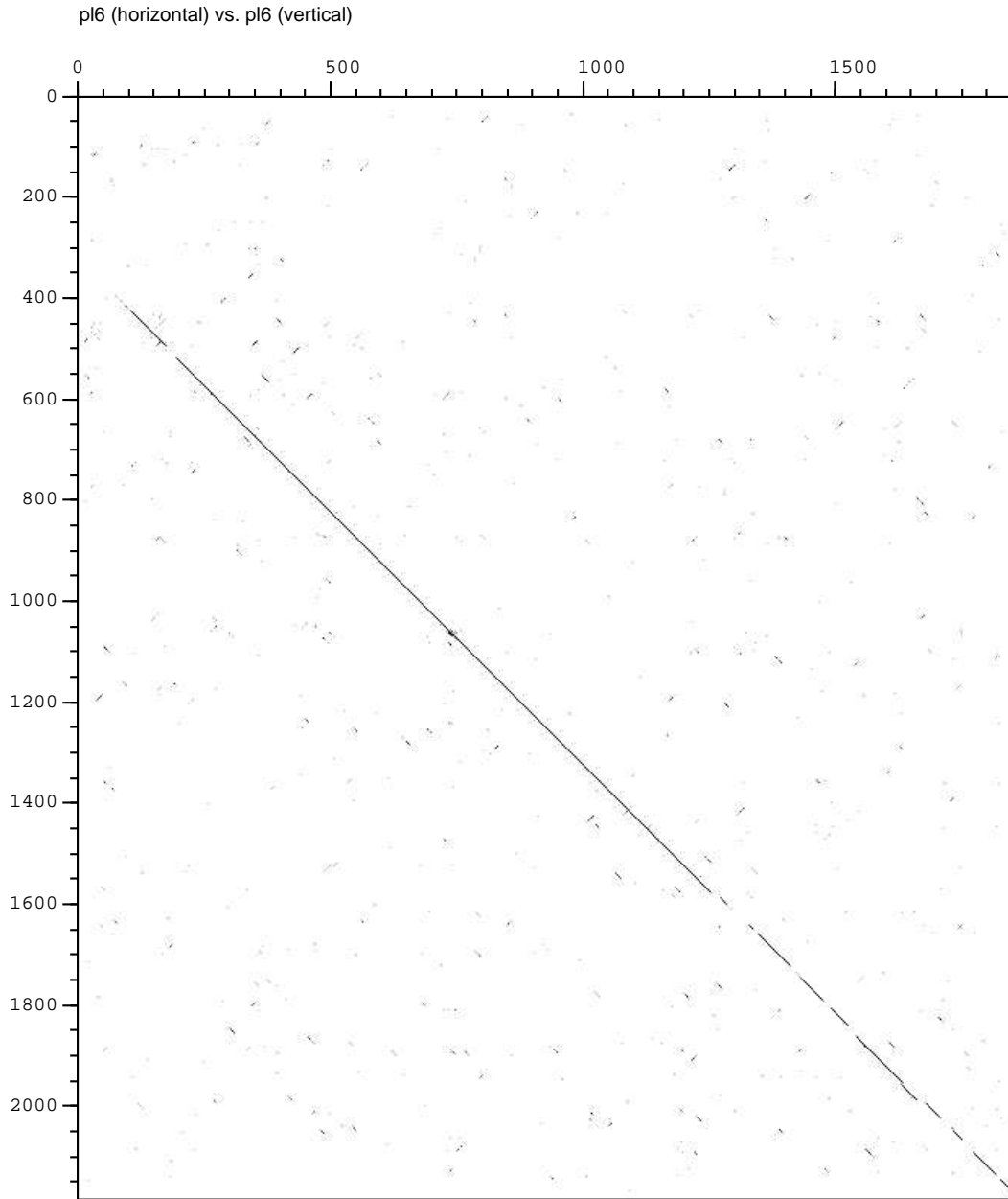


## Cours 3

Alignement local, BLAST



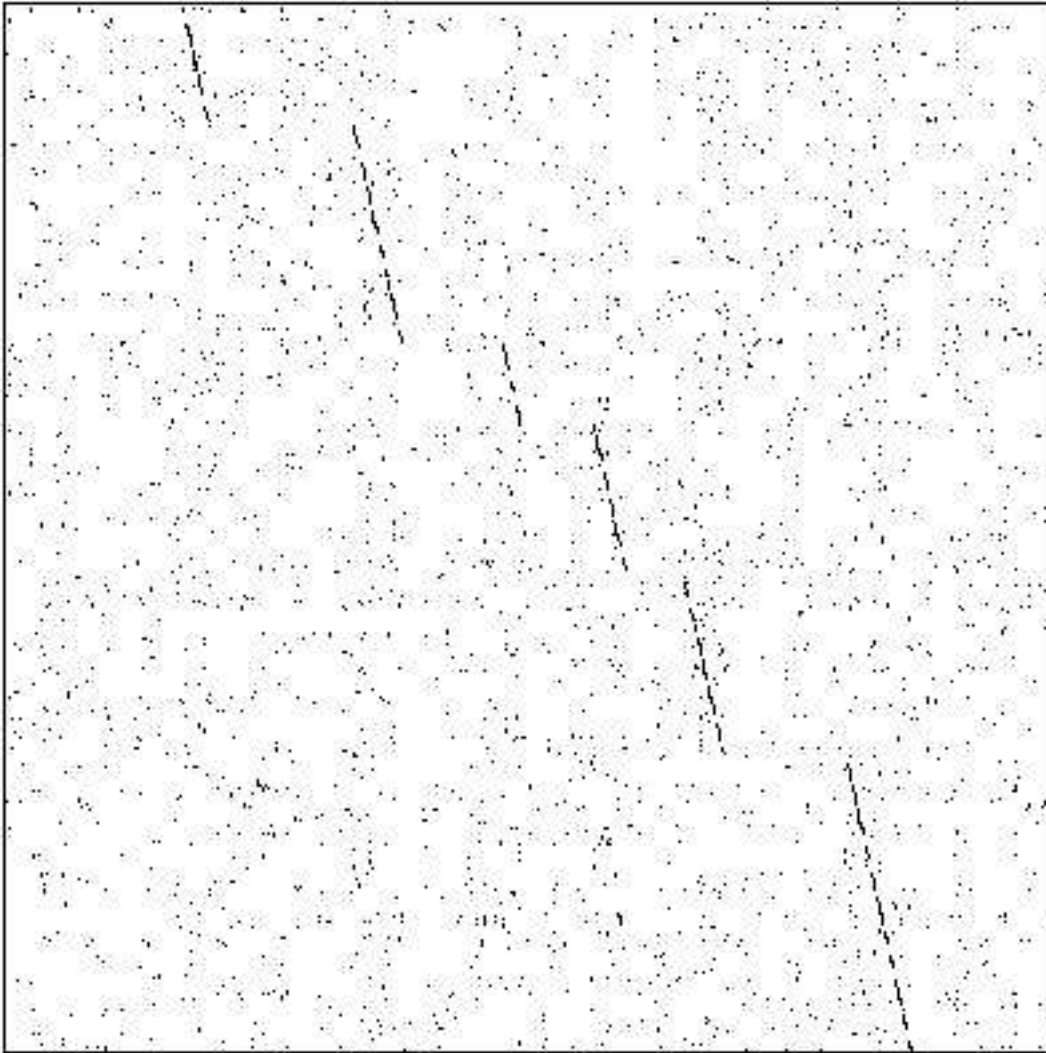
*horizontalement* : gène pl6 chez l'homme

*verticalement* : gène pl6 chez la souris



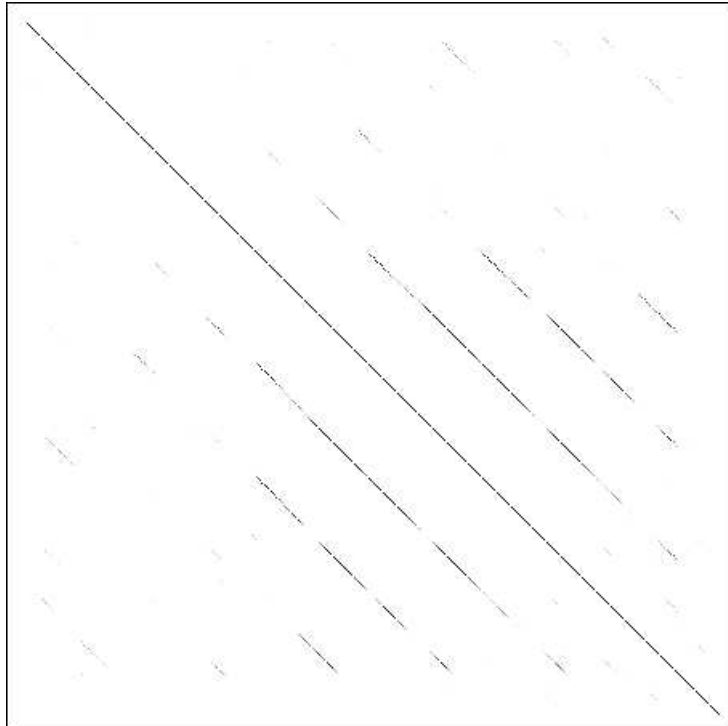
*horizontalement* : ADN codant pour la chaîne  $\alpha$  de l'hémoglobine humaine

*verticalement* : ADN codant pour la chaîne  $\beta$  de l'hémoglobine humaine

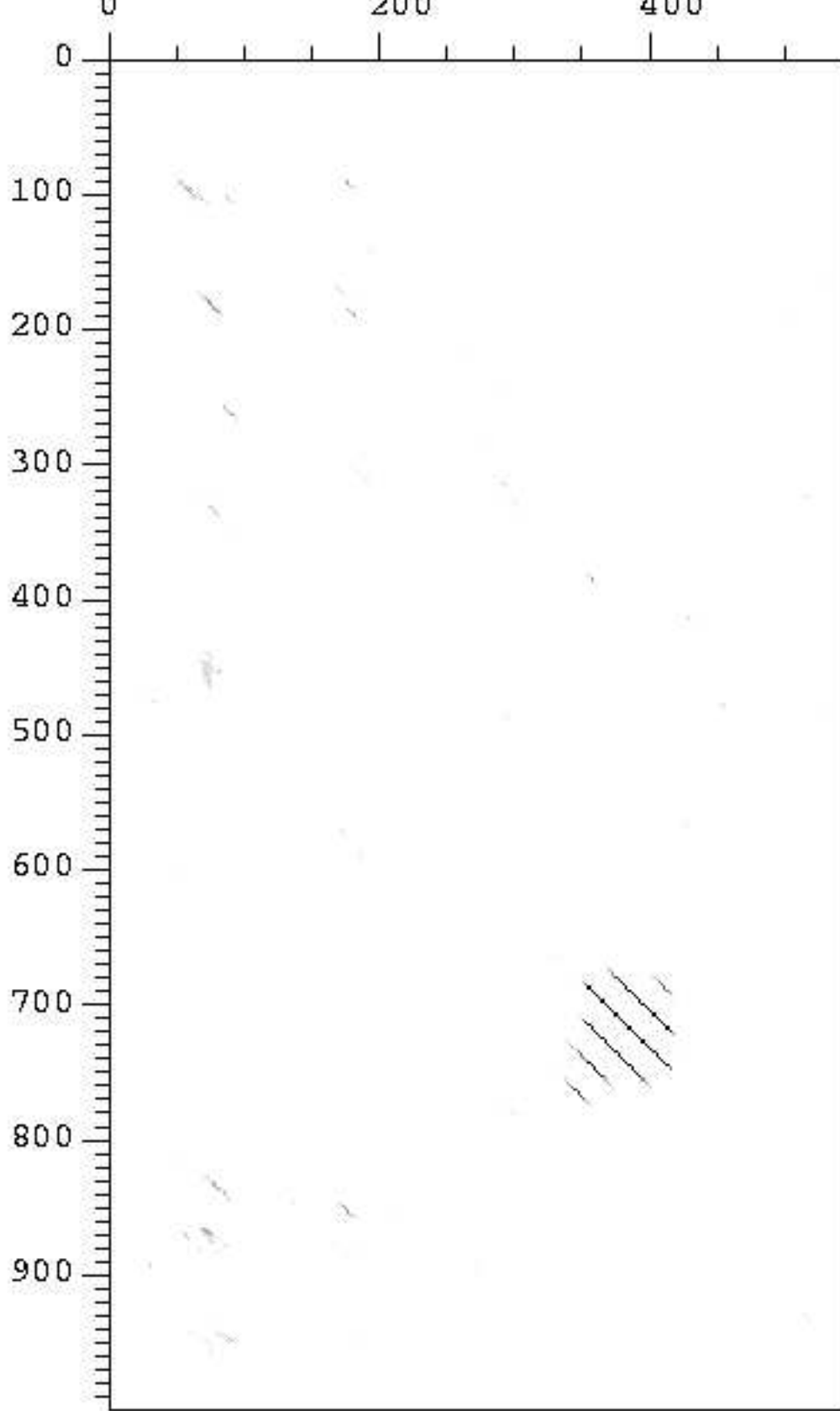


*horizontalement* : séquence nucléaire du gène de l'actine de muscle de *Pisaster ochraceus*

*verticalement* : cDNA de ce même gène



Comparaison de la protéine ribosomique S1 de  
*Escherichia Coli* sur elle-même.



*horizontalement* : early growth response protein 1 (homme)

*verticalement* : protein AZF1 (levure)

## Limites de l'alignement global

**Exemple :** GGCTGACCACCTT et GATCACTTCCATG

Alignement global :

```
1  G G C T G A C C A C C _ T T 13
   |     |  | |     | |  | |
1  G A _ T C A C T T C C A T G 13
```

*Les séquences présentent une similarité que l'alignement global ne révèle pas.*

Résultat souhaité :

```
5  G A C C A C C T T 13
   | |  | | |  | |
1  G A T C A C _ T T 8
```

# Alignement local

*Smith & Waterman - 1981*

## Données

- ▷ deux séquences (nucléotides ou acides aminés),
  - ▷ des scores de similarité.
- 

## Problème

Quelle sont les régions de forte similarité entre les deux séquences ?

$Loc(i, j)$  : score optimal entre un suffixe de  $U(1..i)$  et un suffixe  $V(1..j)$ .

### Formule de récurrence :

$$\left| \begin{array}{l} Loc(0, 0) = 0 \\ Loc(0, j) = 0 \\ Loc(i, 0) = 0 \\ \\ Loc(i, j) = \max \left\{ \begin{array}{l} Loc(i-1, j-1) + Sub(U(i), V(j)) \\ Loc(i-1, j) + Del(U(i)) \\ Loc(i, j-1) + Ins(V(j)) \\ 0 \end{array} \right. \end{array} \right.$$

### Implémentation, complexité :

*cf* alignement global : programmation dynamique

	G	G	C	T	G	A	C	C	A	C	C	T	T
G	2	1	0	0	2	1	0	0	0	0	0	0	0
A	1	1	0	0	0	4	3	2	2	1	0	0	0
T	0	0	0	2	1	0	3	2	1	0	0	2	2
C	0	0	2	1	1	0	2	5	4	3	2	1	1
A	0	0	1	1	0	3	2	4	7	6	5	4	3
C	0	0	2	1	0	2	5	4	6	9	8	7	6
T	0	0	1	4	3	2	4	4	5	8	8	10	9
T	0	0	0	3	3	2	3	3	4	7	7	10	12
C	0	0	2	2	2	2	4	5	4	6	6	9	11
C	0	0	2	1	1	1	4	6	4	6	8	8	10
A	0	0	1	1	0	3	3	5	8	7	7	7	9
T	0	0	0	3	2	2	2	4	7	7	6	9	11
G	2	2	1	2	5	4	3	3	6	6	6	8	10

- score maximal
- les scores qui correspondent à des optimaux locaux
- tous les scores au dessus d'un seuil de similarité

# BLAST

*Basic Local Alignment Search Tool*

*Altschul et al. - 1997*

- ▷ Programme pour la recherche de similarités dans les bases de données
- ▷ Utilise un algorithme heuristique linéaire pour l'alignement local
- ▷ Séquences nucléiques et protéiques
- ▷ Disponible sur le Web
- ▷ Connecté aux principales banques de données

## Algorithme mis en œuvre dans BLAST

- ▷ Ne s'intéresse qu'aux séquences avec un fort taux de similarité

BLAST ne construit pas un alignement avec toutes les séquences de la banque.

- ▷ Tire parti du caractère biologique des séquences:

*des séquences similaires ont des segments communs de taille  $k$  quasi-identiques*

Par défaut : ADN  $k=11$  ou  $13$ , protéines  $k=3$

- ▷ Conçu pour traiter de grandes banques de données

## Étape 0:

Localisation dans la banque de données de tous les mots de longueur  $k$

$4^{11} = 4\,194\,304 \ll$  taille de la banque de données

## Étape 1:

Construction d'une table de hachage recensant tous les mots de longueurs  $k$  apparaissant dans la séquence requête avec un score  $> T$ .

→ **k-mers**

$T$  est déterminé par la configuration de la machine.

Fonction de hachage (pour l'ADN):

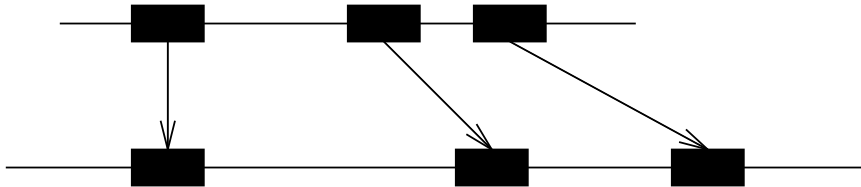
$e : \{A, C, G, T\} \rightarrow \{0, 1, 2, 3\}$

$$\mathcal{H}(i) = \sum_{j=0}^{k-1} e(a_{i+j})4^{k-j-1}$$

$$\mathcal{H}(i+1) = 4 \times \mathcal{H}(i) + e(a_{i+k}) \pmod{4^k}$$

## Étape 2:

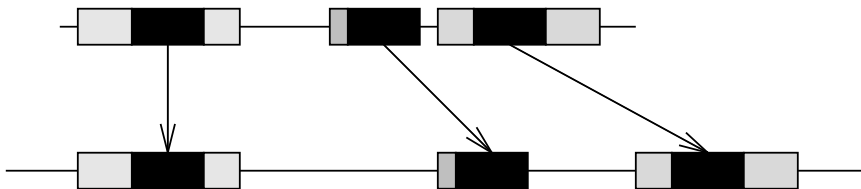
Localisation des k-mers dans la banque de données



→ HSP : *High Scoring Pairs*

## Étape 3 :

Extension de ces points d'ancrage de proche en proche, pour avoir un score **significatif**



Query= Felis catus DRD4 gene fordopamine receptor D4  
(276 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences  
1,174,453 sequences; 5,001,591,585 total letters

Sequences producing significant alignments:		Score	E
		(bits)	Value
gi AB069665	Felis catus DRD4 gene f...	210	5e-52
gi AB069662	Nyctereutes procyonoide...	157	7e-36
gi AB069661	Canis lupus DRD4 gene f...	157	7e-36
gi AB069666	Bos taurus DRD4 gene fo...	143	1e-31
gi 291947	Homo sapiens Dopamine D4 recep...	135	2e-29

>gi|18143632|dbj|AB069662.1|AB069662 Nyctereutes procyonoides  
DRD4 gene fordopamine receptor D4. Length = 393

Score = 157 bits (79), Expect = 7e-36  
Identities = 94/99 (94%)  
Strand = Plus / Plus

Query 1 ttcttcctaccctgcccgctcatgctgctgctctactgggccacgttcc 48  
|||||  
Sbjct 1 ttcttcctaccctgcccgctcatgctgctgctctactgggccacgttcc 48

Query 49 ggggcctgcggcgctgggaggcggctcgccaggccaagctgcactgccgg 99  
|||||  
Sbjct 49 ggggcctgcggcgctgggaggcggctcgccaggccaagctgcactgccgg 99

Score = 107 bits (54), Expect = 5e-21  
Identities = 60/62 (96%)  
Strand = Plus / Plus

Query 215 ggaggcgcgccaagatcaccggccgggagcgcaaggccatgagggtcct 252  
||||  
Sbjct 332 ggagacgcgccaagatcacgggccgggagcgcaaggccatgagggtcct 379

Query 253 tgccggtggtggtc 276  
|||||  
Sbjct 380 tgccggtggtggtc 393

>gi|AB032908 Hylobates pileatus gene for dopamine receptor D4,  
partial cds, drd4, 7-repeat allele. Length = 507

Score = 42.1 bits (21), Expect = 0.27  
Identities = 45/53 (84%)  
Strand = Plus / Plus

52 ggcctgcggcgctgggaggcggctcgccaggccaagctgcactgccgggcgcc 104  
|||||  
4 ggcctgcagcgctgggagggtggcacgtcgcgccaagctgcacggccgcgcgcc 56

```

1      ttcttcctaccctgcccgctcatgctgctgctctactgggccacg 45
      |||||||||||||||||||||||||||||||||||||||||||||
1      ttcttcctaccctgcccgctcatgctgctgctctactgggccacg 45

46     ttccggggcctgcggcgctgggaggcggctcgccaggccaagctg 90
      |||||||||||||||||||||||||||| || || | |||||||||
46     ttccggggcctgcggcgctgggaggccgcgcgctcgggccaaagctg 90

91     cactgccggggcgccctcgtcggcccagcggccccggcccaccgccc 135
      ||| ||||| | || || | ||||||||||||||||||| || |||
91     cacggccggacaccgcgcagaccagcggccccggcccgccacc 135

136    cccga.ggt.....c 144
      ||||| ||| |
136    cccgacggtacccccggccccccgcccccgacggcagccccgac 180

145    ggcgagc..... 151
      ||| |||
181    ggc.agcccggacggcacccccggcccgccgcccccgacggcac 224

152    .....c 152
      |
225    cccgatgacacccccgacgccaccctgcccccgcccccgcc 269

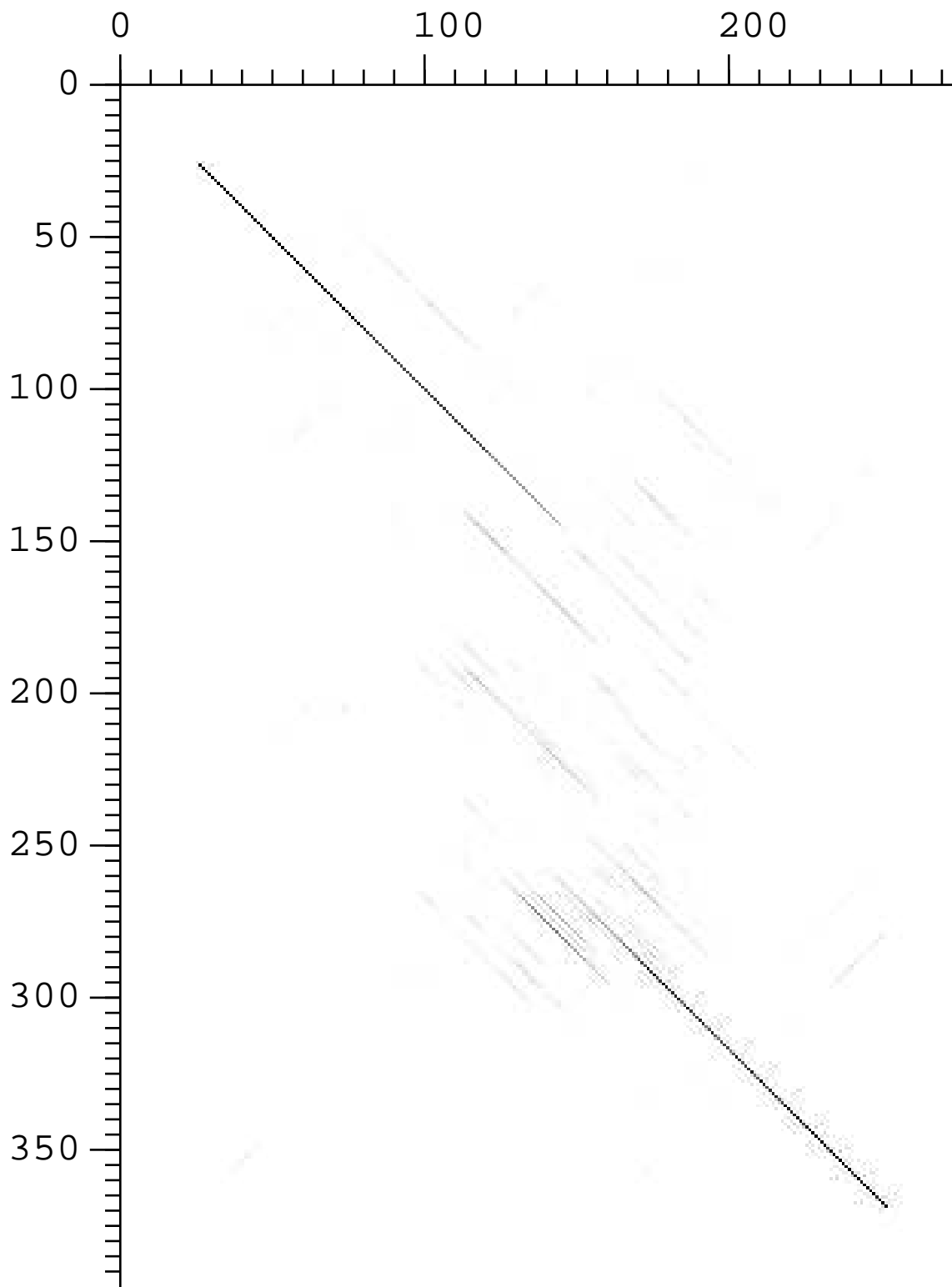
153    cccgacgccgtcgcgccccccgacgccgtcccagccgagccgcc 197
      ||||||||||| ||||||||||| ||||| ||| || ||||| ||
270    cccgacgccgcccgcgccccccgcccggaccctgcggagcccc 314

198    gcggcaggcaccaggaggaggcgcgccaagatcacgggccggga 242
      | ||||| ||| || | |||| | ||||||||||||| |||||||
315    gtggcagccacgcaagcggagacgcgccaagatcacgggccggga 359

243    gcgcaaggccatgagggtcctgccggtggtggtc 276
      ||||||||||||||||||||||||||||||||||||||||
360    gcgcaaggccatgagggtcctgccggtggtggtc 393

```

AB069665. (horizontal) vs. AB069662. (vertical)



# Comment évaluer la pertinence d'un score ?

```
A C C T G A C G T A A G C
| | | | | | | | | | | |
A C C T G A C G T A A G C
```

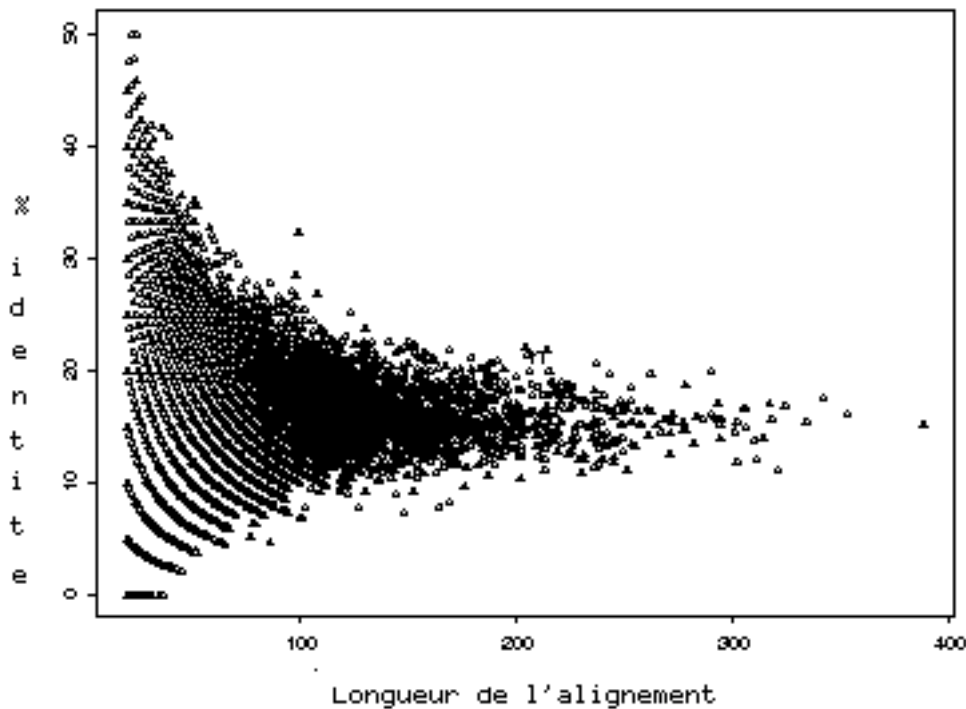
```
A C T T G A C G T - A G C
| |   | | | | |   | | |
A C C T G A C G T A A G C
```

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

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

# Le pourcentage d'identité

- ▷ dépend de la composition en bases, ou acides aminés
- ▷ dépend de la longueur des séquences



Pourcentage d'identité dans des alignements locaux optimaux en fonction de la taille de l'alignement (protéines)

- ▷ comment comparer les % d'alignements de longueur différente ?

*Une fausse bonne idée*

## Approche empirique

### Test de la robustesse du score

$S$  : score de l'alignement entre  $U$  et  $V$

1. Génération de 100 (200, 1000, ...) permutations aléatoires de  $V$   
(même longueur , même composition)
2. Alignements avec  $U$
3. Distribution des scores d'alignement  
Où se situe  $S$  dans cette distribution ?

et ACCTGACGTAAGC

```

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

```

Score : 16

score	s-w	
0	0	:
4	138	:=====
8	166	:=====
12	146	:=====
16	33	:=====
20	7	:==
24	8	:==
28	2	:
32	0	:
36	0	:
40	0	:
44	0	:
48	0	: 500 séquences aléatoires

**Approche statistique : *E-value***

Nombre de fois moyen de trouver un alignement de score supérieur à *S* entre deux séquences de longueur *n* et *m*.

Query= actgagcatag (11 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences  
1,174,453 sequences; 5,001,591,585 total letters

No significant similarity found.

Query= actgagcatagctgga (16 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences  
1,174,453 sequences; 5,001,591,585 total letters

Sequences producing significant alignments:	Score (bits)	E Value
gi AC021642.14  Mus musculus chromosome 10 ...	32	1.0
gi AC079858.8  Homo sapiens BAC clone...	30	4.0
gi AC090032.2  Canis familiaris clone...	30	4.0
gi AF289076 Homo sapiens chromosome 8...	30	4.0
...		

### ALIGNMENTS

>gi|AC021642.14| Mus musculus chromosome 10 clone.  
Length = 203839

```

Query: 1      actgagcatagctgga 16
           |||
Sbjct: 195114 actgagcatagctgga 195129

```

>gi|16973779|gb|AC079858.8| Homo sapiens BAC clone  
Length = 82719

```

Query: 1      actgagcatagctgg 15
           |||
Sbjct: 48150 actgagcatagctgg 48164

```

Query= actgagcatagctggac (17 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences  
1,174,453 sequences; 5,001,591,585 total letters

Sequences producing significant alignments:	Score (bits)	E Value
gi AC021642.14  Mus musculus chromosome ...	34	0.25
gi AL121894.26  Human DNA sequence fro...	32	1.0
gi AC079858.8  Homo sapiens BAC clone ...	30	4.0
gi AC090032.2  Canis familiaris clone...	30	4.0
....		

### ALIGNMENTS

>gi|AC021642.14| Mus musculus chromosome 10 clone  
Length = 203839

Query: 1 actgagcatagctggac 17  
|||||  
Sbjct: 195114 actgagcatagctggac 195130

>gi|9944239| Human DNA sequence. Length = 141079

Query: 2 ctgagcatagctggac 17  
|||||  
Sbjct: 21064 ctgagcatagctggac 21079

>gi|AC079858.8| Homo sapiens BAC clone. Length = 82719

Query: 1 actgagcatagctgg 15  
|||||  
Sbjct: 48150 actgagcatagctgg 48164

Query= actgagcatagctggat (17 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences  
1,174,453 sequences; 5,001,591,585 total letters

Sequences producing significant alignments:			Score	E
			(bits)	Value
gi AC021642.14	Mus musculus chromosome 10 c...		32	1.0
gi AC079858.8	Homo sapiens BAC clone ...		30	4.0
gi AF07784 1	Sulfolobus solfataricus ...		30	4.0
gi AC090032.2	Canis familiaris clone ...		30	4.0
...				

### ALIGNMENTS

>gi|AC021642.14| Mus musculus chromosome 10 clone  
Length= 203839

```

Query: 1      actgagcatagctgga 16
           |||
Sbjct: 195114 actgagcatagctgga 195129

```

>gi|AC079858.8| Homo sapiens BAC clone  
Length = 82719

Score = 30.2 bits (15), Expect = 4.0

```

Query: 1      actgagcatagctgg 15
           |||
Sbjct: 48150  actgagcatagctgg 48164

```

**E-value  $\leq$  1000**

Query= actgagcatag (11 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences

1,174,453 sequences; 5,001,591,585 total letters

Sequences producing significant alignments:		Score	E
		(bits)	Value
gi AC092378.3	Homo sapiens chromosome 16 clo...	22	967
gi NM_131197.1	Danio rerio endothelin recept...	22	967
gi AC084013.5	Homo sapiens BAC clone ...	22	967
gi AC092203.15	Mus musculus clone rp23-422n18...	22	967
gi AP000003	Pyrococcus horikoshii OT...	22	967
gi 10727456	Drosophila melanogaster ...	22	967
...			

### ALIGNMENTS

>gi|AC092378.3| Homo sapiens chromosome 16 clone  
Length = 199869

Query: 1 actgagcatag 11  
          |||||  
Sbjct: 78821 actgagcatag 78811

# Génome de la drosophile

Query= actgagcatag (11 letters)

Database: D. melanogaster genomic nucleotide sequences  
1170 sequences; 122,655,632 total letters

Sequences producing significant alignments:	Score (bits)	E Value
gi AE002770 Drosophila melanogaster g...	22	24
gi AE003609 Drosophila melanogaster g...	22	24
gi AE003450 Drosophila melanogaster g...	22	24
gi AE003426 Drosophila melanogaster g...	22	24
gi AE003484 Drosophila melanogaster g...	22	24
...		

## ALIGNMENTS

>gi|7289299|gb|AE002770.1|AE002770 Drosophila melanogaster  
genomic scaffold 142000013385552

```
Query: 1      actgagcatag 11
          |||
Sbjct: 17834 actgagcatag 17844
```