Alignements multiples

Informatique Génomique - Master 1

Guillaume Blin

IGM-LabInfo UMR 8049,
Bureau 4B066
Université de Marne La Vallée
gblin@univ-mlv.fr
http://igm.univ-mlv.fr/∼gblin

2007-08

Plan

Plan

Définition

Un alignement multiple

- En entrée : un ensemble de k séquences

  * * * * * * * * * * * * * * *

  * * * * * * * * * * *

  * * * * * * * * * * * * * * *

  * * * * * * * * * * *

- En sortie : un ensemble de k séquences pourvues de gaps, de même longueur et tel qu'il n'existe pas de colonne contenant k gaps

  * * * * * * * * * * - * * * *

  * * * - - - * * * * - * * * *

  * * * - * * * * * * * * * * *

  * * * - - * * * - - * * * * *

Définition

Exemple - Opsin Rh2 (Ocellar opsin)

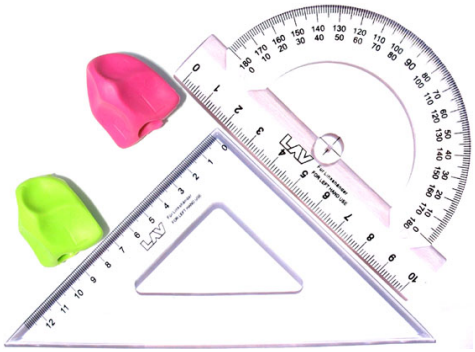
Protéine ayant une rôle dans la pigmentation des yeux de différentes mouches

```
OPS2_DROME   MERSHLPETP FDLAHSGPRF QAQSSGNGSV LDNVLPDMAH LVNPYWSRFA
OPS2_DROPS   MERSLLPEPP LAMALLGPRF EAQTGGNRSV LDNVLPDMAP LVNPHWSRFA
OPS2_LIMPO   ---------- MANQLSYSSL GWPYQPNASV VDTMPKEMLY MIHEHWYAFP
OPS2_HEMSA   ---MTNATGP QMAYYGAASM DFGYPEGVSI VDFVRPEIKP YVHQHWYNYP
OPS2_SCHGR   -----MVNTT DFYPVPAAMA YESSVGLPLL GWNVPTEHLD LVHPHWRSFQ
OPS2_PATYE   ---------- ---------- ---------- --------MP FPLNRTDTAL
```

Objectif

... d'un point de vue informatique

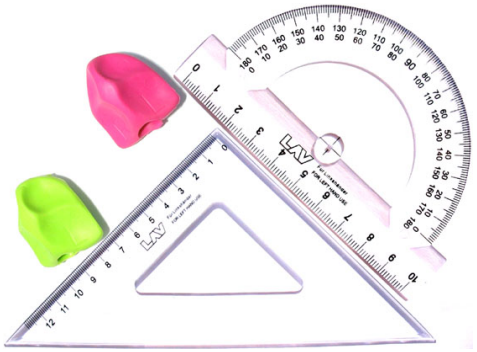
- En entrée : un ensemble de séquences et une méthode de score
- But : mettre en correspondance les séquences en cherchant à maximiser le score



Objectif

... d'un point de vue informatique

- En entrée : un ensemble de séquences et une méthode de score
- But : mettre en correspondance les séquences en cherchant à maximiser le score



... d'un point de vue biologie

- En entrée : un ensemble de séquences
- But : mettre en correspondance les séquences en cherchant à mettre en évidence les relations biologiques

La démarche

... d'un point de vue

- informatique, cela ressemble à l'alignement deux à deux
- biologique, la démarche est totalement différente

Alignement 2 à 2

Deux séquences quelconques

⇓

Détecter une similarité syntaxique

⇓

Y a-t-il une fonction commune ?

Alignement multiple

Famille de séquences ayant la même fonction

⇓

À quelle conservation syntaxique cela correspond-il ?

Utilité de l'alignement multiple

À quoi ça sert ?

- Identifier les régions des séquences qui ont été sujettes à une altération durant l'évolution
- Classifier des séquences en familles – en effet, l'alignement multiple d'un ensemble de séquences peut être vu comme une histoire de l'évolution des séquences
  - un *bon* alignement = forte probabilité de dérivation d'un ancêtre commun
  - un *mauvais* alignement (*i.e.* dont le score de l'alignement est faible) = relation plus complexe et distante dans l'évolution

Evaluer un alignement multiple

Score SP - Sums of Pairs

- On affecte un score à l'alignement correspondant à la somme des scores de ses colonnes
- Quel score donner à une colonne ?
  - adaptable à un nombre quelconque de lignes
  - indépendant de l'ordre des séquences
  - reflètant la similarité
- Solution :

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

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

Evaluer un alignement multiple

Exemple

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	
S_1	A	A	C	G	T	A	C	G	A	T	A	
S_2	A	-	C	G	T	A	-	A	A	T	G	
S_3	G	T	C	G	T	A	-	-	T	T	A	
S_1/S_2	1	-3	1	1	1	1	-3	-2	1	1	-2	
S_1/S_3	-2	-2	1	1	1	1	-3	-3	-2	1	1	
S_2/S_3	-2	-3	1	1	1	1	0	-3	-2	1	-2	
	-3	-8	3	3	3	3	-6	-8	-3	3	-3	=-16

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

Evaluer un alignement multiple

Définition alternative (équivalente)

$$Score(\alpha) = \sum_{1 \leq i < j \leq k} Score2à2(\alpha_i, \alpha_j)$$

où α_i est la $i^{ème}$ séquence de l'alignement α

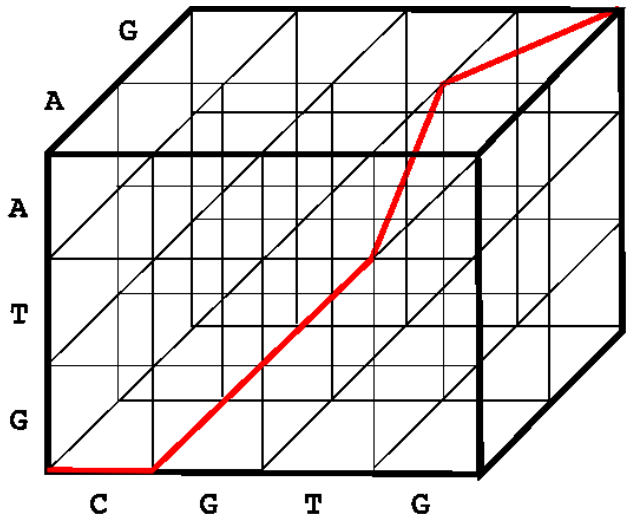
	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	
S_1	A	A	C	G	T	A	C	G	A	T	A	
S_2	A	-	C	G	T	A	-	A	A	T	G	
S_3	G	T	C	G	T	A	-	-	T	T	A	
S_1/S_2	1	-3	1	1	1	1	-3	-2	1	1	-2	=-3
S_1/S_3	-2	-2	1	1	1	1	-3	-3	-2	1	1	=-6
S_2/S_3	-2	-3	1	1	1	1	0	-3	-2	1	-2	=-7
												=-16

Plan

Algorithme exact

Par programmation dynamique

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




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

Algorithme exact

Exemple pour trois séquences (U,V et W)

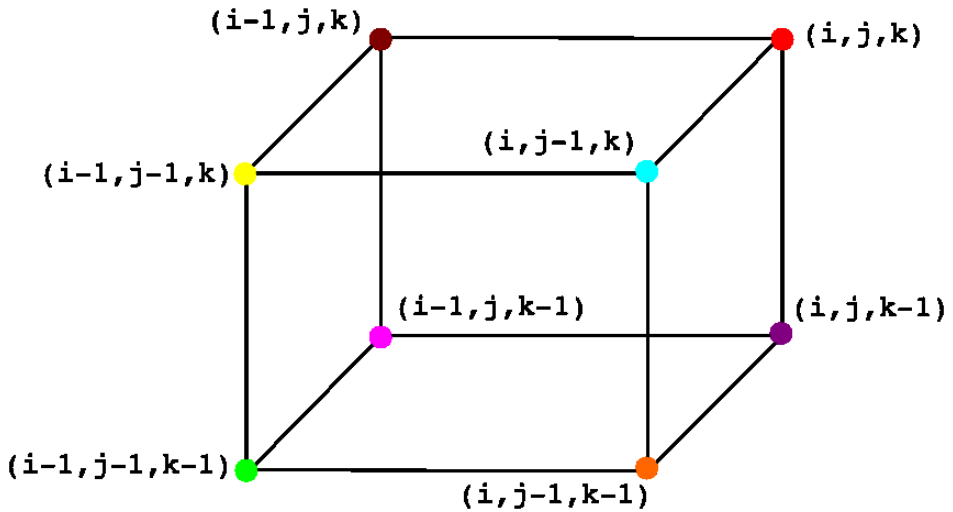
- Matrice de dimension 3
- Sim(i,j,k) : score optimal entre U(1..i), V(1..j) et W(1..k)

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

Algorithme exact

Exemple pour trois séquences (U,V et W)

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



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
- k séquences, $s_1, s_2, \ldots, s_k$:
  - $Sim(i_1, \ldots, i_k)$: score optimal entre les k préfixes $s_1(1 \ldots i_1)$, $\ldots$, $s_k(1 \ldots i_k)$
  - Table de taille n^k
  - Temps de calcul d'une case : dépend de $2^k - 1$ cases précédentes [1]
  - Temps de calcul de chaque score SP : $\frac{k(k-1)}{2} = \sum_{i=1}^{k-1} i$
  - Temps exponentiel : $O(n^k 2^k k^2)$
- Le problème de décision correspondant est NP-complet

[1] Nb de mots binaires de taille k sans la position actuelle

La NP-complétude

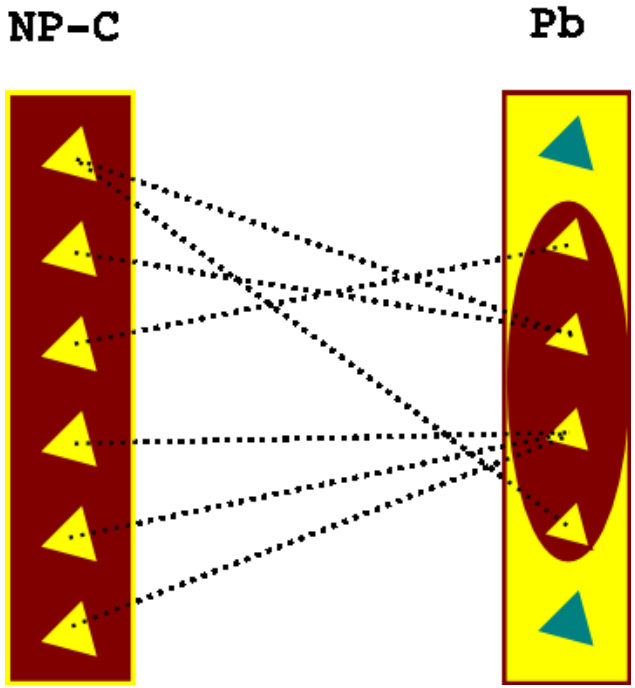
Apparté sur la complexité

- NP-complet (resp. NP-difficile) pour des problèmes de décisions (resp. d'optimisation)
- Définition intuitive :
  - Problème pour lequel la recherche d'une solution consiste à parcourir un arbre *de recherche*
  - Toute solution est représentée par un chemin menant aux feuilles
  - La hauteur d'un tel arbre est polynomiale mais le nombre de ses branches est exponentiel (un noeud = un choix)
  - L'espace des solutions possibles croît exponentiellement en fonction de la profondeur de l'arbre (2^n et 3^n arbre binaire, ternaire resp.)
  - Trouver une solution revient à parcourir l'arbre, ce qui dans le cas le pire peut demander le parcours complet de cet arbre

La NP-complétude

Apparté sur la complexité

- Problèmes NP-complets célèbres
  - Problème SAT et variante 3SAT (mais 2SAT est polynomial)
  - Problème du voyageur de commerce
  - Problème de la clique maximum
  - ...
- Prouver la NP-complétude d'un problème
  - transformation polynomiale
  - équivalence des solutions



Plan

Approches heuristiques

HEURISTIQUE (du grec `heuriskein`, trouver)

- Heuristique en étoile
- Clustal (la plus populaire)
- Dialign2 (complémentaire à Clustal)
- T-coffee, Pima, Multalign, ...

Approches heuristiques

Heuristique en étoile

- Démarche :
  - Sélection d'une séquence de référence (le centre)
  - Construction de l'alignement multiple, en partant de la séquence centrale, puis en incorporant une à une les autres séquences
- Exemple :

S_1 : cgatgagtcattgtgactg
S_2 : cgagccattgtagctactg
S_3 : cgaccattgtagctacctg
S_4 : cgatgagtcactgtgactg

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

Approches heuristiques

Heuristique en étoile

- Étape 1 : Alignements globaux de toutes les séquences 2 par 2

S_1 : cgatgagtcattgt-g--actg
S_2 : cga-g--ccattgtagctactg

S_2 : cgagccattgtagcta-ctg
S_3 : cga-ccattgtagctacctg

S_1 : cgatgagtcattg-tgactg
S_3 : cgacca-ttgtagctacctg

S_2 : cga-g--ccattgtagctactg
S_4 : cgatgagtcactgt-g--actg

S_1 : cgatgagtcattgtgactg
S_4 : cgatgagtcactgtgactg

S_3 : cgaccattgtagctacctg
S_4 : cgatgagtcactgtgactg

	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	

k séquences $\Rightarrow \frac{k(k-1)}{2}$ alignements

Approches heuristiques

Heuristique en étoile

- Étape 2 : Sélection de la séquence de référence à partir du tableau des scores ; i.e. 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

Approches heuristiques

Heuristique en étoile

- Étape 3 : Construction de l'alignement multiple par juxtaposition des alignements 2 à 2 avec la séquence de référence

S_1 : cgatgagtcattgt-g--actg
S_2 : cga-g--ccattgtagctactg

S_1 : cgatgagtcattg-t-g--actg
S_2 : cga-g--ccattg-tagctactg

S_1 : cgatgagtcattg-tgactg
S_3 : cgacca-ttgtagctacctg

S_1 : cgatgagtcattg-t-g--actg
S_3 : cgacca-ttgtagct-ac--ctg

S_1 : cgatgagtcattgtgactg
S_4 : cgatgagtcactgtgactg

S_1 : cgatgagtcactg-t-g--actg
S_4 : cgatgagtcactg-t-g--actg

- L'intégration d'une nouvelle séquence se fait en prenant la séquence de référence comme guide – en étirant les gaps de l'alignement multiple courant.

Approches heuristiques

Le programme Clustal

- Thomson *et al.* en 1994
- Clustal=CLUSTer + ALignment
- Inspiré de 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

Approches heuristiques

Le programme Clustal

- Étape 1 : Alignements globaux de toutes les séquences 2 à 2. Puis, regroupement des séquences suivant leur similarité à partir de la matrice de scores 2 à 2 (ici, UPGMA cf cours sur les phylogénies)

	S_1	S_2	S_3	S_4
S_1		1	0	17
S_2	1		14	0
S_3	0	14		-1
S_4	17	0	-1	

$[(S_1,S_4),\ S_2,S_3]$

Approches heuristiques

Le programme Clustal

- Étape 1 : Alignements globaux de toutes les séquences 2 à 2. Puis, regroupement des séquences suivant leur similarité à partir de la matrice de scores 2 à 2

	S_{14}	S_2	S_3
S_{14}		1	-0,5
S_2	1		14
S_3	-0,5	14	

$[(S_1,S_4), (S_2,S_3)]$

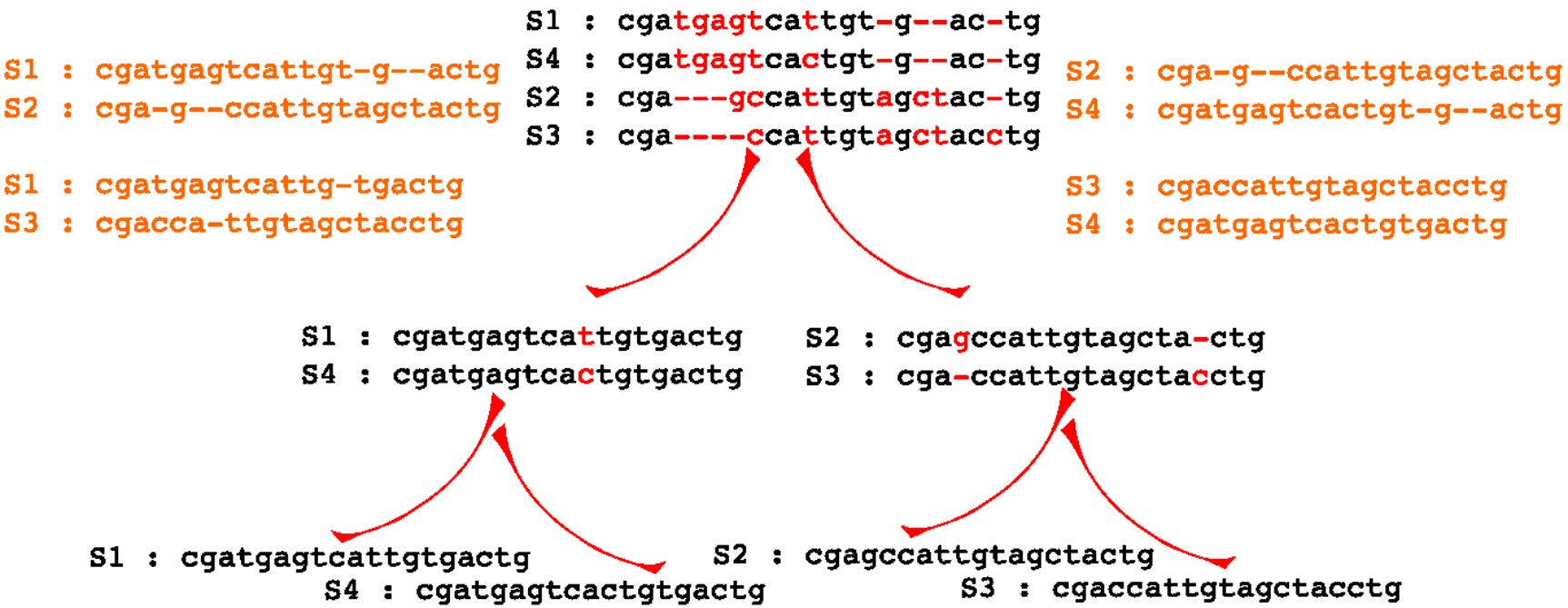
Approches heuristiques

Le programme Clustal

- Étape 2 : Construction de l'alignement à partir de l'arbre guide
  - L'arbre guide correspond à une classification hiérarchique ascendante
  - Alignement entre deux clusters de séquences : alignement 2 à 2 avec les SP pour le score de la colonne
  - L'alignement est obtenu par extensions successives
  - "Once a gap, always a gap"

Approches heuristiques

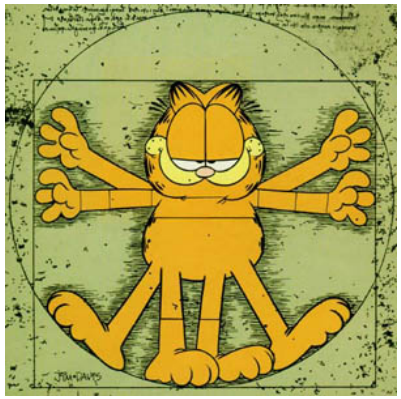
Le programme Clustal



Approches heuristiques

Le programme Dialign2

- Morgenstern *et al.* en 1999
- Dialign = DIagonal + ALIGNment
- Idée : repérer des similarités locales fortes entre les séquences – les diagonales du dotplot
- Incorporer les diagonales dans l'alignement multiple
- Exemple :

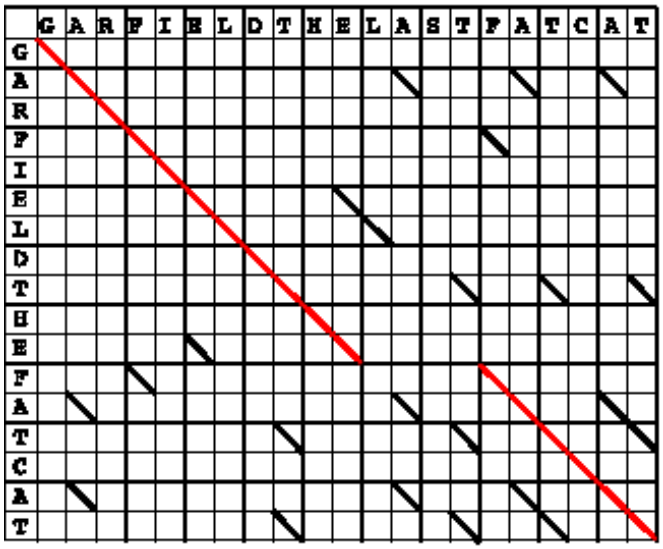


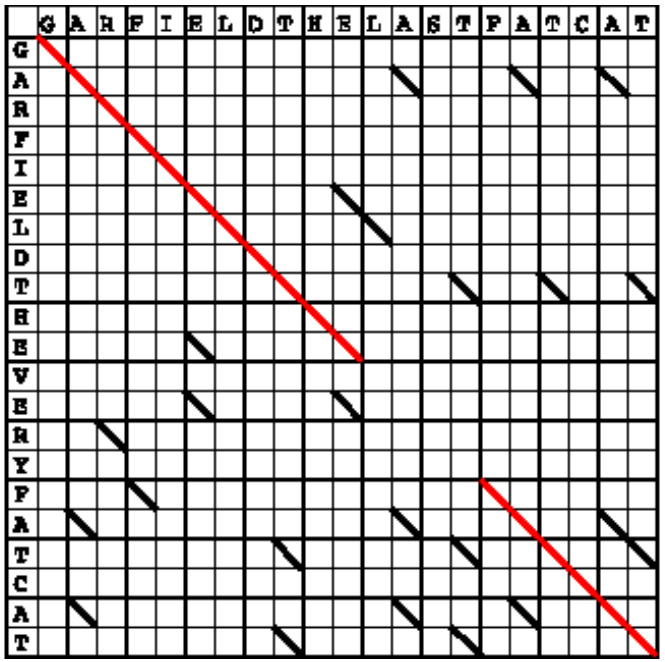
```
GARFIELD THE LAST FAT CAT
GARFIELD THE FAT CAT
GARFIELD THE VERY FAT CAT
THE FAT CAT
```

Approches heuristiques

Le programme Dialign2

- Étape 1 : Calculer les dotplots des alignements 2 à 2 des séquences



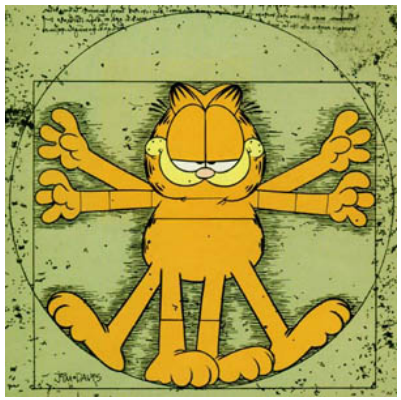


...

Approches heuristiques

Le programme Dialign2

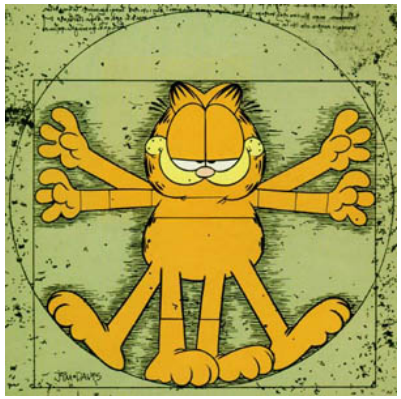
- Étape 1 : Calculer les dotplots des alignements 2 à 2 des séquences
- Étape 2 : Extraire un ensemble de diagonales compatibles dont la somme des scores est maximisée
  - Pas de croisements
  - Pas de chevauchements
  - Score maximal



Approches heuristiques

Le programme Dialign2

- Étape 1 : Calculer les dotplots des alignements 2 à 2 des séquences
- Étape 2 : Extraire un ensemble de diagonales compatibles dont la somme des scores est maximisé
- Étape 3 : Les diagonales restantes sont ordonnées par poids et incorporées une à une tant qu'elles sont compatibles avec l'alignement multiple en construction

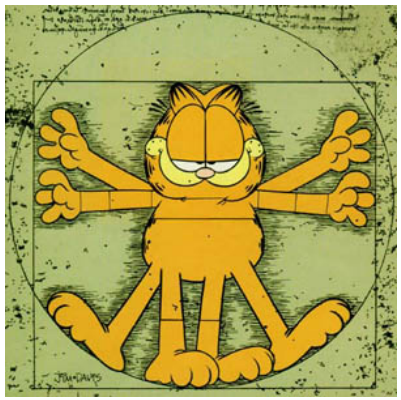


Approches heuristiques

Le programme Dialign2

- Exemple :

```
GARFIELD THE LAST FA-T CAT
GARFIELD THE ---- FA-T CAT
GARFIELD THE VERY FAST CAT
-------- THE ---- FA-T CAT
```



Plan

Représentation graphique

Les principales représentations

- Séquence consensus
- Matrices de positions
- Séquence Logo
- Expression Prosite
- HMM

Représentation graphique

Les séquences consensus

- À chaque position de l'alignement multiple, on retient la lettre majoritaire en considérant le code IUPAC (International Union of Pure and Applied Chemistry)

A		adenine	M	A C	groupe amino
C		cytosine	S	G C	strong
G		guanine	W	A T	weak
T		thymine	B	G T C	pas A
U		uracile	D	G A T	pas C
R	G A	purine	H	A C T	pas G
Y	T C	pyrimidine	V	G C A	pas T
K	G T	groupe keto	N	A G C T	

Représentation graphique

Les séquences consensus

- Exemple :

A	A	M	A C
C	C	S	G C
G	G	W	A T
T	T	B	G T C
U	U	D	G A T
R	G A	H	A C T
Y	T C	V	G C A
K	G T	N	A G C T

```
G C C G G A A G T G
A C C G G A A G C A
G C C G G A T G T A
A C C G G A A G C T
A C C G G A T A T A
C C C G G A A G T G
A C A G G A A G T C
G C C G G A T G C A
T C C G G A A G T A
A C A G G A A G C G
A C A G G A T A T G
T C C G G A A A C C
A C A G G A T A T C
C A A G G A C G A C
T C T G G A C C C T
-----------------
N C M G G A W G Y N
```

Représentation graphique

Matrices de positions

- ligne ⇒ position de l'alignement, colonne ⇒ acide nucléique

```
GCCGGAAGTG
ACCGGAAGCA
GCCGGATGTA
ACCGGAAGCT
ACCGGATATA
CCCGGAAGTG
ACAGGAAGTC
GCCGGATGCA
TCCGGAAGTA
ACAGGAAGCG
ACAGGATATG
TCCGGAAACC
ACAGGATATC
CAAGGACGAC
TCTGGACCCT
```

A	C	G	T	
7	2	3	3	N
1	14	0	0	C
5	9	0	1	M
0	0	15	0	G
0	0	15	0	G
15	0	0	0	A
8	2	0	5	W
4	1	10	0	G
1	6	0	8	Y
5	4	4	2	N

A	C	G	T
0.47	0.13	0.2	0.2
0.07	0.93	0	0
0.33	0.6	0	0.07
0	0	1	0
0	0	1	0
1	0	0	0
0.53	0.13	0	0.33
0.27	0.07	0.67	0
0.07	0.4	0	0.53
0.33	0.27	0.27	0.13

Position Frequency Matrix

Représentation graphique

Des PFM aux PWM

- PWM : Position Weight Matrix
- Poids positif : les bases qui apparaissent plus que la moyenne
- Poids négatif : les bases qui apparaissent moins que la moyenne
- Poids de la base x dans une colonne de l'alignement :

$$\log_2(\frac{f(x)}{0.25})$$

où

- $f(x)$ est la fréquence de x dans la colonne considérée
- 0.25 suppose que les quatre bases ont la même probabilité d'apparition

Représentation graphique

Séquences LOGO

- Représentation graphique de la proportion d'un acide nucléique ou aminé à chaque position
- Plus l'acide est présent à une position donnée, plus haute sera la lettre
- Différents acides à une même position seront représentés par des lettres à différentes échelles en fontion de leurs fréquences

GCCGGAAGTG
ACCGGAAGCA
GCCGGATGTA
ACCGGAAGCT
ACCGGATATA
CCCGGAAGTG
ACAGGAAGTC
GCCGGATGCA
TCCGGAAGTA
ACAGGAAGCG
ACAGGATATG
TCCGGAAACC
ACAGGATATC
CAAGGACGAC
TCTGGACCCT

Représentation graphique

Expression Prosite

- Exemple : hormone pancréatique (PP)

```
NEU CARAU/29-64 AEE..LAKYYSALRHYINLITRQRY
PYY HUMAN/29-64 PEE..LNRYYASLRHYLNLVTRQRY
PMY PETMA/01-36 PEE..LSKYMLAVRNYINLITRQRY
PPY LOPAM/01-36 PED..WASYQAAVRHYVNLITRQRY
PAH BOVIN/30-65 PEQ..MAQYAAELRRYINMLTRPRY
PAH CHICK/26-61 VED..LIRFYNDLQQYLNVVTRHRY
PAH ANSAN/01-36 VED..LRFYYDNLQQYRLNVFRHRY
NPF HELAS/04-39 PNE..LRQYLKELNEYYAIMGRTRF
NPF MONEX/01-39 DNKAALRDYLRQINEYFAIIGRPRF
```

`[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)
  - `[NHG]` : alternative (N, H ou G)

Représentation graphique

HMM

- HMM = Hidden Markov Model = Modèle de Markov caché
- Un ensemble d'états
- Des probabilités de transitions entre les états
- Un ensemble d'observations
- Une probabilité d'émission qui indique pour chaque état la probalité d'y produire une information

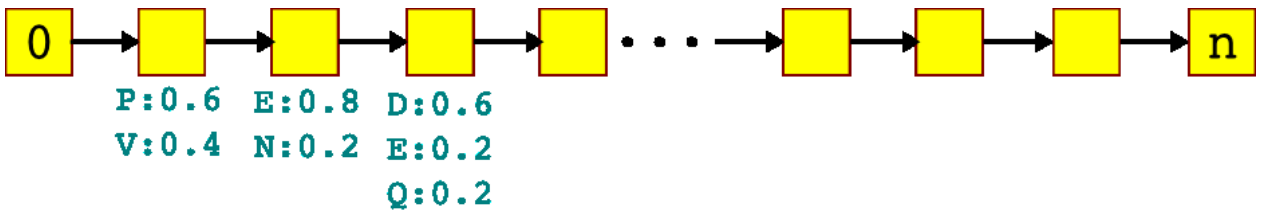
Représentation graphique

Profil HMM

- Si l'alignement n'a pas d'indels
- 1 observation = 1 acide aminé

```
PEDWASYQAAVRHYVNLITRQRY
PEQMAQYAAELRRYINMLTRPRY
VEDLIRFYNDLQQYLNVVTRHRY
VEDLRFYYDNLQQYRLNVFRHRY
PNELRQYLKELNEYYAIMGRTRF
```

- 1 état = 1 colonne de l'alignement multiple
- 1 émission = fréquences des a.a.



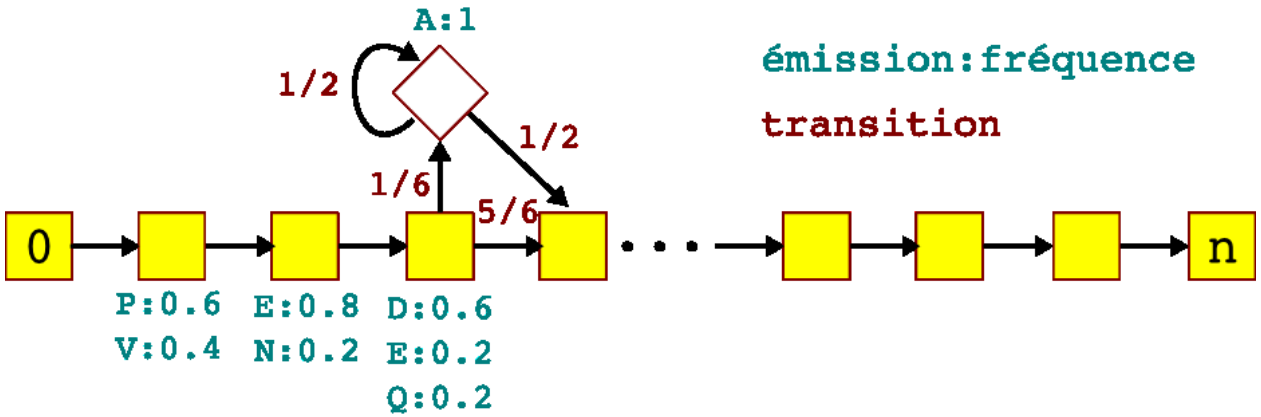
Représentation graphique

Profil HMM

- Avec des insertions – une insertion est un fragment de la séquence qui n'apparait pas dans le modèle

```
PED..WASYQAAVRHYVNLITRQRY
PEQ..MAQYAAELRRYINMLTRPRY
VED..LIRFYNDLQQYLNVVTRHRY
VED..LRFYYDNLQQYRLNVFRHRY
PNEAALRQYLKELNEYYAIMGRTRF
```

- Un nouvel état = un bloc d'insertion



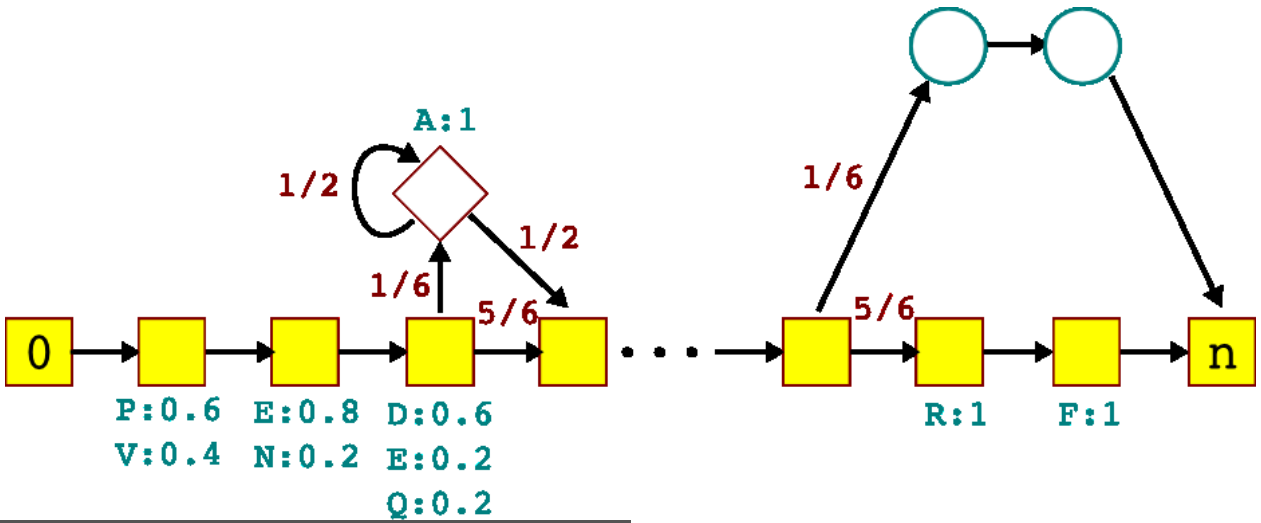
Représentation graphique

Profil HMM

- Avec des délétions – une délétion est un fragment de la séquence qui ne correspond à aucun a.a.

```
PED..WASYQAAVRHYVNLITRQRY
PEQ..MAQYAAELRRYINMLTRPRY
VED..LIRFYNDLQQYLNVVTRHRY
VED..LRFYYDNLQQYRLNVFRHRY
PNEAALRQYLKELNEYYAIMGRT..
```

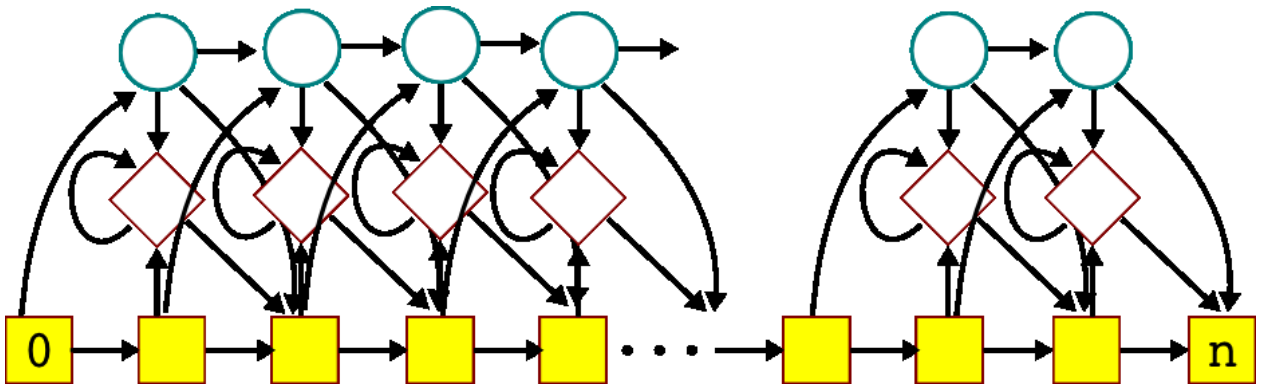
- Option 1 : ajouter des arcs entre les états matchants (nb exp.)
- Option 2 : ajouter un état silencieux ; i.e. qui n'émet rien



Représentation graphique

Profil HMM - en résumé

- États matchants : colonnes avec moins de 50% de –
- États d'insertion : majorité de –
- États de délétion : minorité de –
- Probabilités d'émission : on compte le nombre d'occurrences de chaque acide aminé
- Probabilités de transition : on compte le nombre de séquences empruntant la transition



Plan

Bibliographie

Références

1. Duret, L. and S. Abdeddaim. *Multiple alignment for structural functional or phylogenetic analyses of homologous sequences*, in D. Higgins and W. Taylor : Bioinformatics sequence structure and databanks. Oxford : Oxford University Press, 2000.
2. Notredame, C.. *Recent progresses in multiple sequence alignment : a survey*. Pharmacogenomics 31 (1) : 131 – 144, 2002.
3. J. Setubal and J. Meidanis. *Introduction to Computational Molecular Biology*. PWS Publishing Co, 1997.
4. M. Crochemore, C. Hancart and T. Lecroq. *Algorithmique du texte*. Vuibert, 2001.