, k - 1) + SP(U(i), -, W(k)) \\ \text{Sim}(i - 1, 0, k) + SP(U(i), -, -) \\ \text{Sim}(i, 0, k - 1) + SP(-, -, W(k)) \end{cases}$$

$$\text{Sim}(i, j, 0) = \max \begin{cases} \text{Sim}(i - 1, j - 1, k) + SP(U(i), V(j), -) \\ \text{Sim}(i - 1, j, k) + SP(U(i), -, -) \\ \text{Sim}(i, j - 1, k) + SP(-, V(j), -) \\ \text{Sim}(i, j, k - 1) + SP(-, -, W(k)) \end{cases}$$

$$\text{Sim}(i, j, k) = \max \left\{ \begin{array}{l} \text{Sim}(i-1, j-1, k-1) + SP(U(i), V(j), W(k)) \\ \text{Sim}(i-1, j-1, k) + SP(U(i), V(j), -) \\ \text{Sim}(i-1, j, k-1) + SP(U(i), -, W(k)) \\ \text{Sim}(i-1, j, k) + SP(U(i), -, -) \\ \text{Sim}(i, j-1, k-1) + SP(-, V(j), W(k)) \\ \text{Sim}(i, j-1, k) + SP(-, V(j), -) \\ \text{Sim}(i, j, k-1) + SP(-, -, W(k)) \end{array} \right.$$

Algorithme exact – complexité

- ▶ n : longueur des séquences
- ▶ 2 séquences : $O(n^2)$ en temps et en espace
- ▶ 3 séquences : $O(n^3)$ en temps et en espace
- ▶ ℓ séquences, s_1, \dots, s_ℓ
 - ▶ $\text{Sim}(i_1, \dots, i_\ell)$: score optimal entre les ℓ préfixes $s_1(1..i_1), \dots, s_\ell(1..i_\ell)$
 - ▶ Table de taille n^ℓ
 - ▶ Temps de calcul d'une case : dépend de $2^\ell - 1$ cases précédentes
 - ▶ Temps de calcul de chaque scoreSP candidat : $\ell(\ell - 1)/2$
 - ▶ Temps exponentiel: $O(n^\ell 2^{\ell^2})$
- ▶ **L**e problème de décision associé est NP-complet

Approches heuristiques

HEURISTIQUE (du grec *heuriskein*, trouver)
Qui sert à la découverte

- ▶ Heuristique en étoile
- ▶ Clustal (le plus populaire)
- ▶ Dialign2 (complémentaire à Clustal)
- ▶ T-coffee, Pima, Multalin, ...

Autant d'alignements que de programmes

Heuristique en étoile

- ▶ **S**élection d'une séquence centrale
- ▶ **C**onstruction de l'alignement multiple, en partant de la séquence centrale, puis en incorporant une à une les autres séquences
- ▶ **E**xemple



s_1	cgatgagtcattgtgactg
s_2	cgagccattgtagctactg
s_3	cgaccattgtagctacctg
s_4	cgatgagtcactgtgactg

indel : -2, substitution : -1, identité : 1

► Étape 1 : Alignements globaux de toutes les séquences deux par deux

```

s1 cgatgagtcattgt-g--actg   s2 cgagccattgtagcta-ctg
   ||| |  ||||| |  |||     ||| ||| ||| ||| ||| |||
s2 cga-g--ccattgtagctactg   s3 cga-ccattgtagctacctg

```

```

s1 cgatgagtcattg-tgactg     s2 cga-g--ccattgtagctactg
   ||| | | | | | | |||     ||| |  || ||| |  |||
s3 cgacca-ttgtagctacctg     s4 cgatgagtcactgt-g--actg

```

```

s1 cgatgagtcattgtgactg     s3 cgaccattgtagctacctg
   ||| ||| ||| ||| |||     ||| | | | |||
s4 cgatgagtcactgtgactg     s4 cgatgagtcactgtgactg

```

Tableau des scores

	s_1	s_2	s_3	s_4
s_1		2	0	17
s_2	2		14	0
s_3	0	14		-1
s_4	17	0	-1	

l séquences
 ↓
 $l(l-1)/2$ alignements

- **Étape 2** : sélection de la séquence centrale à partir du tableau des scores: séquence qui maximise la somme des similarités avec l'ensemble des autres séquences

	s_1	s_2	s_3	s_4	
s_1		2	0	17	19
s_2	2		14	0	16
s_3	0	14		-1	13
s_4	17	0	-1		16

- ▶ **Étape 3** : construction de l'alignement multiple par juxtaposition des alignements deux à deux avec la séquence centrale

```

s1 cgatgagtcattgt-g--actg   s1 cgatgagtcattg-tgactg
  ||| |  ||||| |  |||      ||| | | | | | | |||
s2 cga-g--ccattgtagctactg   s3 cgacca-ttgtagctacctg

```

```

      s1 cgatgagtcattgtgactg
          ||||| |||||
s4 cgatgagtcactgtgactg

```

Alignement multiple

```

s1 cgatgagtcattg-t-g--actg
s2 cga-g--ccattg-tagctactg
s3 cgacca-ttgtagct-ac--ctg
s4 cgatgagtcactg-t-g--actg

```

L'intégration d'une nouvelle séquence se fait en prenant la séquence centrale comme guide. C'est toujours possible en étirant les gaps de l'alignement multiple courant.

Le programme Clustal

Thompson *et al.* - 1994

- ▶ Clustal = CLUSTER + ALIGNment
- ▶ Inspiré par la classification hiérarchique ascendante
- ▶ Regroupement progressif des séquences
- ▶ Exemple

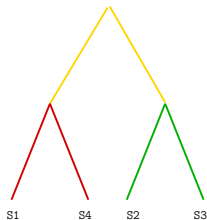
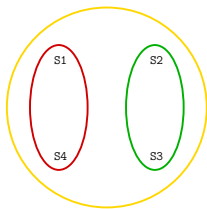
s ₁	cgatgagtcattgtgactg
s ₂	cgagccattgtagctactg
s ₃	cgaccattgtagctacctg
s ₄	cgatgagtcactgtgactg

indel : -2, substitution : -1, identité : 1

- ▶ **Étape 1** : Alignements globaux de toutes les séquences deux par deux
- ▶ Les séquences sont regroupées suivant leur similarité à partir de la matrice des scores 2 à 2.

Tableau des scores

	s_1	s_2	s_3	s_4
s_1		2	0	17
s_2	2		14	0
s_3	0	14		-1
s_4	17	0	-1	



- ▶ **Étape 2** : construction de l'alignement à partir de l'arbre guide
- ▶ Arbre guide : classification hiérarchique ascendante
- ▶ Alignement entre deux clusters de séquences : alignement deux à deux avec le score SP pour le score d'une colonne
- ▶ L'alignement est obtenu par extensions successives.
- ▶ "Once a gap, always a gap"

```
s2 cga---gccattgtagctac-tg
s3 cga---ccattgtagctacctg
s1 cgatgagtcattgt-g--ac-tg
s4 cgatgagtcactgt-g--ac-tg
```

```
s1 cgatgagtcattgtgactg
||||| |||||
s4 cgatgagtcactgtgactg
```

```
s2 cgagccattgtagcta-ctg
||| ||||| |||
s3 cga-ccattgtagctacctg
```

```
s1 cgatgagtcattgtgactg
```

```
s4 cgatgagtcattgtgactg
```

```
s2 cgagccattgtagctactg
```

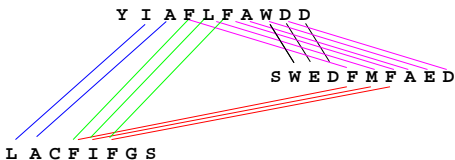
```
s3 cgaccattgtagctacctg
```

Dialign2

Morgenstern *et al.* - 1999

Construire un alignement à partir des diagonales du dot plot

Étape 1: Détection des diagonales dans les paires de séquences



Étape 2: Sélection d'un ensemble cohérent de diagonales pour construire l'alignement

▶ Pas de croisements

▶ Pas de chevauchements

▶ Score maximal

```
y I A - F L F A W D d  
- L A c F I F g s - -  
s w e d F M F A E D -
```

Exemple (C. Notre-Dame)

```
GARFIELD THE LAST FAT CAT
GARFIELD THE FAT CAT
GARFIELD THE VERY FAST CAT
THE FAT CAT
```

► Alignement fourni par Clustal

```
seq1    GARFIELDTHELASTFA-TCAT
seq2    ----GARFIELDTHEFA-TCAT
seq3    GARFIELDTHEVERYFASTCAT
seq4    -----THEFA-TCAT
```

► Alignement fourni par Dialign2

```
seq1    GARFIELD THE LAST FA-T CAT
seq2    GARFIELD THE ---- FA-T CAT
seq3    GARFIELD THE VERY FAST CAT
seq4    ----- THE ---- FA-T CAT
```

