

Comparaison de séquences

Hélène TOUZET

`touzet@lifl.fr`

Les banques de données nucléiques

*International Nucleotide Sequence Database
Collaboration*

▷ trois partenaires



▷ contributions : chercheurs et programmes de séquençage

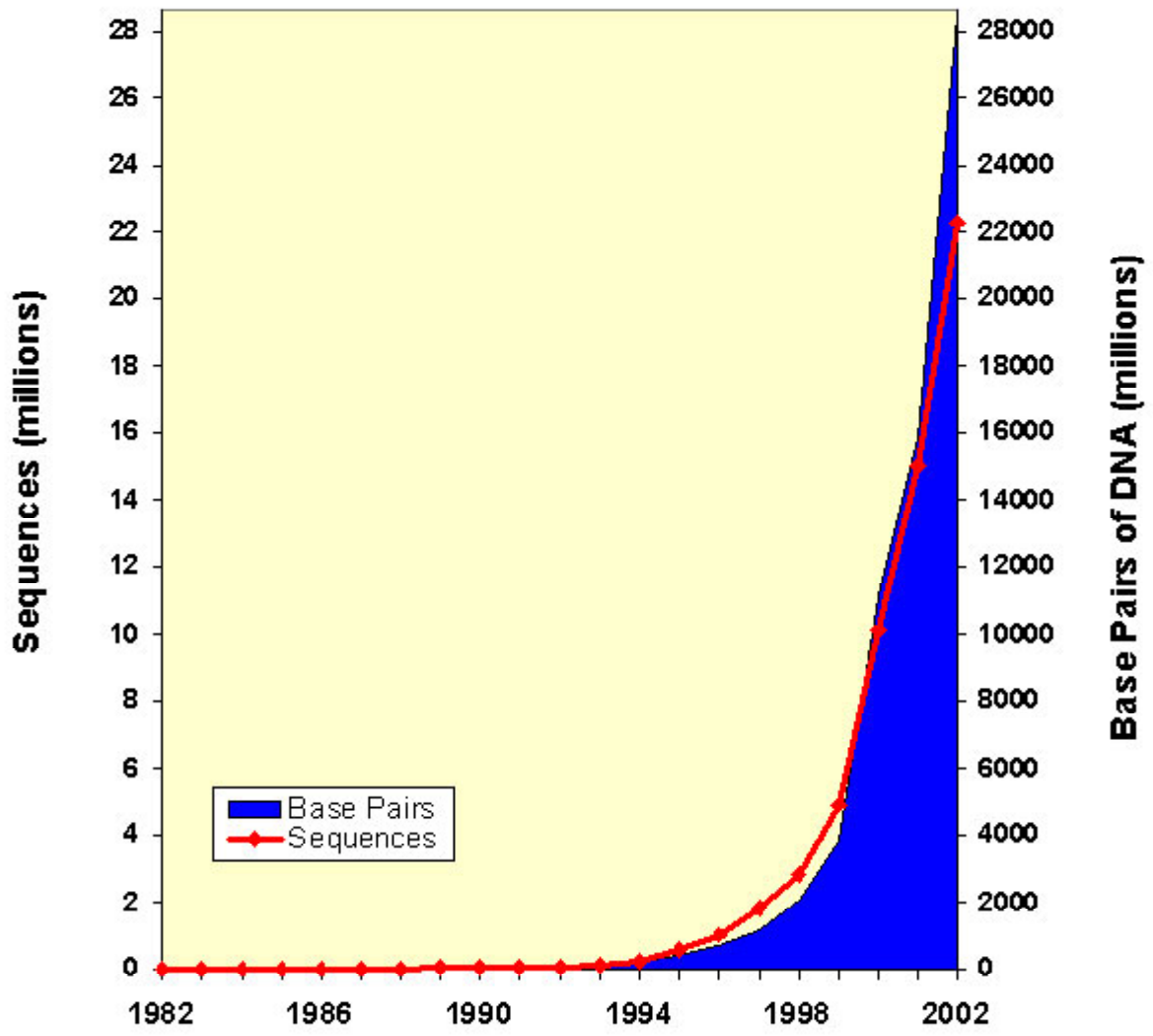
▷ mêmes données, mêmes formats

▷ taxonomie commune

▷ mises à jour quotidiennes

▷ 23 783 987 entrées, 36 903 870 364 bases
(21 février 2003)

Growth of GenBank



Quel type de séquences ?

- ▷ des ARNm
- ▷ régions 5'UTR
- ▷ des EST
- ▷ des clones

▷ des génomes

1978 : séquence du phage phiX174
(premier génome à ADN, 5386 bp)

Virus : plusieurs centaines

Bactéries :

1995 : Haemophilus Influenzae

1996 : Bacillus Subtilis

1996 : Escherichia Coli

Aujourd'hui : environ 100 génomes

Eucaryotes :

1990 : lancement du programme international de séquençage *Génome Humain*

1996 : levure (premier eucaryote)

1998 : Caenorhabditis Elegans (pluri-cellulaire)

2000 : Arabidopsis Thaliana (premier génome de plante)

2000 : brouillon du génome humain

Aujourd'hui : souris, drosophile, rat, zebra fish,
maïs

Quelles informations ?

▷ présentation générale

LOCUS HSA000073 17484 bp DNA linear PRI 09-DEC-2002

DEFINITION TPA: Homo sapiens lymphocyte-specific protein tyrosine kinase (lck) gene baseline reference (distal and proximal promoters, exon 1, 1'-12, 3' UTR).

ACCESSION BN000073

VERSION BN000073.1 GI:28317392

KEYWORDS Third Party Annotation; TPA; LCK gene; protein tyrosine kinase.

SOURCE Homo sapiens (human)

ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata;
Vertebrata; Euteleostomi; Mammalia; Eutheria;
Primates; Catarrhini; Hominidae; Homo.

▷ références bibliographiques

REFERENCE 1

AUTHORS Nervi,S., Nicodeme,S., Gartoux,C., Atlan,C.,
Lathrop,M., Reviron,D., Naquet,P., Matsuda,F.,
Imbert,J. and Vialettes,B.

TITLE No association between lck gene polymorphisms
and protein level in type 1 diabetes

JOURNAL Diabetes 51 (11), 3326-3330 (2002)

MEDLINE 22289034

...

Medline : base de données bibliographique

- 4500 revues depuis 1960: santé, biologie, biotechnologies
- Accessible via **Pubmed**

▷ Annotation de la séquence

```

FEATURES             Location/Qualifiers
source               1..17484
                    /organism="Homo sapiens"
                    /db_xref="taxon:9606"
                    /note="baseline reference generated from
                    the working draft sequence AL121991.33"
promoter            <3319..4011
                    /note="type II distal promoter"
promoter            <4533..5218
                    /note="type I proximal promoter"
gene                5428..16817
                    /gene="LCK"
CDS                 join(5428..5532,5835..5916,6091..6181,
                    6419..6517, 6667..6770,7012..7161,7435..7587,
                    7705..7884,10772..10848, 10942..11095,
                    11180..11311,16615..16817)
                    /gene="LCK"
                    /function="immune response"
                    /codon_start=1
                    /product="protein tyrosine kinase"
                    /protein_id="CAD55807.1"
                    /db_xref="GI:28317393"
                    /translation="MGCGCSSHPEDDWMENIDVCENCHYPIVP
                    ...
                    EELYQLMRLCWKERPEDRPTFDYLRSVLEDDFTATEGQYQPQP"
exon                5428..5532
                    /gene="LCK"
                    /number=1
intron              5533..5834
                    /gene="LCK"
                    /number=1
                    ...
polyA_signal        17240..17245
polyA_site          17261
                    /evidence=experimental

```

▷ la séquence elle-même

BASE COUNT 4014 a 4351 c 4808 g 4211 t 100 others

ORIGIN

```
  1 ccctggggct actgaaattc caggcagtgg gtgaagagga cgaggaggat gaggaggggg
 61 agagcctgga ctctgtgaag gcactgacag ccaagctgca gctgcagact cggcggccct
121 catatctgga gtggacagcc caggtccaga gccaggcctg gcgcagggcc caagccaaac
    . . .

4021 cccaggctcc ctagggatgc agcagccctt tgtggctggg gagagaagat cctcgctcaa
4081 ggtcagnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn
4141 nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnnnnnn nnnnnngggt gtgtgcatga
4201 atgtgtgtgt gggtagctgt gtgagagtgg gtgcctgtgt gtgggggggt agtgtgtgtg
4261 gtgggggggc acttgtggag ggtgagtgta tgtgtttact gagtgtgagt gtgggtgcct
4321 gtgtgtggga gggtagtct gtgtgtgagt gtgtggggga gtacctgtga ggggtgagt
    . . .

5401 gccccctctt ccattccctc agggaccatg ggctgtggct gcagctcaca cccggaagat
5461 gactggatgg aaaacatcga tgtgtgtgag aactgccatt atcccatagt cccactggat
5521 ggcaagggca cggtaagagg cgagacaggg gccttgggtga gggagttagg tagagaatgc
5581 aaccaggag aaagaaatga ccagcactac aggccttga aagaatagag tggccctctc
    . . .

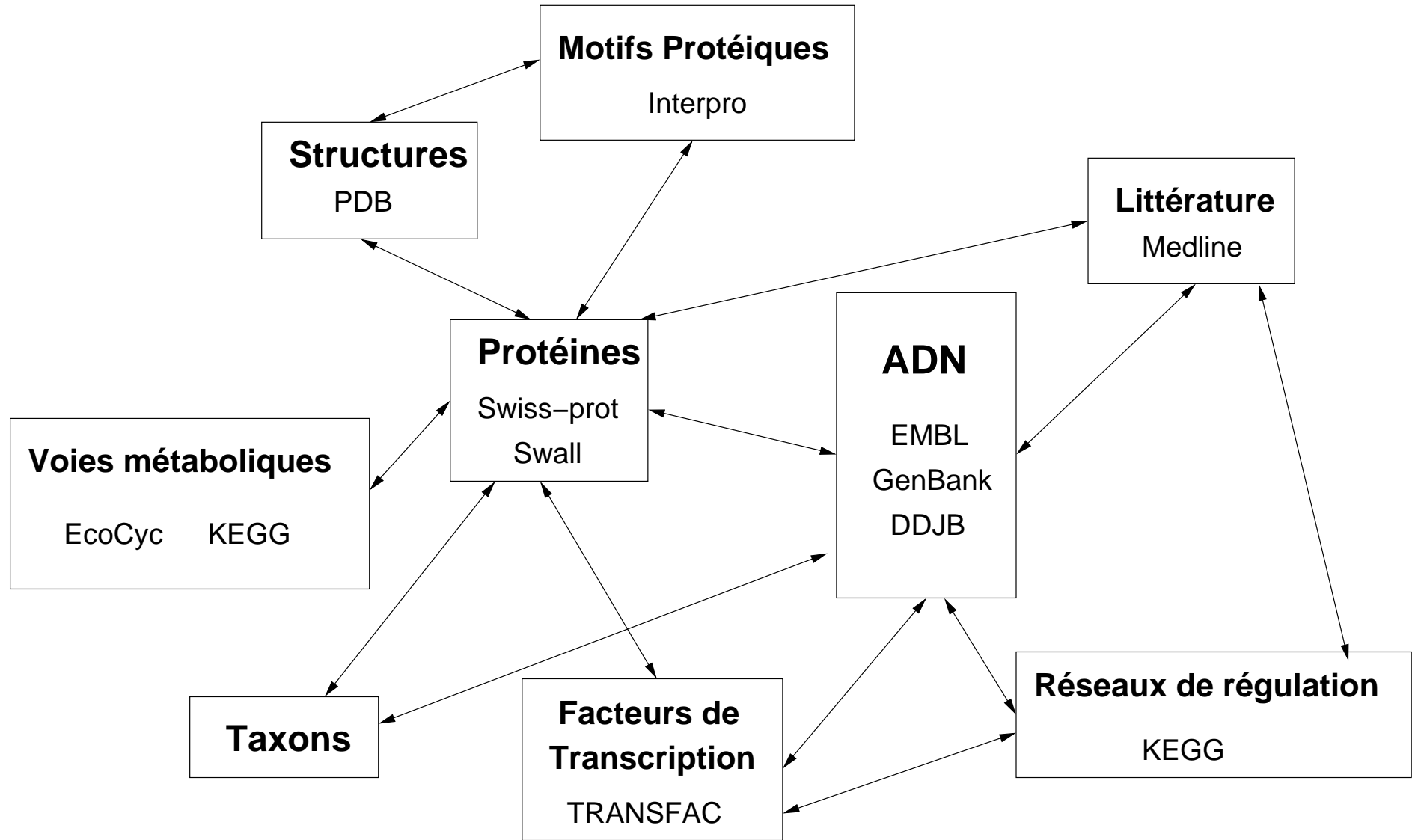
17281 ccactctttg tgggtgggca gtgggggtta agaaaatggt aattaggtca ccctgagttg
17341 gggtgaaaga tgggatgagt ggatgtctgg aggctctgca gacccttca aatgggacag
17401 tgctcctcac ccctcccaa aggattcagg gtgactccta cctggaatcc cttagggaat
17461 gggtagctca aaggaccttc ctcc
```

Exemple 2 : opéron

```
source          1..9430
                /organism="Lactococcus sp."
                /strain="MG1234"
-35_signal      160..165
                /gene="galA"
                /evidence=EXPERIMENTAL
-10_signal      179..184
                /gene="galA"
                /evidence=EXPERIMENTAL
CDS             405..1934
                /gene="galA"
                /product="galactose permease"
                /function="galactose transporter"
                /evidence=EXPERIMENTAL
CDS             2003..3001
                /gene="galM"
                /product="aldose 1-epimerase"
                /EC_number="5.1.3.3"
                /function="mutarotase"
CDS             3235..4537
                /gene="galK"
                /product="galactokinase"
                /EC_number="2.7.1.6"
                /evidence=EXPERIMENTAL
```

Format FASTA

```
> Mus musculus, carbonic anhydrase 14, mRNA
GAGATTACGAAGGGACATACGGAGAGGGCAGAGAGAGGAAGAG
AGAGAAAGTGAGAGAGGAAGAAATTTATCCTGGGATCCAGAAG
TTCGTGTTAACCTGTGGAAAATCAAGTCCCTGGAAGTCTGCAG
AGACAGAGACAAGAAGAAAAGAGATAGAACGCCAGAAACTCCT
CTCTCTCCCTCCCTCTCCACCTCTCTCTTCTAAACCCCAAATT
CCTGGTCCCTTGTACCCCATTTGTGGGGATAATATGTTGTTCTT
CGCTCTCCTGTTAAAGGTGACTTGGATCCTGGCTGCAGATGGG
GGTCACCACTGGACATATGAAGGCCACACGGTCAGGACCATT
GGCCAACCTCTTATCCTGAGTGTGGAGGCGATGCCAGTCCCC
CATCAATATCCAGACAGACAGTGTGATATTTGACCCCGATCTG
CCTGCTGTACAGCCCCATGGATATGACCAGCTTGGGACTGAGC
CTTTGGATCTACACAATAATGGCCATACAGTGCAGCTTCCCT
GCCCCAACCCCTGCACCTGGGTGGACTGCCCCGAAAATACACA
...
```



Pourquoi comparer des séquences ?

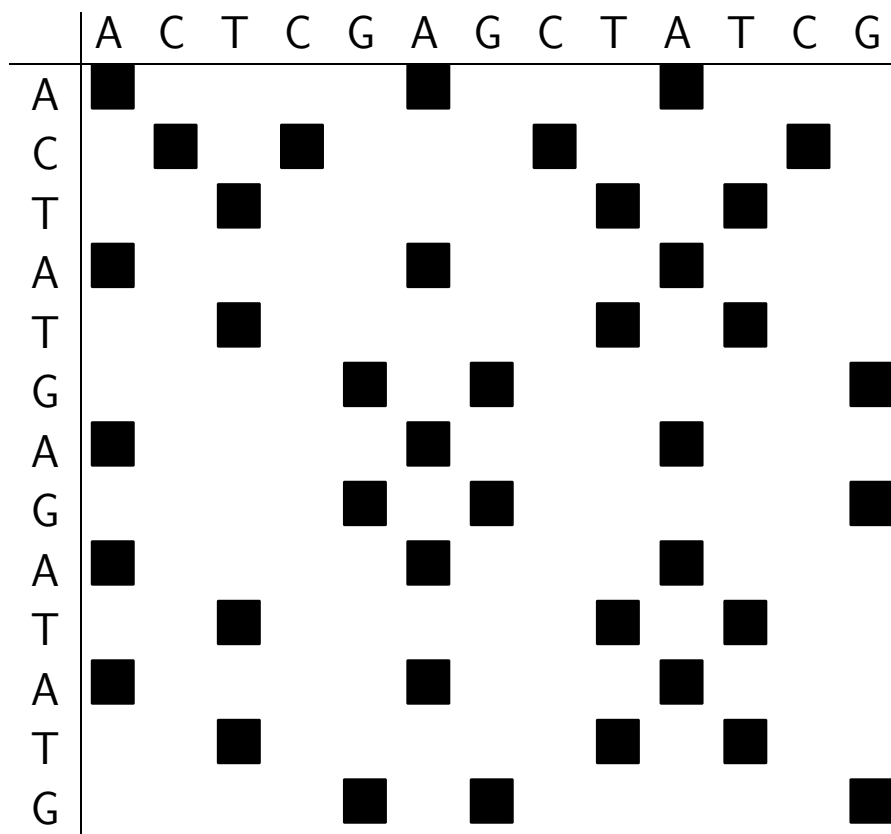
- l'évolution se fait par mutations successives
- homologie \Rightarrow similarité
- même séquence \Rightarrow même fonction ?

Le dot plot

Maizel & Lenk - 1981, Staden - 1982

Tableau de points, indexé par les deux séquences.

match (identité) → ■
mismatch → □



Les similarités apparaissent le long des segments diagonaux.

Mise en pratique

▷ Filtrage

Les éléments ne sont pas traités un par un, mais par fenêtre (de taille 25 généralement). Seules les fenêtres avec un score suffisamment élevé sont retenues.

→ Elimination des similitudes courtes, non significatives

▷ Matrice de score

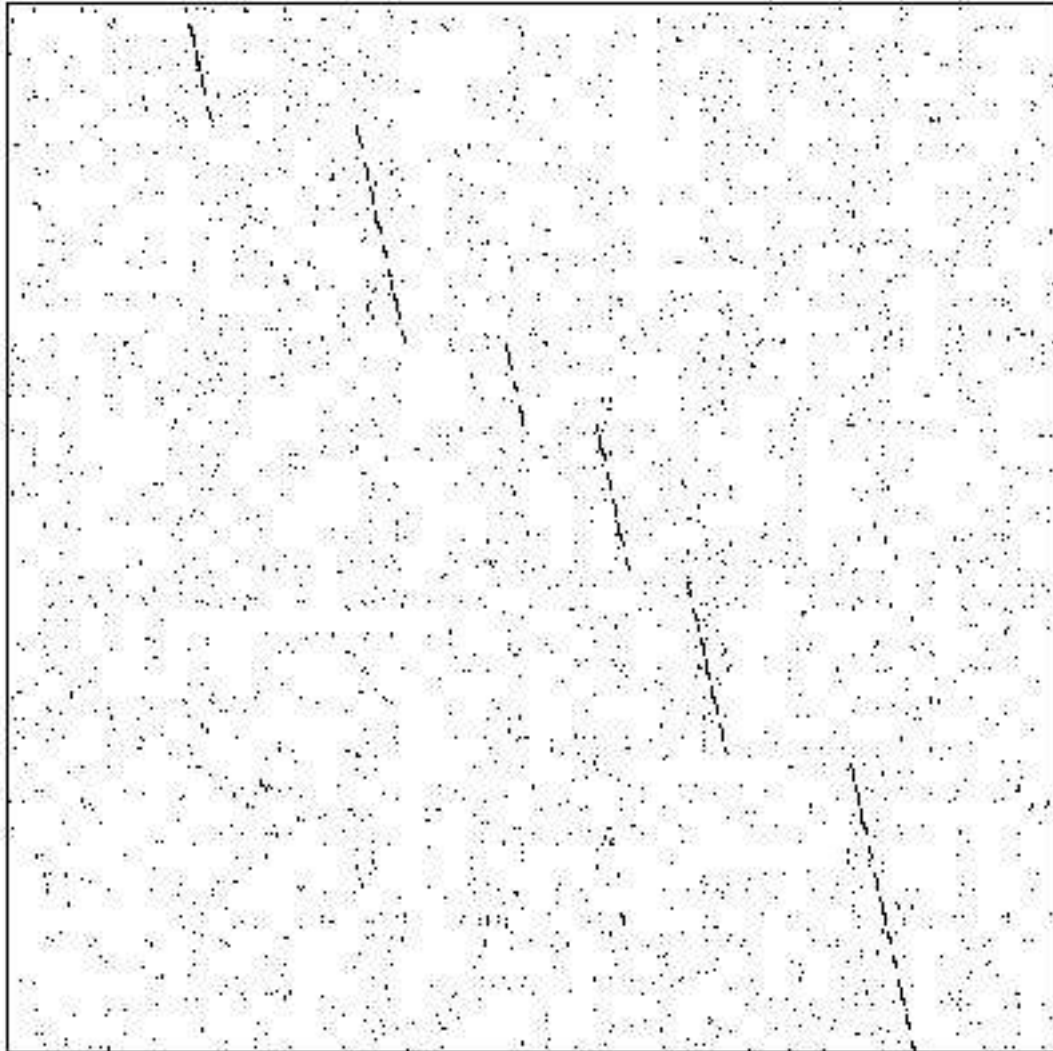
Variation d'intensité : on mesure la similarité entre deux bases ou deux acides aminés plus finement que par ■ et □.

→ Prise en compte des propriétés des acides aminés avec des matrices de pour les correspondances entre acides aminés (PAM, BLOSUM)



horizontalement : ADN codant pour la chaîne α de l'hémoglobine humaine

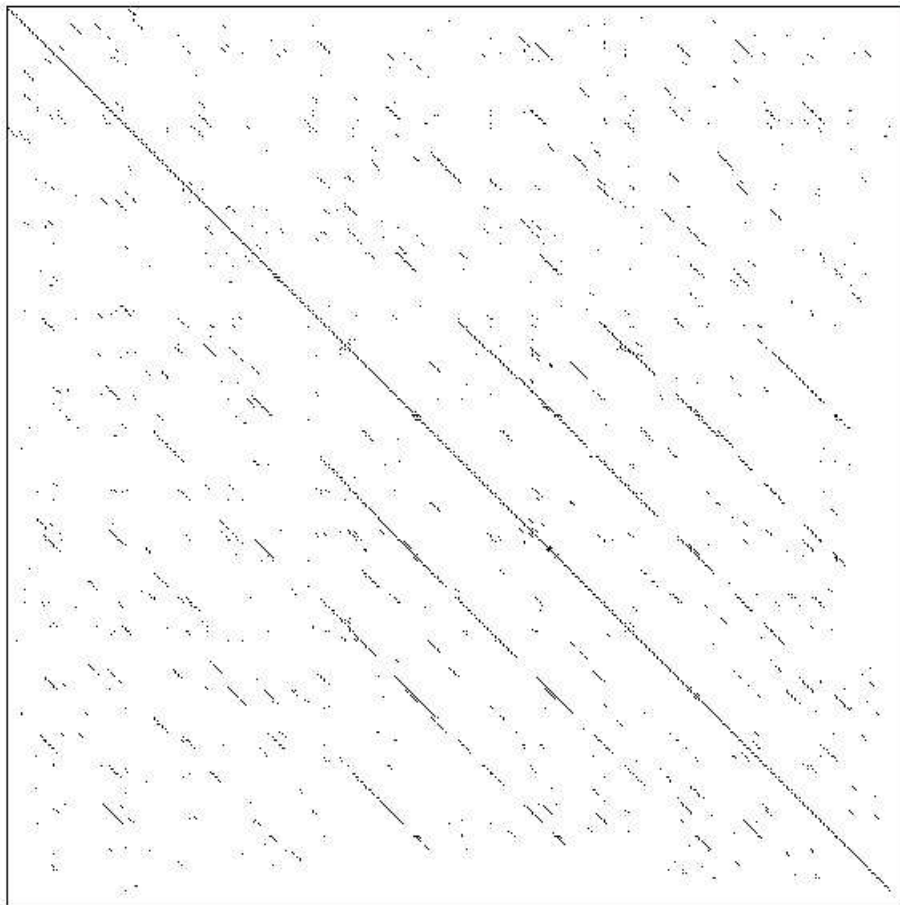
verticalement : ADN codant pour la chaîne β de l'hémoglobine humaine



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

verticalement : cDNA de ce même gène

Localisation des répétitions:
une séquence contre elle-même



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

Avantage du dot plot

- ▷ Simple
- ▷ Très informatif : on repère toutes les similarités

Mais ...

- ▷ Pas d'*historique* général sur les séquences
- ▷ Interprétation : pas de mesure objective
L'apparition des zones de similarité dépend du filtrage.
- ▷ Traitement : pas d'automatisation du résultat

Nécessité d'avoir une mesure quantitative de la similarité.

Alignement

- ▷ Mise en correspondance de deux séquences (ADN ou protéines)

```
R D I S L V - - - K N A G I
|   |   | |       | |   | |
R N I - L V S D A K N V G I
```

- ▷ 3 événements mutationnels élémentaires :

- substitution
 - insertion
 - délétion
- } *indel*

- ▷ Score

- substitution : matrice de similarité
- indel : pénalité

Le score de l'alignement est la somme des scores des événements élémentaires.

▷ 2 séquences

→ plusieurs alignements possibles

```
R D I S L V - - - K N A G I
| | | | | | | |
R N I - L V S D A K N V G I
```

```
R D I - - S L V K N A - - - G I
| | | | | | | |
R N I L V S - - - D A K N V G I
```

```
R D I - - S L V K N A G I
| | | | | | | |
R N I L V S D A K N V G I
```

▷ Bon/mauvais alignement? *Score*

Exemple inspiré de BLOSUM62

Mismatch :	Match :
DN : 1	G, N : 6
AV, LD : 0	R, K : 5
	A, I, L, S, V : 4
Indel : -5	

Alignement global

Needleman & Wunsch - 1970

Évaluation d'une ressemblance globale entre deux séquences

Données

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

Problème

Quel est l'alignement de score maximal ?

Amélioration du modèle : traitement des gaps

Gap: succession de délétions ou d'insertions

Un gap correspond à un seul événement mutationnel (insertion ou disparition d'un bloc).

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

Nouvelles pénalités :

- pénalité d'ouverture de gap (exemple : -3)
- pénalité d'extension de gap (exemple : -0.5)

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

Sensibilité aux paramètres

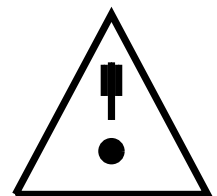
- ▷ Match 2, mismatch -1, indel -1

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

- ▷ Match 1, mismatch -1, indel -2

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

*L'alignement optimal dépend
de la matrice de similarité et des
pénalités pour les indels.*



Alignement local

Smith & Waterman -1981

Données

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

Problème

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

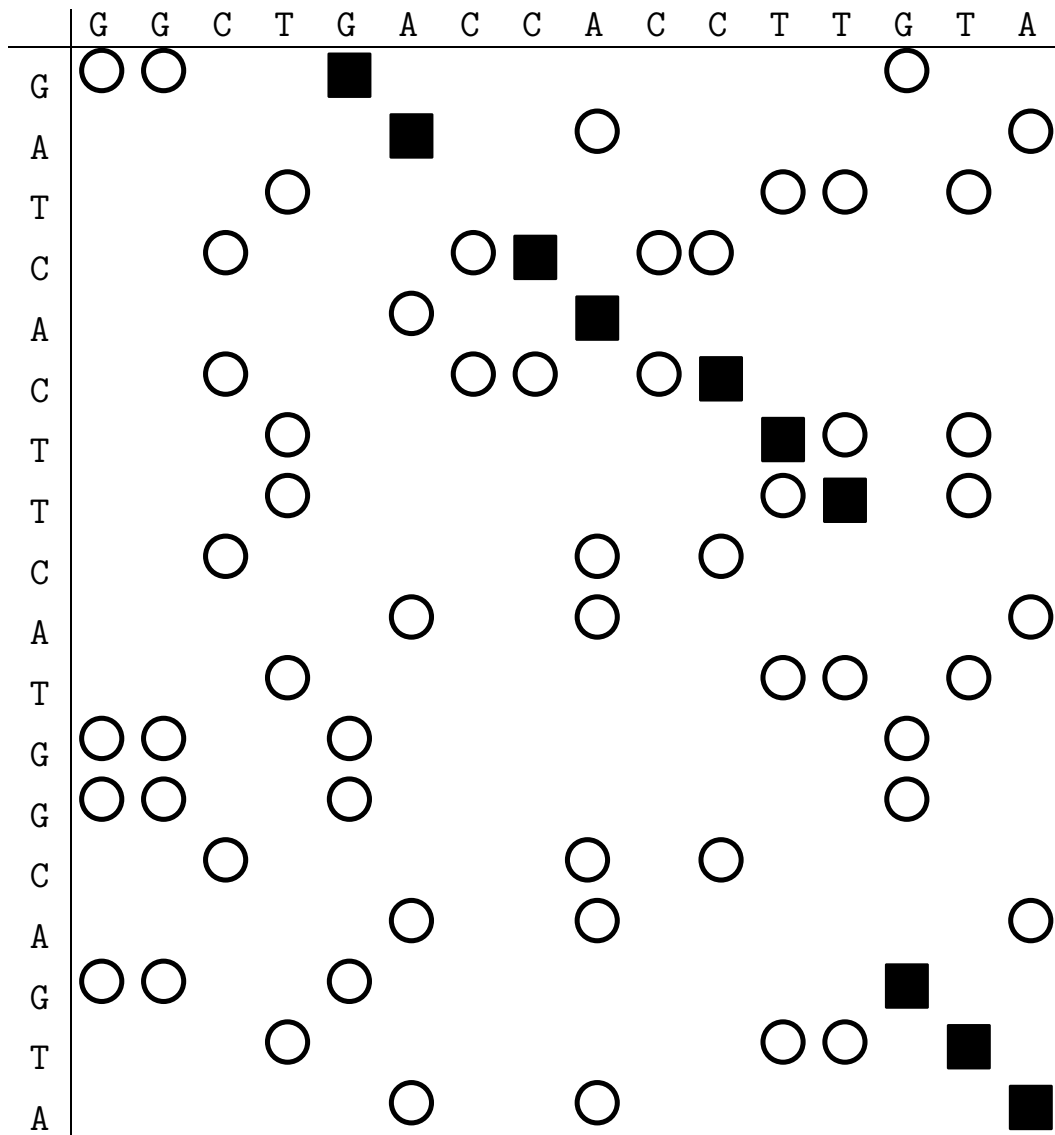
Exemple : GGCTGACCACCTTGTA et GATCACTTCCATGGCAGTA

Alignement global :

```

1  G G C T G A C C A C C _ T T G T A - - -   16
    |   |   | |   | |   |   |   |
1  G A _ T C A C T T C C A T G G C A G T A   19
  
```

Dot plot :



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

Alignement local :

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

```
14   G T A   16
     | | |
17   G T A   19
```

Comment savoir un un score est significatif ?

Approche empirique

Test de la robustesse du score

S : score de l'alignement entre U et V

1. *Génération de 200 (500, 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 ?

Exemple : Alignement local pour ACCAGTGCAGTTC 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

BLAST

Basic Local Alignment Search Tool

Altschul et al. - 1997

- ▷ Programme pour la recherche de similarités dans les bases de données
- ▷ Utilise un algorithme très rapide pour construire des alignements locaux approchés
- ▷ Séquences nucléiques et protéiques
- ▷ Connecté aux principales banques de données

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

ALIGNMENTS

>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 ggggcctgcggcgctgggaggccgcgctcgggccaagctgcacggccgg 99

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

Query 215 ggaggcgcgccaagatcacgggccgggagcgcaaggccatgagggtcct 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 ggcctgcggcgctgggaggcggctcgccaggccaagctgcactgccgggccc 104
|||||

4 ggcctgcagcgctgggaggtggcacgtcgcgccaagctgcacggccgcgccc 56

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

