

Comparaisons locales et matrices de score

Équipe Bonsai

<http://www.lifl.fr/bonsai>

année 2013



- Matrices de scores
- Recherches locales : BLAST et FastA

Exemples pour l'ADN

Identité (similarité)

	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

BLAST (similarité)

	A	C	G	T
A	1	-3	-3	-3
C	-3	1	-3	-3
G	-3	-3	1	-3
T	-3	-3	-3	1

Transition/Transversion

	A	C	G	T
A	5	-4	-3	-4
C	-4	5	-4	-3
G	-3	-4	5	-4
T	-4	-3	-4	5

Matrice BLOSUM-62 pour les protéines

```
# Matrix made by matblas from blosum62.iij
# * column uses minimum score
# BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
# Blocks Database = /data/blocks.5.0/blocks.dat
# Cluster Percentage: >= 62
# Entropy = 0.6979, Expected = -0.5209
# A R N D C Q E G H I L K M F P S T W Y V B Z X *
A 4 -1 -2 -2 0 -1 -1 0 -2 -1 -1 -1 -1 -2 -1 1 0 -3 -2 0 -2 -1 0 -4
R -1 5 0 -2 -3 1 0 -2 0 -3 -2 2 -1 -3 -2 -1 -1 -3 -2 -3 -1 0 -1 -4
N -2 0 6 1 -3 0 0 0 1 -3 -3 0 -2 -3 -2 1 0 -4 -2 -3 3 0 -1 -4
D -2 -2 1 6 -3 0 2 -1 -1 -3 -4 -1 -3 -3 -1 0 -1 -4 -3 -3 4 1 -1 -4
C 0 -3 -3 -3 9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1 1 0 0 -3 5 2 -2 0 -3 -2 1 0 -3 -1 0 -1 -2 -1 -2 0 3 -1 -4
E -1 0 0 2 -4 2 5 -2 0 -3 -3 1 -2 -3 -1 0 -1 -3 -2 -2 1 4 -1 -4
G 0 -2 0 -1 -3 -2 -2 6 -2 -4 -4 -2 -3 -3 -2 0 -2 -2 -3 -3 -1 -2 -1 -4
H -2 0 0 1 -1 -3 0 0 -2 8 -3 -3 -1 -2 -1 -2 -1 -2 -2 2 -3 0 0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 3 -4 3 2 -3 1 0 -3 -2 -1 -3 -1 3 -3 -3 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3 2 4 -2 2 0 -3 -2 -1 -2 -1 1 -4 -3 -1 -4
K -1 2 0 -1 -3 1 1 -2 -1 -3 -2 5 -1 -3 -1 0 -1 -3 -2 -2 0 1 -1 -4
M -1 -1 -2 -3 -1 0 -2 -3 -2 1 2 -1 5 0 -2 -1 -1 -1 -1 1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1 0 0 -3 0 6 -4 -2 -2 1 3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4 7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S 1 -1 1 0 0 1 0 0 0 -1 -2 -2 0 -1 -2 -1 4 1 -3 -2 -2 0 0 0 -4
T 0 -1 0 0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1 1 5 -2 -2 0 -1 -1 0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1 1 -4 -3 -2 11 2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3 2 -1 -1 -2 -1 3 -3 -2 -2 2 7 -1 -3 -2 -1 -4
V 0 -3 -3 -3 -1 -2 -2 -3 -3 3 1 -2 1 -1 -2 -2 0 -3 -1 4 -3 -2 -1 -4
B -2 -1 3 4 -3 0 1 -1 0 -3 -4 0 -3 -3 -2 0 -1 -4 -3 -3 4 1 -1 -4
Z -1 0 0 0 1 -3 3 4 -2 0 -3 -3 1 -1 -3 -1 0 -1 -3 -2 -2 1 4 -1 -4
X 0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2 0 0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 1
```

Matrice BLOSUM-62 pour les protéines

```
# Matrix made by matblas from blosum62.iiij
# * column uses minimum score
# BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
# Blocks Database = /data/blocks.5.0/blocks.dat
# Cluster Percentage: >= 62
# Entropy = 0.6979, Expected = -0.5209
# A R N D C Q E G H I L K M F P S T W Y V B Z X *
A 4 -1 -2 -2 0 -1 -1 0 -2 -1 -1 -1 -1 -2 -1 1 0 -3 -2 0 -2 -1 0 -4
R -1 5 0 -2 -3 1 0 -2 0 -3 -2 2 -1 -3 -2 -1 -1 -3 -2 -3 -1 0 -1 -4
N -2 0 6 1 -3 0 0 0 1 -3 -3 0 -2 -3 -2 1 0 -4 -2 -3 3 0 -1 -4
D -2 -2 1 6 -3 0 2 -1 -1 -3 -4 -1 -3 -3 -1 0 -1 -4 -3 -3 4 1 -1 -4
C 0 -3 -3 -3 9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1 1 0 0 -3 5 2 -2 0 -3 -2 1 0 -3 -1 0 -1 -2 -1 -2 0 3 -1 -4
E -1 0 0 2 -4 2 5 -2 0 -3 -3 1 -2 -3 -1 0 -1 -3 -2 -2 1 4 -1 -4
G 0 -2 0 -1 -3 -2 -2 6 -2 -4 -4 -2 -3 -3 -2 0 -2 -2 -3 -3 -1 -2 -1 -4
H -2 0 0 1 -1 -3 0 0 -2 8 -3 -3 -1 -2 -1 -2 -1 -2 -2 2 -3 0 0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 3 4 2 -3 1 0 -3 -2 -1 -3 -1 3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3 2 4 -2 2 0 -3 -2 -1 -2 -1 1 -4 -3 -1 -4
K -1 2 0 0 -1 -3 1 1 -2 -1 -3 -2 5 -1 -3 -1 0 -1 -3 -2 -2 0 1 -1 -4
M -1 -1 -2 -3 -1 0 -2 -3 -2 1 2 -1 5 0 -2 -1 -1 -1 -1 1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1 0 0 -3 0 6 -4 -2 -2 1 3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4 7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S 1 -1 1 0 0 1 0 0 0 -1 -2 -2 0 -1 -2 -1 4 1 -3 -2 -2 0 0 0 -4
T 0 -1 0 0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1 1 5 -2 -2 0 -1 -1 0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1 1 -4 -3 -2 11 2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3 2 -1 -1 3 -2 -1 3 -3 -2 -2 2 7 -1 -3 -2 -1 -4
V 0 -3 -3 -3 -1 -2 -2 -3 -3 3 1 -2 1 -1 -2 -2 0 -3 -1 4 -3 -2 -1 -4
B -2 -1 3 4 -3 0 1 -1 0 -3 -4 0 -3 -3 -2 0 -1 -4 -3 -3 4 1 -1 -4
Z -1 0 0 0 1 -3 3 4 -2 0 -3 -3 1 -1 -3 -1 0 -1 -3 -2 -2 1 4 -1 -4
X 0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2 0 0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 1
```

Matrice BLOSUM-62 pour les protéines

```
# Matrix made by matblas from blosum62.iij
# * column uses minimum score
# BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
# Blocks Database = /data/blocks.5.0/blocks.dat
# Cluster Percentage: >= 62
# Entropy = 0.6979, Expected = -0.5209
# A R N D C Q E G H I L K M F P S T W Y V B Z X *
A 4 -1 -2 -2 0 -1 -1 0 -2 -1 -1 -1 -1 -2 -1 1 0 -3 -2 0 -2 -1 0 -4
R -1 5 0 -2 -3 1 0 -2 0 -3 -2 2 -1 -3 -2 -1 -1 -3 -2 -3 -1 0 -1 -4
N -2 0 6 1 -3 0 0 0 1 -3 -3 0 -2 -3 -2 1 0 -4 -2 -3 3 0 -1 -4
D -2 -2 1 6 -3 0 2 -1 -1 -3 -4 -1 -3 -3 -1 0 -1 -4 -3 -3 4 1 -1 -4
C 0 -3 -3 -3 9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1 1 0 0 -3 5 2 -2 0 -3 -2 1 0 -3 -1 0 -1 -2 -1 -2 0 3 -1 -4
E -1 0 0 2 -4 2 5 -2 0 -3 -3 1 -2 -3 -1 0 -1 -3 -2 -2 1 4 -1 -4
G 0 -2 0 -1 -3 -2 -2 6 -2 -4 -4 -2 -3 -3 -2 0 -2 -2 -3 -3 -1 -2 -1 -4
H -2 0 0 1 -1 -3 0 0 -2 8 -3 -3 -1 -2 -1 -2 -1 -2 -2 2 3 0 0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -3 -4 3 4 2 -3 1 0 -3 -2 -1 -3 -1 3 -3 -3 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3 2 4 -2 2 0 -3 -2 -1 -2 -1 1 -4 -3 -1 -4
K -1 2 0 0 -1 -3 1 1 -2 -1 -3 -2 5 -1 -3 -1 0 -1 -3 -2 -2 0 1 -1 -4
M -1 -1 -2 -3 -1 0 -2 -3 -2 1 2 -1 5 0 -2 -1 -1 -1 -1 1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1 0 0 -3 0 6 -4 -2 -2 1 3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4 7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S 1 -1 1 0 0 -1 0 0 0 -1 -2 -2 0 -1 -2 -1 4 1 -3 -2 -2 0 0 0 -4
T 0 -1 0 0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1 1 5 -2 -2 0 -1 -1 0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1 1 -4 -3 -2 11 2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3 2 -1 -1 -2 -1 3 -3 -2 -2 2 7 -1 -3 -2 -1 -4
V 0 -3 -3 -3 -1 -2 -2 -3 -3 3 1 -2 1 -1 -2 -2 0 -3 -1 4 -3 -2 -1 -4
B -2 -1 3 4 -3 0 1 -1 0 -3 -4 0 -3 -3 -2 0 -1 -4 -3 -3 4 1 -1 -4
Z -1 0 0 0 1 -3 3 4 -2 0 -3 -3 1 -1 -3 -1 0 -1 -3 -2 -2 1 4 -1 -4
X 0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2 0 0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 1
```

Matrice BLOSUM-62 pour les protéines

```
# Matrix made by matblas from blosum62.iij
# * column uses minimum score
# BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
# Blocks Database = /data/blocks.5.0/blocks.dat
# Cluster Percentage: >= 62
# Entropy = 0.6979, Expected = -0.5209
# A R N D C Q E G H I L K M F P S T W Y V B Z X *
A 4 -1 -2 -2 0 -1 -1 0 -2 -1 -1 -1 -1 -2 -1 1 0 -3 -2 0 -2 -1 0 -4
R -1 5 0 -2 -3 1 0 -2 0 -3 -2 2 -1 -3 -2 -1 -1 -3 -2 -3 -1 0 -1 -4
N -2 0 6 1 -3 0 0 0 1 -3 -3 0 -2 -3 -2 1 0 -4 -2 -3 3 0 -1 -4
D -2 -2 1 6 -3 0 2 -1 -1 -3 -4 -1 -3 -3 -1 0 -1 -4 -3 -3 4 1 -1 -4
C 0 -3 -3 -3 9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1 1 0 0 -3 5 2 -2 0 -3 -2 1 0 -3 -1 0 -1 -2 -1 -2 0 3 -1 -4
E -1 0 0 2 -4 2 5 -2 0 -3 -3 1 -2 -3 -1 0 -1 -3 -2 -2 1 4 -1 -4
G 0 -2 0 -1 -3 -2 -2 6 -2 -4 -4 -2 -3 -3 -2 0 -2 -2 -3 -3 -1 -2 -1 -4
H -2 0 0 1 -1 -3 0 0 -2 8 -3 -3 -1 -2 -1 -2 -1 -2 -2 2 -3 0 0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 3 -4 2 -3 1 0 -3 -2 -1 -3 -1 3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3 2 4 -2 2 0 -3 -2 -1 -2 -1 1 -4 -3 -1 -4
K -1 2 0 0 -1 -3 1 1 -2 -1 -3 -2 5 -1 -3 -1 0 -1 -3 -2 -2 0 1 -1 -4
M -1 -1 -2 -3 -1 0 -2 -3 -2 1 2 -1 5 0 -2 -1 -1 -1 -1 1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1 0 0 -3 0 6 -4 -2 -2 1 3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4 7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S 1 -1 1 0 0 -1 0 0 0 -1 -2 -2 0 -1 -2 -1 4 1 -3 -2 -2 0 0 0 -4
T 0 -1 0 0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1 1 5 -2 -2 0 -1 -1 0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1 1 -4 -3 -2 11 2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3 2 -1 -1 -2 -1 3 -3 -2 -2 2 7 -1 -3 -2 -1 -4
V 0 -3 -3 -3 -1 -2 -2 -3 3 1 -2 1 -1 -2 -2 0 -3 -1 4 -3 -2 -1 -4
B -2 -1 3 4 -3 0 1 -1 0 -3 -4 0 -3 -3 -2 0 -1 -4 -3 -3 4 1 -1 -4
Z -1 0 0 0 1 -3 3 4 -2 0 -3 -3 1 -1 -3 -1 0 -1 -3 -2 -2 1 4 -1 -4
X 0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2 0 0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 1
```

Matrice BLOSUM-62 pour les protéines

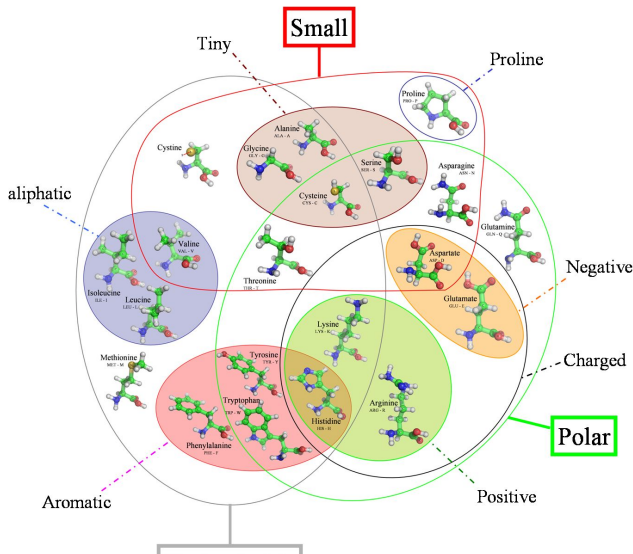
```
# Matrix made by matblas from blosum62.iij
# * column uses minimum score
# BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
# Blocks Database = /data/blocks.5.0/blocks.dat
# Cluster Percentage: >= 62
# Entropy = 0.6979, Expected = -0.5209
# A R N D C Q E G H I L K M F P S T W Y V B Z X *
A 4 -1 -2 -2 0 -1 -1 0 -2 -1 -1 -1 -1 -2 -1 1 0 -3 -2 0 -2 -1 0 -4
R -1 5 0 -2 -3 1 0 -2 0 -3 -2 2 -1 -3 -2 -1 -1 -3 -2 -3 -1 0 -1 -4
N -2 0 6 1 -3 0 0 0 1 -3 -3 0 -2 -3 -2 1 0 -4 -2 -3 3 0 -1 -4
D -2 -2 1 6 -3 0 2 -1 -1 -3 -4 -1 -3 -3 -1 0 -1 -4 -3 -3 4 1 -1 -4
C 0 -3 -3 -3 9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1 1 0 0 -3 5 2 -2 0 -3 -2 1 0 -3 -1 0 -1 -2 -1 -2 0 3 -1 -4
E -1 0 0 2 -4 2 5 -2 0 -3 -3 1 -2 -3 -1 0 -1 -3 -2 -2 1 4 -1 -4
G 0 -2 0 -1 -3 -2 -2 6 -2 -4 -4 -2 -3 -3 -2 0 -2 -2 -3 -3 -1 -2 -1 -4
H -2 0 0 1 -1 -3 0 0 -2 8 -3 -3 -1 -2 -1 -2 -1 -2 -2 2 -3 0 0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 3 -4 2 -3 1 0 -3 -2 -1 -3 -1 3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3 2 4 -2 2 0 -3 -2 -1 -2 -1 1 -4 -3 -1 -4
K -1 2 0 0 -1 -3 1 1 -2 -1 -3 -2 5 -1 -3 -1 0 -1 -3 -2 -2 0 1 -1 -4
M -1 -1 -2 -3 -1 0 -2 -3 -2 1 2 -1 5 0 -2 -1 -1 -1 -1 1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1 0 0 -3 0 6 -4 -2 -2 1 3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4 7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S 1 -1 1 0 0 1 0 0 0 -1 -2 -2 0 -1 -2 -1 4 1 -3 -2 -2 0 0 0 -4
T 0 -1 0 0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1 1 5 -2 -2 0 -1 -1 0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1 1 -4 -3 -2 11 2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3 2 -1 -1 -2 -1 3 -3 -2 -2 2 7 -1 -3 -2 -1 -4
V 0 -3 -3 -3 -1 -2 -2 -3 -3 3 1 -2 1 -1 -2 -2 0 -3 -1 4 -3 -2 -1 -4
B -2 -1 3 4 -3 0 1 -1 0 -3 -4 0 -3 -3 -2 0 -1 -4 -3 -3 4 1 -1 -4
Z -1 0 0 0 1 -3 3 4 -2 0 -3 -3 1 -1 -3 -1 0 -1 -3 -2 -2 1 4 -1 -4
X 0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2 0 0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 1
```


Matrice BLOSUM-62 pour les protéines

```
# Matrix made by matblas from blosum62.iij
# * column uses minimum score
# BLOSUM Clustered Scoring Matrix in 1/2 Bit Units
# Blocks Database = /data/blocks.5.0/blocks.dat
# Cluster Percentage: >= 62
# Entropy = 0.6979, Expected = -0.5209
# A R N D C Q E G H I L K M F P S T W Y V B Z X *
A 4 -1 -2 -2 0 -1 -1 0 -2 -1 -1 -1 -1 -2 -1 1 0 -3 -2 0 -2 -1 0 -4
R -1 5 0 -2 -3 1 0 -2 0 -3 -2 2 -1 -3 -2 -1 -1 -3 -2 -3 -1 0 -1 -4
N -2 0 6 1 -3 0 0 0 1 -3 -3 0 -2 -3 -2 1 0 -4 -2 -3 3 0 -1 -4
D -2 -2 1 6 -3 0 2 -1 -1 -3 -4 -1 -3 -3 -1 0 -1 -4 -3 -3 4 1 -1 -4
C 0 -3 -3 -3 9 -3 -4 -3 -3 -1 -1 -3 -1 -2 -3 -1 -1 -2 -2 -1 -3 -3 -2 -4
Q -1 1 0 0 -3 5 2 -2 0 -3 -2 1 0 -3 -1 0 -1 -2 -1 -2 0 3 -1 -4
E -1 0 0 2 -4 2 5 -2 0 -3 -3 1 -2 -3 -1 0 -1 -3 -2 -2 1 4 -1 -4
G 0 -2 0 -1 -3 -2 -2 6 -2 -4 -4 -2 -3 -3 -2 0 -2 -2 -3 -3 -1 -2 -1 -4
H -2 0 0 1 -1 -3 0 0 -2 8 -3 -3 -1 -2 -1 -2 -1 -2 -2 2 -3 0 0 -1 -4
I -1 -3 -3 -3 -1 -3 -3 -4 3 -4 2 -3 1 0 -3 -2 -1 -3 -1 3 -3 -3 -1 -4
L -1 -2 -3 -4 -1 -2 -3 -4 -3 2 4 -2 2 0 -3 -2 -1 -2 -1 1 -4 -3 -1 -4
K -1 2 0 -1 -3 1 1 -2 -1 -3 -2 5 -1 -3 -1 0 -1 -3 -2 -2 0 1 -1 -4
M -1 -1 -2 -3 -1 0 -2 -3 -2 1 2 -1 5 0 -2 -1 -1 -1 -1 1 -3 -1 -1 -4
F -2 -3 -3 -3 -2 -3 -3 -3 -1 0 0 -3 0 6 -4 -2 -2 1 3 -1 -3 -3 -1 -4
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4 7 -1 -1 -4 -3 -2 -2 -1 -2 -4
S 1 -1 1 0 0 -1 0 0 0 -1 -2 -2 0 -1 -2 -1 4 1 -3 -2 -2 0 0 0 -4
T 0 -1 0 0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -2 -1 1 5 -2 -2 0 -1 -1 0 -4
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1 1 -4 -3 -2 11 2 -3 -4 -3 -2 -4
Y -2 -2 -2 -3 -2 -1 -2 -3 2 -1 -1 -2 -1 3 -3 -2 -2 2 7 -1 -3 -2 -1 -4
V 0 -3 -3 -3 -1 -2 -2 -3 -3 3 1 -2 1 -1 -2 -2 0 -3 -1 4 -3 -2 -1 -4
B -2 -1 3 4 -3 0 1 -1 0 -3 -4 0 -3 -3 -2 0 -1 -4 -3 -3 4 1 -1 -4
Z -1 0 0 0 1 -3 3 4 -2 0 -3 -3 1 -1 -3 -1 0 -1 -3 -2 -2 1 4 -1 -4
X 0 -1 -1 -1 -2 -1 -1 -1 -1 -1 -1 -1 -1 -1 -2 0 0 -2 -1 -1 -1 -1 -1 -4
* -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 -4 1
```

Matrice BLOSUM-62 pour les protéines

source http://apodtele.googlepages.com/aa_venn_diagram.jpg



De l'importance des matrices de scores

- impliquées dans toutes les analyses par comparaison de séquences
- résultats **fortement** dépendants de la matrice
- représentent implicitement une théorie de l'évolution (matrices protéiques)
- la compréhension d'une matrice \Rightarrow un bon choix

Similarité vs. distance

- un élément de la matrice représente :
 - le coût du remplacement d'une base par une autre (*distance*)
 - la mesure de la *similarité* du remplacement

- un élément de la matrice représente :
 - le coût du remplacement d'une base par une autre (**distance**)
 - la mesure de la **similarité** du remplacement
- association entre
 - distance → phylogénie
 - similarité → recherche dans des bases de données

- un élément de la matrice représente :
 - le coût du remplacement d'une base par une autre (**distance**)
 - la mesure de la **similarité** du remplacement
- association entre
 - distance → phylogénie
 - similarité → recherche dans des bases de données
- même principe de recherche :
 - **maximiser un score \equiv minimiser une distance**
 - matrice de distance et de similarité peuvent être déduites l'une de l'autre

Comment obtenir une telle matrice ?

pour l'ADN
 \Rightarrow
souvent données de manière *ad hoc*

Comment obtenir une telle matrice ?

pour les protéines

Comment obtenir une telle matrice ?

pour les protéines

- matrices log odds ratio

$$S_{ij} = \log \frac{q_{ij}}{p_i p_j}$$

Comment obtenir une telle matrice ?

pour les protéines

- matrices **log odds ratio**

$$S_{ij} = \log \frac{q_{ij}}{p_i p_j}$$

- exprime le ratio entre :
 - la probabilité que deux résidus i et j soient alignés par **descendance**
 - et la probabilité que ceux-ci soient alignés par **chance**

Comment obtenir une telle matrice ?

pour les protéines

- matrices **log odds ratio**

$$S_{ij} = \log \frac{q_{ij}}{p_i p_j}$$

- exprime le ratio entre :
 - la probabilité que deux résidus i et j soient alignés par **descendance**
 - et la probabilité que ceux-ci soient alignés par **chance**
- explication :
 - q_{ij} = la fréquence que l'alignement de i et j soit observé dans des séquences homologues.
 - p_i = la fréquence d'occurrence de i
 - un score est > 0 si la proba d'un match significatif est $>$ à la proba d'un match aléatoire

Comment obtenir une telle matrice ?

pour les protéines

- matrices **log odds ratio**

$$S_{ij} = \log \frac{q_{ij}}{p_i p_j}$$

- exprime le ratio entre :
 - la probabilité que deux résidus i et j soient alignés par **descendance**
 - et la probabilité que ceux-ci soient alignés par **chance**
- explication :
 - q_{ij} = la fréquence que l'alignement de i et j soit observé dans des séquences homologues.
 - p_i = la fréquence d'occurrence de i
 - un score est > 0 si la proba d'un match significatif est $>$ à la proba d'un match aléatoire

Comment obtenir une telle matrice ?

pour les protéines

- matrices **log odds ratio**

$$S_{ij} = \log \frac{q_{ij}}{p_i p_j}$$

- exprime le ratio entre :
 - la probabilité que deux résidus i et j soient alignés par **descendance**
 - et la probabilité que ceux-ci soient alignés par **chance**
- explication :
 - q_{ij} = la fréquence que l'alignement de i et j soit observé dans des séquences homologues.
 - p_i = la fréquence d'occurrence de i
 - un score est > 0 si la proba d'un match significatif est $>$ à la proba d'un match aléatoire

⇒ matrices PAM et BLOSUM

BLOCKS SUBstitutions Matrices Henikoff & Henikoff, 1992

BLOCKS SUBstitutions Matrices

Henikoff & Henikoff, 1992

- fréquence de changements entre deux acides aminés avec conservation de structure
échantillon : BLOCKS (alignements multiples sans gaps)

BLOCKS SUBstitutions Matrices

Henikoff & Henikoff, 1992

- fréquence de changements entre deux acides aminés avec conservation de structure
échantillon : BLOCKS (alignements multiples sans gaps)
- BLOSUM- N : seuil de similarité, $N = \%$ de similarité

BLOCKS SUBstitutions Matrices Henikoff & Henikoff, 1992

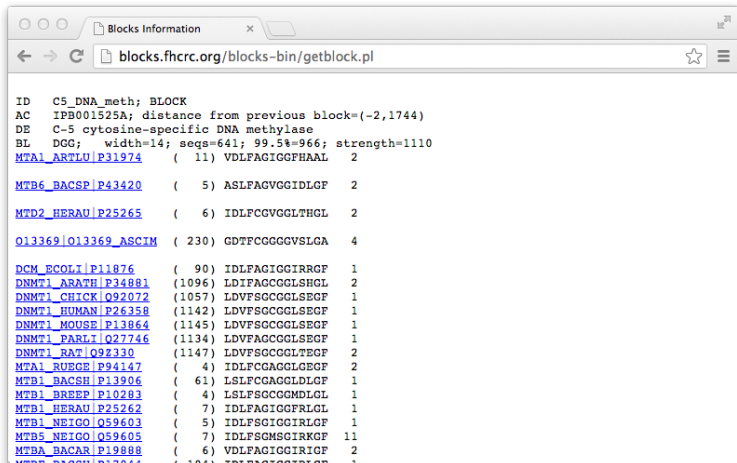
- fréquence de changements entre deux acides aminés avec conservation de structure
échantillon : BLOCKS (alignements multiples sans gaps)
- BLOSUM- N : seuil de similarité, N = % de similarité
- convient bien pour la recherche de similarités locales

BLOCKS SUBstitutions Matrices

Henikoff & Henikoff, 1992

- fréquence de changements entre deux acides aminés avec conservation de structure
échantillon : BLOCKS (alignements multiples sans gaps)
- BLOSUM- N : seuil de similarité, N = % de similarité
- convient bien pour la recherche de similarités locales
- la plus courante : BLOSUM-62 (matrice par défaut de BLAST)

Matrice BLOSUM - Matériel



```
ID    C5_DNA_meth; BLOCK
AC    IPB001525A; distance from previous block=(-2,1744)
DE    C-5 cytosine-specific DNA methylase
BL    DGG; width=14; seqs=641; 99.5%=966; strength=1110
MTA1_ARTLU|P31974 ( 11) VDLFAGIGGFHAAL 2
MTB6_BACSP|P43420 ( 5) ASLFAGVGGIDLGF 2
MTD2_HERAU|P25265 ( 6) IDLFCGVGGLTHGL 2
O13369|O13369_ASCIM ( 230) GDTFCGGGGVSLGA 4
DCM_ECOLI|P11876 ( 90) IDLFAGIGGIRRGF 1
DNMT1_ARATH|P34881 (1096) LDIFAGCGGLSHGL 2
DNMT1_CHICK|Q92072 (1057) LDVFSGCCGLSEGF 1
DNMT1_HUMAN|P26358 (1142) LDVFSGCCGLSEGF 1
DNMT1_MOUSE|P13864 (1145) LDVFSGCCGLSEGF 1
DNMT1_PARLI|Q27746 (1134) LDVFAGCGGLSEGF 1
DNMT1_RAT|Q92330 (1147) LDVFSGCCGLTEGF 2
MTA1_RUEGE|P94147 ( 4) IDLFCGAGGLGEF 2
MTB1_BACSH|P13906 ( 61) LSLFCGAGGLDLGF 1
MTB1_BREPE|P10283 ( 4) LSLFSGCCGMDLGL 1
MTB1_HERAU|P25262 ( 7) IDLFAGIGGFRLGL 1
MTB1_NEIGO|Q59603 ( 5) IDLFSGIGGIRLGF 1
MTB5_NEIGO|Q59605 ( 7) IDLFSGMSGIRKGF 11
MTBA_BACAR|P19888 ( 6) VDLFAGIGGIRIGF 2
MTB5_BACSH|P13864 (104) IDLFAGIGGIRIGF 1
```

Matrice BLOSUM - Schéma de Construction

Matrice BLOSUM - Schéma de Construction

- 1 éliminer les alignements d'identité $< n\%$ \Rightarrow **BLOSUM-n**.

Matrice BLOSUM - Schéma de Construction

- 1 éliminer les alignements d'identité $< n\% \Rightarrow$ **BLOSUM-n**.
- 2 compter le nombre f_{ij} de paires d'aa i et j .

Matrice BLOSUM - Schéma de Construction

- 1 éliminer les alignements d'identité $< n\%$ \Rightarrow **BLOSUM-n**.
- 2 compter le nombre f_{ij} de paires d'aa i et j .
- 3 calcul de la fréquence q_{ij} des paires d'aa i et j :

$$q_{ij} = \frac{f_{ij}}{\# \text{ total de paires}}$$

Matrice BLOSUM - Schéma de Construction

- 1 éliminer les alignements d'identité $< n\% \Rightarrow$ **BLOSUM-n**.
- 2 compter le nombre f_{ij} de paires d'aa i et j .
- 3 calcul de la fréquence q_{ij} des paires d'aa i et j :

$$q_{ij} = \frac{f_{ij}}{\# \text{ total de paires}}$$

- 4 calcul de la fréquence *marginale* p_i des aa i :

$$p_i = \sum_j q_{ij}$$

(plus exactement $p_i = q_{ii} + \sum_{j \neq i} q_{ij}/2$)

Matrice BLOSUM - Schéma de Construction

- 1 éliminer les alignements d'identité $< n\% \Rightarrow$ **BLOSUM-n**.
- 2 compter le nombre f_{ij} de paires d'aa i et j .
- 3 calcul de la fréquence q_{ij} des paires d'aa i et j :

$$q_{ij} = \frac{f_{ij}}{\# \text{ total de paires}}$$

- 4 calcul de la fréquence *marginale* p_i des aa i :

$$p_i = \sum_j q_{ij}$$

(plus exactement $p_i = q_{ii} + \sum_{j \neq i} q_{ij}/2$)

- 5 calcul de la matrice log odds S_{ij} :

$$S_{ij} = c \times \log \frac{q_{ij}}{p_i p_j}$$

(plus exactement $S_{i \neq j} = c \times \log \frac{q_{ij}}{2p_i p_j}$ et $S_{ii} = c \times \log \frac{q_{ii}}{p_i p_i}$)

Accepted Point Mutation / Percent Accepted Mutation
Dayhoff, 1979

Accepted Point Mutation / Percent Accepted Mutation Dayhoff, 1979

- fréquence de changements entre acides aminés
Reconstitution de l'évolution avec la construction d'arbres phylogénétiques pour
71 familles de protéines

Accepted Point Mutation / Percent Accepted Mutation Dayhoff, 1979

- fréquence de changements entre acides aminés
Reconstitution de l'évolution avec la construction d'arbres phylogénétiques pour 71 familles de protéines
- convient bien pour les séquences avec un ancêtre commun

Accepted Point Mutation / Percent Accepted Mutation Dayhoff, 1979

- fréquence de changements entre acides aminés
Reconstitution de l'évolution avec la construction d'arbres phylogénétiques pour 71 familles de protéines
- convient bien pour les séquences avec un ancêtre commun
- possibilité de choix d'une matrice en fonction de l'évolution supposée
PAM- N : N mutations acceptées par 100 acides aminés

Accepted Point Mutation / Percent Accepted Mutation Dayhoff, 1979

- fréquence de changements entre acides aminés
Reconstitution de l'évolution avec la construction d'arbres phylogénétiques pour 71 familles de protéines
- convient bien pour les séquences avec un ancêtre commun
- possibilité de choix d'une matrice en fonction de l'évolution supposée
PAM- N : N mutations acceptées par 100 acides aminés
- si la distance mutationnelle n'est pas connue, faire plusieurs essais
PAM 40, PAM 120, PAM 250, par exemple.

Matrice PAM - Construction

- 1 alignements globaux de familles de protéines (identité $> 85\%$)

Matrice PAM - Construction

- 1 alignements globaux de familles de protéines (identité $> 85\%$)
- 2 reconstruction d'une phylogénie et des ancêtres (71 familles)

Matrice PAM - Construction

- 1 alignements globaux de familles de protéines (identité $> 85\%$)
- 2 reconstruction d'une phylogénie et des ancêtres (71 familles)
- 3 compter le nombre de fois A_{ij} où un aa i est remplacé par un aa j dans toutes les comparaisons 2 à 2

Matrice PAM - Construction

- 1 alignements globaux de familles de protéines (identité $> 85\%$)
- 2 reconstruction d'une phylogénie et des ancêtres (71 familles)
- 3 compter le nombre de fois A_{ij} où un aa i est remplacé par un aa j dans toutes les comparaisons 2 à 2
- 4 estimation de la mutabilité m_j d'un aa j

- 1 alignements globaux de familles de protéines (identité > 85%)
- 2 reconstruction d'une phylogénie et des ancêtres (71 familles)
- 3 compter le nombre de fois A_{ij} où un aa i est remplacé par un aa j dans toutes les comparaisons 2 à 2
- 4 estimation de la mutabilité m_j d'un aa j
- 5 calcul de la matrice de probabilité de mutations

$$M_{ij} = \lambda \frac{m_j A_{ij}}{\sum_i A_{ij}} \text{ et } M_{jj} = 1 - \lambda m_j$$

m_j : probabilité pour j de muter

$\frac{A_{ij}}{\sum_i A_{ij}}$: probabilité conditionnelle pour j , s'il mute, de muter en i

- 1 alignements globaux de familles de protéines (identité > 85%)
- 2 reconstruction d'une phylogénie et des ancêtres (71 familles)
- 3 compter le nombre de fois A_{ij} où un aa i est remplacé par un aa j dans toutes les comparaisons 2 à 2
- 4 estimation de la mutabilité m_j d'un aa j
- 5 calcul de la matrice de probabilité de mutations

$$M_{ij} = \lambda \frac{m_j A_{ij}}{\sum_i A_{ij}} \text{ et } M_{jj} = 1 - \lambda m_j$$

m_j : probabilité pour j de muter

$\frac{A_{ij}}{\sum_i A_{ij}}$: probabilité conditionnelle pour j , s'il mute, de muter en i

- 6 calcul de la matrice log odds

$$S_{ij} = \log \frac{M_{ij}}{p_i}$$

$S_{ij} = \log \frac{p_j \times M_{ij}}{p_i p_j}$: comme p_j est la proba de j dans une séq., et M_{ij} la proba, pour un occ. j , de muter en i durant un laps de temps donné, le

Matrice PAM - Construction (1/6)

alignements de 71 groupes de l'*Atlas of protein sequence*

KAPPA

1 HUMAN EU
2 MOUSE MOPC 21
3 QAT S211
4 84 RA881T 4135
S B9 RA881T

LAMBDA

```

6 HUMAN SH
7 PIG
8 1 MOUSE MOPC 104E
9 2 MOUSE MOPC 315

```

BETA-2 MICROGLOBULIN

10 HUMAN

HEAVY CHAIN FIRST
11. GAMMA-1 FU11 GAMMA- 1 EU
12 EPSILON ND12 EPSILON ND
13 ALPHA-T 8UR

13 ALPHA-1 BOR
14 MU GAL

HEAVY CHAIN EXTRA

15 EPSILON ND

16 MU GAL
HEAVY CHAIN MIDDLE

HEAVY CHAIN M1
17 GAMMA-I EU

```

17 GAMMA=1 EO
18 EPSILON ND

```

19 ALPHA-I BUR

20 MU GAL

HEAVY CHAIN LAST

21 GAMMA-I EU
22 EFGIION ND

```

22 EPSILON ND
23 ALPHA-T BUR

```

23 ALPHA-1 BOR
24 MU GAL

24 MU GAL

[illegible]

CONSERVED

P V

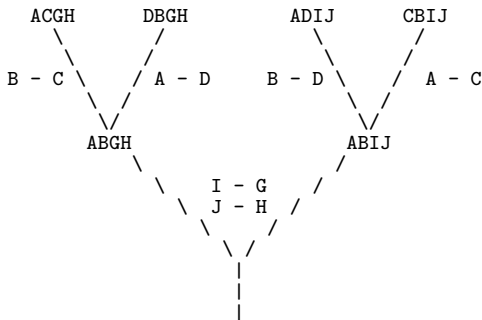
P

L. C L V G F P

V V W

Matrice PAM - Construction (2/6)

phylogenie et reconstruction des ancetres pour chacun des 71 groupes



Matrice PAM - Construction (3/6)

nombre d'accepted point mutations

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
ala																				
arg	30																			
asn	109	17																		
asp	154	0	532																	
cys	33	10	0	0																
gln	93	120	50	76	0															
glu	266	0	94	831	0	422														
gly	579	10	156	162	10	30	112													
his	21	103	226	43	10	243	23	10												
ile	66	30	36	13	17	8	35	0	3											
leu	95	17	37	0	0	75	15	17	40	253										
lys	57	477	322	85	0	147	104	60	23	43	39									
met	29	17	0	0	0	20	7	7	0	57	207	90								
phe	20	7	7	0	0	0	0	17	20	90	167	0	17							
pro	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
ser	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
thr	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
trp	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
tyr	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
val	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	

Matrice PAM - Construction (4/6)

Exemple de calcul :

Alignement	A	D	A
	A	D	B
Acides aminés	A	B	D
Chgt. observés	1	1	0
Fréquence d'occ.	3	1	2
Mutabilité relative	.33	1	0

Mutabilités relatives de tous les aa :

Ser	149	Ala	100	Gln	98
Met	122			Asp	90
Asn	111			Thr	90
Ile	110			Gap	84
Glu	102			Val	80
				Lys	57
				Pro	56
				His	50
				Gly	48
				Phe	45
				Arg	44
				Leu	38
				Tyr	34
				Cys	27
				Trp	22

Matrice PAM - Construction (5/6)

probabilités de mutation 1-PAM (une mutation pour 100 aa),

probabilités à l'échelle $\times 10000$

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
Arg R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
Asn N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
Asp D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
Cys C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Gln Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
Glu E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
Gly G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
His H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
Ile I	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
Leu L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
Lys K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
Met M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
Phe F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
Pro P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
Ser S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
Thr T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
Trp W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Tyr Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
Val V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Matrice PAM

probabilités de mutation PAM-250 = $(1 - \text{PAM})^{250}$

probabilités à l'échelle $\times 100$

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
Arg R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
Asn N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
Asp D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
Cys C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Gln Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
Glu E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
Gly G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
His H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
Ile I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
Leu L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
Lys K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
Met M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
Phe F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
Pro P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
Ser S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
Thr T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
Trp W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Tyr Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
Val V	7	4	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	72	4	17

Matrice PAM - Propriétés

- la probabilité de mutation d'un aa est fixé à 1% (1-PAM)

Matrice PAM - Propriétés

- la probabilité de mutation d'un aa est fixé à 1% (1-PAM)
- probabilités de mutation PAM-1 \Rightarrow définition d'une unité de changement évolutif

Matrice PAM - Propriétés

- la probabilité de mutation d'un aa est fixé à 1% (1-PAM)
- probabilités de mutation PAM-1 \Rightarrow définition d'une unité de changement évolutif
- applications successives \Rightarrow 2,3,4,... changements $\Rightarrow \neq$ modèles

- la probabilité de mutation d'un aa est fixé à 1% (1-PAM)
- probabilités de mutation PAM-1 \Rightarrow définition d'une unité de changement évolutif
- applications successives \Rightarrow 2,3,4,... changements $\Rightarrow \neq$ modèles
- les opérations suivantes sont équivalentes :
 - appliquer n fois les probabilités de mutation PAM-1
 - appliquer 1 fois les probabilités de mutation PAM-1 ^{n}
 - modifier les élt de PAM-1 par une constante multiplicative λ si celle ci est proche de 1.
($M_{ij} = \lambda m_j A_{ij} / \sum_i A_{ij}$ et $M_{jj} = 1 - \lambda m_j$)

- la probabilité de mutation d'un aa est fixé à 1% (1-PAM)
- probabilités de mutation PAM-1 \Rightarrow définition d'une unité de changement évolutif
- applications successives \Rightarrow 2,3,4,... changements $\Rightarrow \neq$ modèles
- les opérations suivantes sont équivalentes :
 - appliquer n fois les probabilités de mutation PAM-1
 - appliquer 1 fois les probabilités de mutation PAM-1 ^{n}
 - modifier les élt de PAM-1 par une constante multiplicative λ si celle ci est proche de 1.
($M_{ij} = \lambda m_j A_{ij} / \sum_i A_{ij}$ et $M_{jj} = 1 - \lambda m_j$)
- PAM contient de l'information sur la composition :
 - PAM-0 : $M_{ij} = 0$ et $M_{ii} = 1$
 - PAM- ∞ : approche asymptotique de la composition en aa

PAM vs. BLOSUM

BLOSUM-80
PAM-1
faible divergence

BLOSUM-62
PAM-120

BLOSUM-45
PAM-250
forte divergence

% \neq observé	dist. évolutive PAM
1	1
5	5
10	11
15	17
20	23
25	30
30	38
34	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328

- on se donne :
 - une séquence requête q
 - une banque de séquences $T = \{t_1, \dots, t_n\}$

- on se donne :
 - une séquence requête q
 - une banque de séquences $T = \{t_1, \dots, t_n\}$
- on veut :
 - trouver des alignements **significatifs** entre q et les t_i

- on se donne :
 - une séquence requête q
 - une banque de séquences $T = \{t_1, \dots, t_n\}$
- on veut :
 - trouver des alignements **significatifs** entre q et les t_i
- les algorithmes classiques ne fonctionnent pas : prennent trop de temps, il faut trouver des parades

Pearson et Lipman, 1988

Pearson et Lipman, 1988

- alignement **global** avec gaps

Pearson et Lipman, 1988

- alignement **global** avec gaps
- traite les séquences de la banque les unes après les autres

Pearson et Lipman, 1988

- alignement **global** avec gaps
- traite les séquences de la banque les unes après les autres
- fonctionnement :

Pearson et Lipman, 1988

- alignement **global** avec gaps
- traite les séquences de la banque les unes après les autres
- fonctionnement :
 - 1 trouve tous les mots identiques de longueur $\geq l$ communs à q et t_i

Pearson et Lipman, 1988

- alignement **global** avec gaps
- traite les séquences de la banque les unes après les autres
- fonctionnement :
 - 1 trouve tous les mots identiques de longueur $\geq l$ communs à q et t_i
 - 2 sélectionne ceux de score suffisamment élevé (score PAM par exemple)

Pearson et Lipman, 1988

- alignement **global** avec gaps
- traite les séquences de la banque les unes après les autres
- fonctionnement :
 - 1 trouve tous les mots identiques de longueur $\geq l$ communs à q et t_i
 - 2 sélectionne ceux de score suffisamment élevé (score PAM par exemple)
 - 3 sélectionne une diagonale d (du dotplot) contenant le maximum de mots identiques de longueur $\geq l$

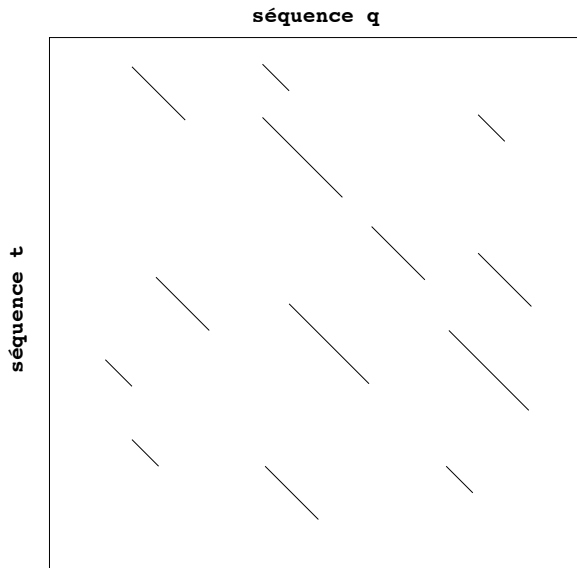
Pearson et Lipman, 1988

- alignement **global** avec gaps
- traite les séquences de la banque les unes après les autres
- fonctionnement :
 - 1 trouve tous les mots identiques de longueur $\geq l$ communs à q et t_i
 - 2 sélectionne ceux de score suffisamment élevé (score PAM par exemple)
 - 3 sélectionne une diagonale d (du dotplot) contenant le maximum de mots identiques de longueur $\geq l$
 - 4 procède à un alignement global "classique" dans une bande de largeur $2k$ autour de la diagonale d

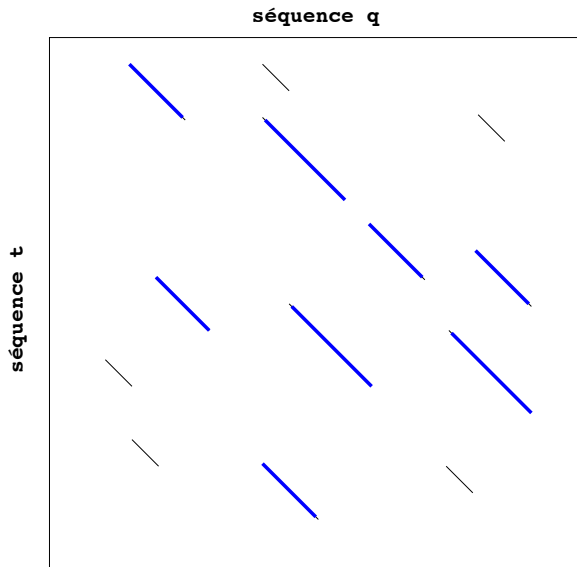
Pearson et Lipman, 1988

- alignement **global** avec gaps
- traite les séquences de la banque les unes après les autres
- fonctionnement :
 - 1 trouve tous les mots identiques de longueur $\geq l$ communs à q et t_i
 - 2 sélectionne ceux de score suffisamment élevé (score PAM par exemple)
 - 3 sélectionne une diagonale d (du dotplot) contenant le maximum de mots identiques de longueur $\geq l$
 - 4 procède à un alignement global "classique" dans une bande de largeur $2k$ autour de la diagonale d
- deux paramètres : k et l , l généralement de longueur 6 pour l'ADN et 2 pour les protéines

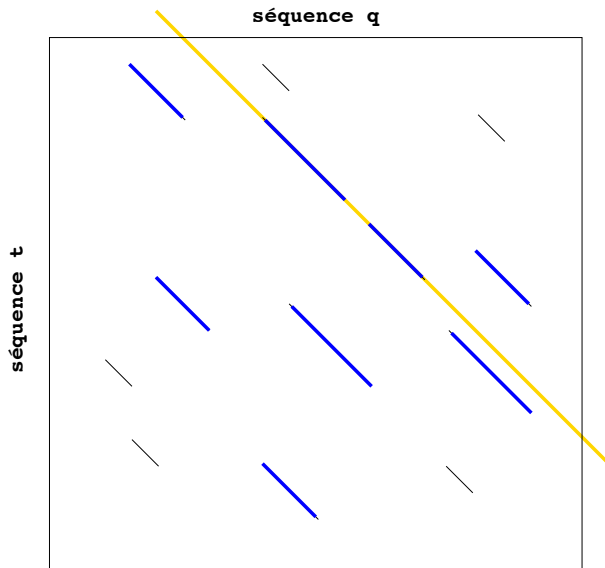
Schématiquement



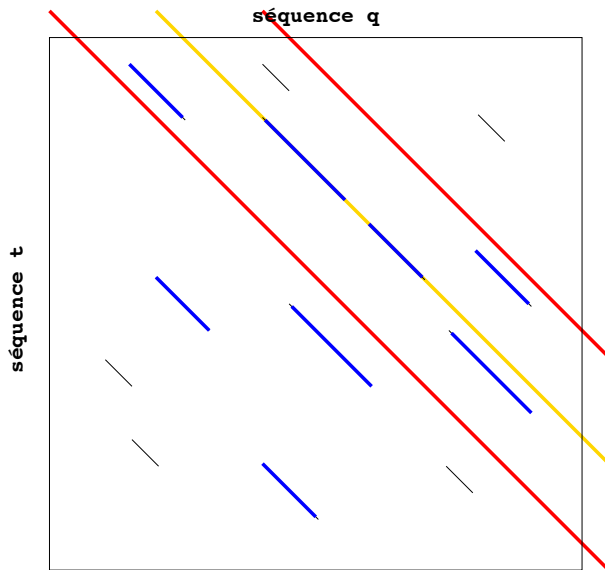
Schématiquement



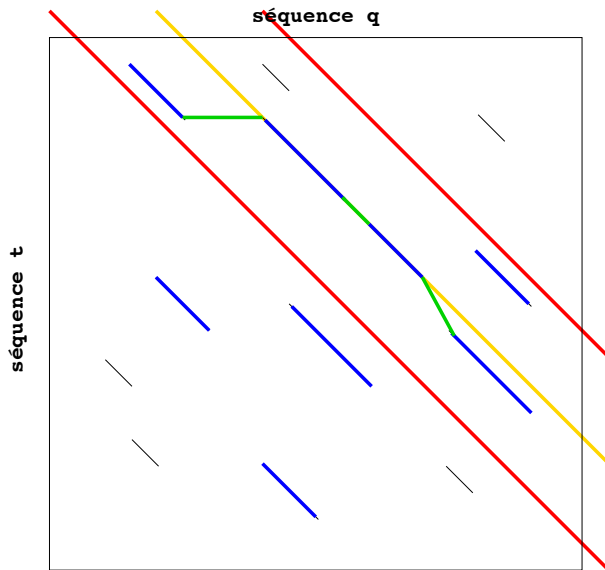
Schématiquement



Schématiquement



Schématiquement



- naît en 1990 : trouve des matchs significatifs sans gaps

- naît en 1990 : trouve des matches significatifs sans gaps
- évolution vers une version 2, avec gaps

- naît en 1990 : trouve des matches significatifs sans gaps
- évolution vers une version 2, avec gaps
 - NCBI-Blast

- naît en 1990 : trouve des matches significatifs sans gaps
- évolution vers une version 2, avec gaps
 - NCBI-Blast
 - WU-Blast : très similaire à NCBI-Blast (mix entre Blast1 et FASTA pour la dernière étape)

- naît en 1990 : trouve des matches significatifs sans gaps
- évolution vers une version 2, avec gaps
 - NCBI-Blast
 - WU-Blast : très similaire à NCBI-Blast (mix entre Blast1 et FASTA pour la dernière étape)
- évolution vers des versions avec raffinement des résultats

Blast 1

- Trouve les mots **similaires** de taille fixe entre q et t_i .
(taille par défaut : ADN \rightarrow 11 , Protéines \rightarrow 3)

- Trouve les mots **similaires** de taille fixe entre q et t_i .
(taille par défaut : ADN \rightarrow 11 , Protéines \rightarrow 3)
- Ne considère que les couples de mots ...

- Trouve les mots **similaires** de taille fixe entre q et t_i .
(taille par défaut : ADN \rightarrow 11 , Protéines \rightarrow 3)

- Ne considère que les couples de mots ...

(Protéines) **similaires** \rightarrow score des mots alignés \geq seuil T
(à l'origine $T = 13$, actuellement $T = 11$ sur BLOSUM-62)

- Trouve les mots **similaires** de taille fixe entre q et t_i .
(taille par défaut : ADN \rightarrow 11 , Protéines \rightarrow 3)

- Ne considère que les couples de mots ...

(Protéines) **similaires** \rightarrow score des mots alignés \geq seuil T

(à l'origine $T = 13$, actuellement $T = 11$ sur BLOSUM-62)

(ADN) **identiques** \rightarrow pas de seuil T donc **moins sensible**.

- Trouve les mots **similaires** de taille fixe entre q et t_i .
(taille par défaut : ADN \rightarrow 11 , Protéines \rightarrow 3)
- Ne considère que les couples de mots ...
(Protéines) **similaires** \rightarrow score des mots alignés \geq seuil T
(à l'origine $T = 13$, actuellement $T = 11$ sur BLOSUM-62)
- (ADN) **identiques** \rightarrow pas de seuil T donc **moins sensible**.
- Chaque couple de mots entre q et un t_i forme un *hit*

- Trouve les mots **similaires** de taille fixe entre q et t_i .
(taille par défaut : ADN \rightarrow 11 , Protéines \rightarrow 3)

- Ne considère que les couples de mots ...

(Protéines) **similaires** \rightarrow score des mots alignés \geq seuil T
(à l'origine $T = 13$, actuellement $T = 11$ sur BLOSUM-62)

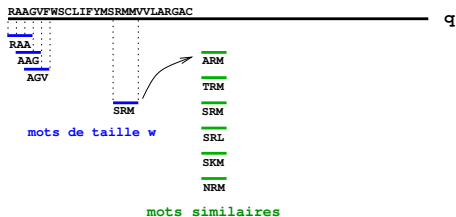
(ADN) **identiques** \rightarrow pas de seuil T donc **moins sensible**.

- Chaque couple de mots entre q et un t_i forme un *hit*
- Chaque hit est étendu à gauche et à droite : l'extension est stoppée lorsque le score du hit décroît de plus de X (X -drop)

Blast 1 - Schématiquement

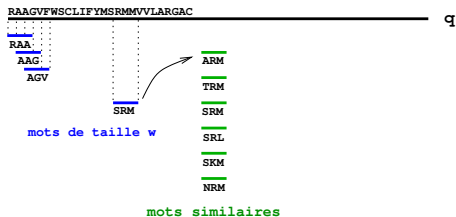
1

-

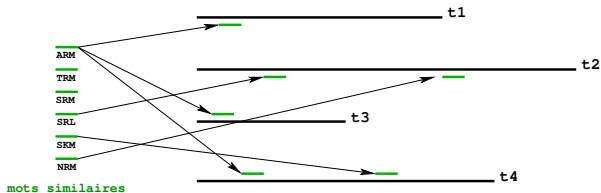


Blast 1 - Schématiquement

1 -

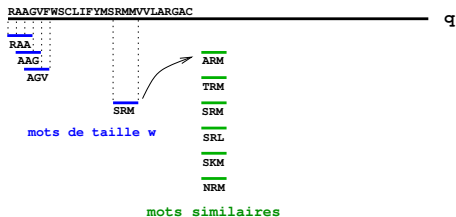


2 -

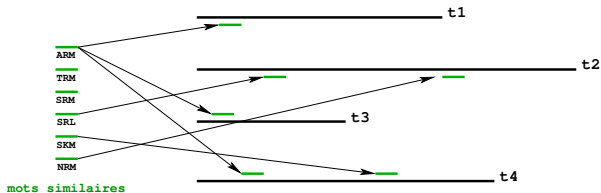


Blast 1 - Schématiquement

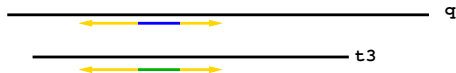
1 -



2 -



3 -



Blast 1

- un **hit** est un mot “commun” de taille fixée w (et de score supérieur à un seuil T dans le cas de BLAST-P) sur les deux séquences q et t_i

- un **hit** est un mot “commun” de taille fixée w (et de score supérieur à un seuil T dans le cas de BLAST-P) sur les deux séquences q et t_i
- chaque hit **étendu** (X-drop) forme un **LMSP** : *Locally Maximal scoring Segment Pair*

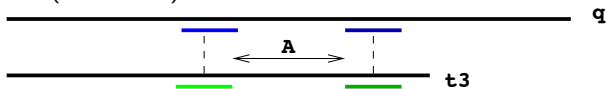
- un **hit** est un mot “commun” de taille fixée w (et de score supérieur à un seuil T dans le cas de BLAST-P) sur les deux séquences q et t_i
- chaque hit **étendu** (X-drop) forme un **LMSP** : *Locally Maximal scoring Segment Pair*
- ne conserve que les LMSP de score supérieur à un score seuil donné, les **HSP** : *High scoring Segment Pair*

- un **hit** est un mot “commun” de taille fixée w (et de score supérieur à un seuil T dans le cas de BLAST-P) sur les deux séquences q et t_i
- chaque hit **étendu** (X-drop) forme un **LMSP** : *Locally Maximal scoring Segment Pair*
- ne conserve que les LMSP de score supérieur à un score seuil donné, les **HSP** : *High scoring Segment Pair*
- significativité évaluée (pour chaque t_i) sur le meilleur HSP trouvé nommé **MSP** : *Maximun scoring Segment Pair*

NCBI - Blast 2 (Gapped-blast)

NCBI - Blast 2 (Gapped-blast)

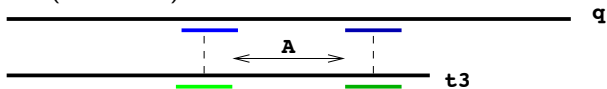
idée 1 : se baser sur 2 hits distants au maximum de A sur la même diagonale (BLASTP)



→ baisser le seuil de score de chaque hit $T = 13 \rightarrow 11$
pour conserver une bonne sensibilité

NCBI - Blast 2 (Gapped-blast)

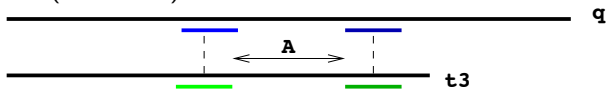
idée 1 : se baser sur 2 hits distants au maximum de A sur la même diagonale (BLASTP)



→ baisser le seuil de score de chaque hit $T = 13 \rightarrow 11$
pour conserver une bonne sensibilité

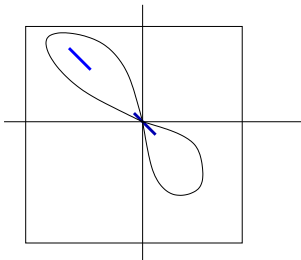
NCBI - Blast 2 (Gapped-blast)

idée 1 : se baser sur 2 hits distants au maximum de A sur la même diagonale (BLASTP)



→ baisser le seuil de score de chaque hit $T = 13 \rightarrow 11$
pour conserver une bonne sensibilité

idée 2 : étendre les hits comme dans Blast 1 (X-drop) mais en **autorisant les gaps**



MegaBLAST

pour l'ADN

- idée : un Blast plus rapide lorsqu'on recherche une grande similarité

MegaBLAST

pour l'ADN

- idée : un Blast plus rapide lorsqu'on recherche une grande similarité
- mise en œuvre : utiliser des mots de taille plus grande (28 contre 11)

MegaBLAST

pour l'ADN

- idée : un Blast plus rapide lorsqu'on recherche une grande similarité
- mise en œuvre : utiliser des mots de taille plus grande (28 contre 11)
- à réserver à des requêtes du style : trouver la séquence dans la banque

MegaBLAST

pour l'ADN

- idée : un Blast plus rapide lorsqu'on recherche une grande similarité
- mise en œuvre : utiliser des mots de taille plus grande (28 contre 11)
- à réserver à des requêtes du style : trouver la séquence dans la banque
- **évolution** : Discontiguous MegaBLAST

MegaBLAST

pour l'ADN

- idée : un Blast plus rapide lorsqu'on recherche une grande similarité
- mise en œuvre : utiliser des mots de taille plus grande (28 contre 11)
- à réserver à des requêtes du style : trouver la séquence dans la banque
- **évolution** : Discontiguous MegaBLAST
 - principe : utiliser une *graine espacée* plutôt qu'un *mot exact* (graine contiguë)

MegaBLAST

pour l'ADN

- idée : un Blast plus rapide lorsqu'on recherche une grande similarité
- mise en œuvre : utiliser des mots de taille plus grande (28 contre 11)
- à réserver à des requêtes du style : trouver la séquence dans la banque
- **évolution** : Discontiguous MegaBLAST
 - principe : utiliser une *graine espacée* plutôt qu'un *mot exact* (graine contiguë)
 - **exemple** : graine espacée 100101100101100101101 plutôt que graine contiguë 1111111111

- idée : un Blast plus rapide lorsqu'on recherche une grande similarité
- mise en œuvre : utiliser des mots de taille plus grande (28 contre 11)
- à réserver à des requêtes du style : trouver la séquence dans la banque
- **évolution** : Discontiguous MegaBLAST
 - principe : utiliser une *graine espacée* plutôt qu'un *mot exact* (graine contiguë)
 - **exemple** : graine espacée 100101100101100101101 plutôt que graine contiguë 11111111111
 - peut se révéler meilleur que BLAST (en particulier avec *graines espacées multiples*).

Définition des graines contiguës vs espacées

graine contiguë : 111111

```
ATCAGTGCAAATGCGCAAGA
|||||:|||||||.|||||
ATCAGCGCAAATGCTCAAGA
```

Définition des graines contiguës vs espacées

graine contiguë : 111111

111111

```
ATCAGTGCAAATGCGCAAGA
|||||:|||||||.|||||
ATCAGCGCAAATGCTCAAGA
```


Définition des graines contiguës vs espacées

graine contiguë : 111111

111111

```
ATCAGTGCAAATGCGCAAGA
|||||:|||||||.|||||
ATCAGCGCAAATGCTCAAGA
```

Définition des graines contiguës vs espacées

graine contiguë : 111111

111111

```
ATCAGTGCAAATGCGCAAGA
|||||:|||||||.|||||
ATCAGCGCAAATGCTCAAGA
```

Définition des graines contiguës vs espacées

graine contiguë : 111111

111111

```
ATCAGTGCAAATGCGCAAGA
|||||:|||||||.|||||
ATCAGCGCAAATGCTCAAGA
```

Définition des graines contiguës vs espacées

graine contiguë : 111111

111111

```
ATCAGTGCAAATGCGCAAGA
|||||:|||||||.|||||
ATCAGCGCAAATGCTCAAGA
```

Définition des graines contiguës vs espacées

graine contiguë : 111111

111111

```
ATCAGTGCAAATGCGCAAGA
|||||:|||||||.|||||
ATCAGCGCAAATGCTCAAGA
```

graine espacée : 11101011

```
ATCAGTGCGAATGCGCAAGA
|||||:|:|||||.|||||
ATCAGCGCAAAATGCTCAAGA
```

Définition des graines contiguës vs espacées

graine contiguë : 111111

111111

```
ATCAGTGCAAATGCGCAAGA
|||||:|||||||.|||||
ATCAGCGCAAATGCTCAAGA
```

graine espacée : 11101011

11101011

```
ATCAGTGCGAATGCGCAAGA
|||||:|:|||||.|||||
ATCAGCGCAAAATGCTCAAGA
```

Définition des graines contiguës vs espacées

graine contiguë : 111111

111111

```
ATCAGTGCAAATGCGCAAGA
|||||:|||||||.|||||
ATCAGCGCAAATGCTCAAGA
```

graine espacée : 11101011

11101011

```
ATCAGTGCGAATGCGCAAGA
|||||:|:|||||.|||||
ATCAGCGCAAAATGCTCAAGA
```

Définition des graines contiguës vs espacées

graine contiguë : 111111

111111

```
ATCAGTGCAAATGCGCAAGA
|||||:|||||||.|||||
ATCAGCGCAAATGCTCAAGA
```

graine espacée : 11101011

11101011

```
ATCAGTGCGAATGCGCAAGA
|||||:|:|||||.|||||
ATCAGCGCAAAATGCTCAAGA
```


Définition des graines contiguës vs espacées

graine contiguë : 111111

111111

```
ATCAGTGCAAATGCGCAAGA
|||||:|||||||.|||||
ATCAGCGCAAATGCTCAAGA
```

graine espacée : 11101011

11101011

```
ATCAGTGCGAATGCGCAAGA
|||||:|:|||||.|||||
ATCAGCGCAAAATGCTCAAGA
```

Définition des graines contiguës vs espacées

graine contiguë : 111111

111111

```
ATCAGTGCAAATGCGCAAGA
|||||:|||||||.|||||
ATCAGCGCAAATGCTCAAGA
```

graine espacée : 11101011

11101011

```
ATCAGTGCGAATGCGCAAGA
|||||:|:|||||.|||||
ATCAGCGCAAAATGCTCAAGA
```

Définition des graines contiguës vs espacées

graine contiguë : 111111

111111

```
ATCAGTGCAAATGCGCAAGA
|||||:|||||||.|||||
ATCAGCGCAAATGCTCAAGA
```

graine espacée : 11101011

11101011

```
ATCAGTGCGAATGCGCAAGA
|||||:|:|||||.|||||
ATCAGCGCAAAATGCTCAAGA
```

Définition des graines contiguës vs espacées

graine contiguë : 111111

111111
ATCAGTGCAAATGCGCAAGA
|||||:|||||||.|||||
ATCAGCGCAAATGCTCAAGA

graine espacée : 11101011

11101011
ATCAGTGCGAATGCGCAAGA
|||||:|:|||||.|||||
ATCAGCGCAAAATGCTCAAGA

- Les graines espacées peuvent être *bien choisies* pour mieux détecter les alignements (*Keich Li Ma Tromp DAM 2004*)

Définition des graines contiguës vs espacées

graine contiguë : 111111

111111
ATCAGTGCAAATGCGCAAGA
|||||:|||||||.|||||
ATCAGCGCAAATGCTCAAGA

graine espacée : 11101011

11101011
ATCAGTGCGAATGCGCAAGA
|||||:|:|||||.|||||
ATCAGCGCAAAATGCTCAAGA

- Les graines espacées peuvent être *bien choisies* pour mieux détecter les alignements (*Keich Li Ma Tromp DAM 2004*)
- Il est possible d'utiliser plusieurs graines espacées de formes différentes pour améliorer la sensibilité de la recherche

$$\pi_c = 111111$$

$$\pi_s = 11101011$$

$$\pi_c = 111111$$

$$\pi_s = 11101011$$

$$\alpha = \begin{array}{c} \text{ATCAGTGCAAATGCTCAAGA} \\ ||||| \\ \text{ATCAGTGCAAATGCTCAAGA} \end{array}$$

$$\pi_c = 111111$$

$$\pi_s = 11101011$$

$$\alpha = \begin{array}{cccccccccccccccc} \text{ATCAGT} & \text{GCAAATGCTCAAGA} \\ | & | & | & | & | & | & | & | & | & | & | & | & | & | & | & | \\ \text{ATCAGT} & \text{GCAAATGCTCAAGA} \end{array}$$

$$\pi_c = 111111$$

$$\pi_s = 11101011$$

$$\alpha = \begin{array}{cccccccccccccccc} \text{ATCAGT} & \text{GCAAATGCTCAAGA} \\ ||||| & . ||||| \\ \text{ATCAGC} & \text{GCAAATGCTCAAGA} \end{array}$$

$$\pi_c = 111111$$

$$\pi_s = 11101011$$

$$\alpha = \begin{array}{cccccccccccc} \text{ATCAGTGC} & \text{AAATGC} & \text{TCAAGA} \\ |||||. & ||||| & ||||| \\ \text{ATCAGCGC} & \text{AAATGC} & \text{TCAAGA} \end{array}$$

$$\pi_c = 111111$$

$$\pi_s = 11101011$$

$$\alpha = \begin{array}{cccccccccccc} \text{ATCAGTGCAAATGCGCAAGA} \\ |||||.|||||||.||||| \\ \text{ATCAGCGCAAATGCTCAAGA} \end{array}$$

$$\pi_c = 111111$$

$$\pi_s = 11101011$$

$$\alpha = \begin{array}{cccccccccccccccc} \text{ATCAGTGC} & \text{AAATGC} & \text{CAAGA} \\ ||||| & .||| & ||||| & .||||| \\ \text{ATCAGCGC} & \text{AAATGCT} & \text{CAAGA} \end{array}$$

$$\pi_c = 111111$$

$$\pi_s = 11101011$$

$$\alpha = \begin{array}{cccccccccccccccc} \text{ATCAGTGC} & \text{GAATGC} & \text{CAAGA} \\ ||||| & .|| & .||| & ||| & .||| & ||| \\ \text{ATCAGCGC} & \text{AAATGC} & \text{CAAGA} \end{array}$$

ATCAGTGCAAATGCTCAAGA
| | | | | | | | | | | | | | | | | |
ATCAGTGCAAATGCTCAAGA

ATCAGTGCAAATGCTCAAGA
||||||||||||||||
ATCAGTGCAAATGCTCAAGA

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

ATCAGTGCAAATGCTCAAGA
||||||||||||||||
ATCAGTGCAAATGCTCAAGA

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

ATCAGTGCAAATGCTCAAGA
| | | | | | | | | | | | | | | |
ATCAGTGCAAATGCTCAAGA

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

ATCAGTGCAAATGCTCAAGA
| | | | | | | | | | | | | | | |
ATCAGTGCAAATGCTCAAGA

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

ATCAGTGCAAATGCTCAAGA
|||||.|||||||
ATCAGCGCAAATGCTCAAGA

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

ATCAGTGCAAATGCTCAAGA
|||||.|||||||
ATCAGCGCAAATGCTCAAGA

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

ATCAGTGCAAATGCTCAAGA
|||||.|||||||
ATCAGCGCAAATGCTCAAGA

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

ATCAGTGCAAATGCTCAAGA
|||||.|||||||
ATCAGCGCAAATGCTCAAGA

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

ATCAGTGCAAATGCTCAAGA
|||||.|||||||
ATCAGCGCAAATGCTCAAGA

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

ATCAGTGCAAATGCTCAAGA
|||||.|||||||
ATCAGCGCAAATGCTCAAGA

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

ATCAGTGCAAATGCTCAAGA
|||||.|||||||
ATCAGCGCAAATGCTCAAGA

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

ATCAGTGCAAATGCTCAAGA
|||||.|||||||
ATCAGCGCAAATGCTCAAGA

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

ATCAGTGCAAATGCGCAAGA
| | | | | . | | | | | | | . | | | | |
ATCAGCGCAAATGCTCAAGA

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

ATCAGTGCAAATGCGCAAGA
| | | | | . | | | | | | | . | | | | |
ATCAGCGCAAATGCTCAAGA

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

ATCAGTGCAAATGCGCAAGA
|||||.|||||||.|||||
ATCAGCGCAAATGCTCAAGA

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

ATCAGTGCAAATGCGCAAGA
|||||.|||||||.|||||
ATCAGCGCAAATGCTCAAGA

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

ATCAGTGCAAATGC^GCAAGA
| | | | | . | | | | | | | . | | | | |
ATCAG^CGCAAATGCT^TCAAGA

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

ATCAGTGCAAATGC^GCAAGA
| | | | | . | | | | | | | . | | | | |
ATCAG^CGCAAATGCT^TCAAGA

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

ATCAGTGC AAATGC GCAAGA
| | | | . | | | | | . | | | |
ATCAGCGC AAATGCT CAAGA

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

ATCAGTGC AAATGC GCAAGA
| | | | . | | | | | . | | | |
ATCAGCGC AAATGCT CAAGA

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

ATCAGTGCGAATGCGCAAGA
| | | | | . | | . | | | | | . | | | | |
ATCAGCGCAAAATGCTCAAGA

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

ATCAGTGCGAATGCGCAAGA
| | | | | . | | . | | | | | . | | | | |
ATCAGCGCAAAATGCTCAAGA

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

ATCAGTGCGAATGCGCAAGA
| | | | | . | | . | | | | | . | | | | |
ATCAGCGCAAAATGCTCAAGA

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

ATCAGTGCGAATGCGCAAGA
| | | | | . | | . | | | | | . | | | | |
ATCAGCGCAAAATGCTCAAGA

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

ATCAGTGCGAATGCGCAAGA
|||||.|||.|||||.|||||
ATCAGCGCAAAATGCTCAAGA

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

ATCAGTGCGAATGCGCAAGA
|||||.|||.|||||.|||||
ATCAGCGCAAAATGCTCAAGA

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

ATCAGTGCGAATGCGCAAGA
|||||.|||.|||||.|||||
ATCAGCGCAAAATGCTCAAGA

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

111111

ATCAGTGCGAATGCGCAAGA
|||||.|||.|||||.|||||
ATCAGCGCAAAATGCTCAAGA

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011

11101011


11101011

11101011

11101011

11101011

BLAST : Page d'accueil

**BLAST®**
Basic Local Alignment Search Tool

HomeRecent ResultsSaved StrategiesHelp

• [NCBI/BLAST Home](#)

BLAST finds regions of similarity between biological sequences. [more...](#)

New DELTA-BLAST, a more sensitive protein-protein search Go

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or [list all genomic BLAST databases](#).

Human	Oryza sativa	Gallus gallus
Mouse	Bos taurus	Pan troglodytes
Rat	Danio rerio	Microbes
Arabidopsis thaliana	Drosophila melanogaster	Apis mellifera

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a nucleotide database using a nucleotide query <i>Algorithms: blastn, megablast, discontiguous megablast</i>
protein blast	Search protein database using a protein query <i>Algorithms: blastp, psi-blast, phi-blast, delta-blast</i>
blastx	Search protein database using a translated nucleotide query
tblastn	Search translated nucleotide database using a protein query
tblastx	Search translated nucleotide database using a translated nucleotide query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with [Primer-BLAST](#)
- Search [trace archives](#)
- Find [conserved domains](#) in your sequence (cds)
- Find sequences with similar [conserved domain architecture](#) (cdart)
- Search sequences that have [gene expression profiles](#) (GEO)
- Search [immunoglobulins](#) (IgBLAST)
- Search using [SNP flanks](#)

Équipe Bonsai — Comparaisons locales et matrices de score

36/57

Exemple : *BLAST Assembled RefSeq Genomes* → *Human*

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastn suite **Homo sapiens (human) Nucleotide BLAST**

blastn blastp blastx tblastn tblastx

Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

Query subrange [From](#) [To](#)

Or, upload file [Choisissez un fichier](#) [Aucun fichier choisi](#)

Job Title [Enter a descriptive title for your BLAST search](#)

Choose Search Set

Database ☒ Genome (all assemblies) 4900 sequences [Optional](#)

Exclude ☐ Genome (reference only)

Optional ☐ HTCS

Optional ☐ RefSeq Genomic

Optional ☐ RefSeq RNA

Optional ☐ Non-RefSeq RNA

Optional ☐ Build RNA

Optional ☐ Ab initio RNA

Optional ☐ ESTs

Optional ☐ Clone end sequences

Optional ☐ SNPs

Optional ☐ CRC latest patch release

Optional ☐ Genome (reference only, previous build)

Optional ☐ Genome (all assemblies, previous build)

Optional ☐ Build RNA (previous build)

Optional ☐ Ab initio RNA (previous build)

Optional ☐ CH17 BAC Ends

Optional ☐ Genome (reference assembly scaffolds)

Optional ☐ Genome (reference assembly top-level)

BLAST

[Algorithm parameters](#)

[Homo sapiens using Megablast \(Optimize for highly similar sequences\)](#)

Exemple : Specialized BLAST → Search Trace Archive

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

NCBI/BLAST/blastn suite **Trace Archive Nucleotide BLAST**

blastn

BLASTn programs search Trace Archive databases using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

Query subrange [From](#) [To](#)

Or, upload file [Choisissez un fichier](#) [Aucun fichier choisi](#)

Job Title [Enter a descriptive title for your BLAST search](#)

Choose Search Set

Database

- ✓ 10632 methanococcus maripaludis - other
- 10632 methanococcus maripaludis - WGS
- 10752 listeria monocytogenes str 4b h7858 - WGS
- 10753 listeria monocytogenes str 1 2a f6854 - WGS
- 10774 burkholderia thailandensis e264 - WGS
- 10784 bacillus anthracis str ames 0581 - other
- 10784 bacillus anthracis str ames 0581 - WGS
- 10788 bacillus cereus g9241 - other
- 10788 bacillus cereus g9241 - WGS
- 10797 bacillus anthracis str vollum - other
- 10797 bacillus anthracis str vollum - WGS
- 10799 bacillus anthracis str a0039 - other
- 10799 bacillus anthracis str a0039 - WGS
- 11785 mus musculus - WGS
- 118 medicago truncatula - other
- 12516 campylobacter coli rm2228 - WGS
- 12517 campylobacter lari rm2100 - WGS
- 12518 campylobacter upsaliensis rm3195 - WGS
- 12521 clostridium perfringens sm101 - other
- 12521 clostridium perfringens sm101 - WGS
- 12554 borrelia garinii pbi - WGS
- 12926 entamoeba invadens ip1 - WGS
- 13044 ostreococcus lucimarinus CCE9901 - WGS
- 13392 shewanella putrefaciens 200 - WGS
- 13623 bacillus weihenstephanensis kba4 - other
- 13694 environmental sequence - WGS
- 13696 environmental sequence - WGS

Program Select [Optimize for](#)

BLAST

[Algorithm parameters](#)

[\(Optimize for highly similar sequences\)](#)

[are highlighted in yellow and marked with ♣ sign](#)

Library of Medicine.

NCBI | NLM | NIH | DHHS

Exemple : Specialized BLAST → Search SRA Transcript&Genomic

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

NCBI/BLAST/ blastn suite **Sequence Read Archive Nucleotide BLAST** [Status of the NCBI Sequence Read Archive](#)

blastn

BLASTN programs search SRA databases using a nucleotide query.

[Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

Query subrange [From](#) [To](#)

Or, upload file [Choisissez un fichier](#) [Aucun fichier choisi](#)

Job Title [Enter a descriptive title for your BLAST search](#)

Choose Search Set

Database ☒ Transcript ☐ WGS

- ✓ SRA Abies balsamea
- SRA Acanthamoeba polyphaga mimivirus
- SRA Acremonium alcalophilum JCM 7366
- SRA Acropora (2 entries)
- SRA Actinopus
- SRA Aedes (2 entries)
- SRA Aegilops (2 entries)
- SRA Agaricus bisporus U1
- SRA Agkistrodon contortrix
- SRA Agrius planipennis
- SRA Alasmidonta varicosa
- SRA Alexandrium (3 entries)
- SRA Allium cepa
- SRA Amanita thiersii Skay4041
- SRA Amaranthus (2 entries)
- SRA Amblyomma maculatum
- SRA Amborella trichopoda
- SRA Ambrosia (2 entries)
- SRA Ambystoma mexicanum
- SRA Amerana
- SRA Amorphophallus konjac
- SRA Ampelomyces quisqualis
- SRA Amsonia hubrichtii
- SRA Ancistrus
- SRA Ancylostoma caninum
- SRA Anguilla (2 entries)

Program Select

Optimize for [Optimize for highly similar sequences](#)

BLAST


[Algorithm parameters](#)

Copyright | Disclaimer | Privacy

trademark of the National Library of Medicine.

NCBI | NLM | NIH | DHHS

Exemple : Basic BLAST → nucleotide blast → nr

 **BLAST®**
Basic Local Alignment Search Tool

HomeRecent ResultsSaved StrategiesHelp


•NCBI/ BLAST/ blastn suite



Standard Nucleotide BLAST

blastnblastblastblastnblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter Query Sequence



Enter accession number(s), gi(s), or FASTA sequence(s) 

 Query subrange 


From


To

>Felis catus DRD4 gene for dopamine receptor D4, partial cds.
TTCTTCCTACCTGCCCCCTCAAGCTGCTGCTACTGGGCCACGTTCCGGGGCCCTCGGGCGC
TGGGAGGGGCTCGCCAGGCCAAGCTGCACGTGCGGGGCGCTCGTCTGGCCACAGCGGCCCGCCG
CCACCGCCCCCGAGGTCGGGAGCCCCCGACGCCGTCGCGCCCCCGACGCCGTCCTCAGCC
GAGCCCGCCGGCAGGCACCCAGGAGGGGCGGCCCAAGATCACCGGCCGGGAGCGCAAGGCC
ATGAGGGTCTCTGCCGGTGGTGGTC

Or, upload file  



Job Title


Enter a descriptive title for your BLAST search 


☐ Align two or more sequences 


Choose Search Set


Database ☐ Human genomic + transcript ☐ Mouse genomic + transcript ☒ Others (nr etc.):

 Nucleotide collection (nr/nt) 


Organism 

Optional 


Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. 

Exclude 

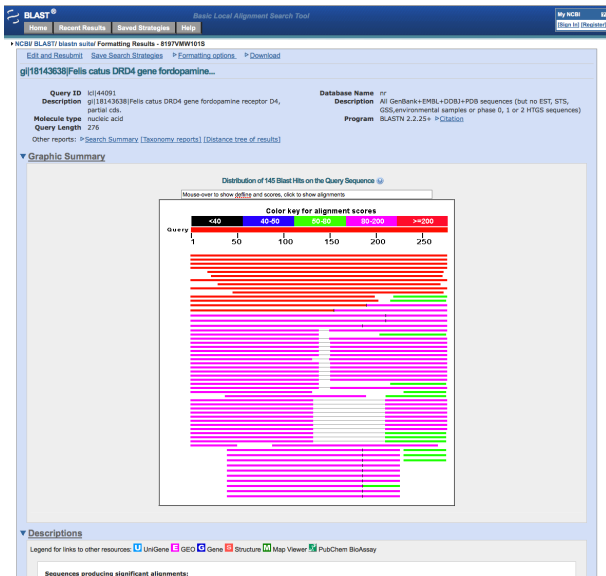
Optional ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query 

Optional

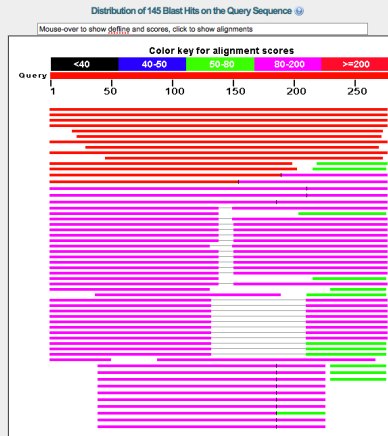
Enter an Entrez query to limit search 

Exemple : Résultats



Exemple : Résultats

▼ Graphic Summary



Exemple : Résultats

▼ Descriptions

Legend for links to other resources: [U](#) UniGene [E](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer [A](#) PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
AB069665.1	Felis catus DRD4 gene fordopamine receptor D4, partial cds	499	499	100%	5e-138	100%	
AB069664.1	Ursus thibetanus DRD4 gene fordopamine receptor D4, partial cds	315	315	100%	1e-82	81%	
AB069663.1	Procyon lotor DRD4 gene fordopamine receptor D4, partial cds	302	302	100%	8e-79	79%	
AY394848.1	Mustela putorius furo dopamine receptor D4 (DRD4) mRNA, complete	291	291	100%	1e-75	82%	
DQ029098.1	Lutra lutra dopamine D4 receptor (DRD4) gene, exon 3 and partial cds	291	291	92%	1e-75	85%	
DQ029100.1	Phoca groenlandica dopamine D4 receptor (DRD4) gene, exon 3 and p	262	262	90%	7e-67	78%	
XM_003423233.1	PREDICTED: Loxodonta africana D(4) dopamine receptor-like (LOC100	250	250	100%	4e-63	80%	G
DQ071548.1	Halichoerus grypus dopamine D4 receptor (DRD4) gene, exon 3 and p	237	237	86%	3e-59	77%	
AB069666.1	Bos taurus DRD4 gene fordopamine receptor D4, partial cds	233	233	100%	3e-58	73%	
AY611807.1	Ursus maritimus dopamine receptor D4 gene, exon 3 and partial cds	226	226	82%	5e-56	78%	
XM_003122390.1	PREDICTED: Sus scrofa D(4) dopamine receptor-like (LOC100512329)	215	215	71%	9e-53	84%	G M
AY611808.1	Mustela lutreola dopamine receptor D4 gene, exon 3 and partial cds	214	214	73%	3e-52	81%	
DQ132799.1	Lagenorhynchus albobrostris dopamine D4 receptor (DRD4) gene, part	210	336	100%	4e-51	84%	
AB069661.1	Canis lupus DRD4 gene fordopamine receptor D4, partial cds	201	365	100%	2e-48	89%	
AB069662.1	Nyctereutes procyonoides DRD4 gene fordopamine receptor D4, part	199	368	100%	7e-48	82%	
AB167773.1	Canis lupus familiaris DRD4 gene for dopamine receptor D4, partial cd	197	352	100%	2e-47	87%	G
EF561289.1	Equus caballus dopamine receptor D4 (DRD4) gene, exons 1 through	192	363	100%	1e-45	85%	
XM_003281326.1	PREDICTED: Nomascus leucogenys dopamine receptor D4 (DRD4), pa	181	319	96%	2e-42	89%	G M
XM_001149749.2	PREDICTED: Pan troglodytes dopamine receptor D4 (DRD4), partial m	178	178	50%	2e-41	88%	G M
XM_002821327.1	PREDICTED: Pongo abelii D(4) dopamine receptor-like (LOC10046226	178	310	96%	2e-41	88%	G M
NG_021241.1	Homo sapiens dopamine receptor D4 (DRD4), RefSeqGene on chromo	178	314	95%	2e-41	88%	G
NM_000797.3	Homo sapiens dopamine receptor D4 (DRD4), mRNA	178	314	95%	2e-41	88%	U E G M
BC172267.1	Synthetic construct Homo sapiens clone IMAGE:100068961, MGC:198	178	314	95%	2e-41	88%	G M
EU432112.1	Homo sapiens dopamine receptor D4 (DRD4) mRNA, complete cds	178	314	95%	2e-41	88%	U G M
XM_001087197.1	PREDICTED: Macaca mulatta dopamine receptor D4 (DRD4), mRNA	178	301	93%	2e-41	90%	G M
L12392.1	Homo sapiens Dopamine D4 receptor (DRD4) gene, partial cds	178	314	95%	2e-41	88%	U E G M
L12398.1	Homo sapiens dopamine receptor D4 (DRD4) mRNA, complete cds	178	314	95%	2e-41	88%	U E G M
AC138374.2	Homo sapiens chromosome 11, clone CTD-2647G13, complete sequen	178	314	95%	2e-41	88%	
AC131934.13	Homo sapiens chromosome 11, clone RP11-754B17, complete sequen	178	314	95%	2e-41	88%	
AP006284.2	Homo sapiens genomic DNA, chromosome 11 clone:RP11-49619, com	178	314	95%	2e-41	88%	
AJ336085.1	Homo sapiens genomic sequence surrounding NotI site, clone NR1-1P1	178	178	50%	2e-41	88%	
AB065765.1	Homo sapiens gene for seven transmembrane helix receptor, complete	172	309	95%	1e-39	87%	E G
AJ335893.1	Homo sapiens genomic sequence surrounding NotI site, clone NL3-CO	168	168	47%	1e-38	88%	
AY615863.1	Physeter catodon dopamine D4 (DRD4) gene, DRD4-3R allele, exon 3	158	158	55%	2e-35	82%	
NM_007678.2	Mus musculus dopamine receptor D4 (Drd4), mRNA >gb BC016086.1	152	233	72%	9e-34	86%	U E G M
BC051421.1	Mus musculus dopamine receptor 4, mRNA (cDNA clone MGC:58931.1	152	233	72%	9e-34	86%	U E G M
AC102547.8	Mus musculus chromosome 7, clone RP23-134L4, complete sequence	152	233	72%	9e-34	86%	
AC163434.5	Mus musculus chromosome 7, clone RP23-134L4, complete sequence	152	233	72%	9e-34	86%	

Exemple : Résultats

▼ Alignments

☐ Select All [Get selected sequences](#) [Distance tree of results](#)

>[db1|AB069665.1](#) Felis catus DRD4 gene fordopamine receptor D4, partial cds
Length=276

Score = 499 bits (552), Expect = 5e-138
Identities = 276/276 (100%), Gaps = 0/276 (0%)
Strand=Plus/Plus

```
Query 1      TTTCTTCTACCTGCCCGCTCATGCTGCTGCTTACTGGGCCACGTTCCGGGGCCTGCGG 60
Sbjct 1      TTTCTTCTACCTGCCCGCTCATGCTGCTGCTTACTGGGCCACGTTCCGGGGCCTGCGG 60

Query 61     CGCTGGGAGCGCGCTCGCCAGGCCAAGCTGCAC TGCCGGGCGCCTCGTcgagccagcggc 120
Sbjct 61     CGCTGGGAGCGCGCTCGCCAGGCCAAGCTGCAC TGCCGGGCGCCTCGTGCGCCACGCGC 120

Query 121    cccggcccaaccgccccccgaggtcgggcgagccccccgagcgccgtcgccgcccccgagcgc 180
Sbjct 121    CCCGGCCCAACGCCCCCGGAGGTGCGCGAGCCCCCGAGCGCCGTGCGCCGCCCGCGACGCC 180

Query 181    gtccagcggcgagcgccggcggcAGCAGCCACCCAGGAGGAGCGCGCCAAGATCACCGGCCGG 240
Sbjct 181    GTCCAGCGCGAGCGCCCGCGGACAGCCACCCAGGAGGAGCGCGCCAAGATCACCGGCCGG 240

Query 241    GAGCGCAAGGCCATGAGGGTCTCTGCCGCTGGTGGTC 276
Sbjct 241    GAGCGCAAGGCCATGAGGGTCTCTGCCGCTGGTGGTC 276
```

>[db1|AB069664.1](#) Ursus thibetanus DRD4 gene fordopamine receptor D4, partial cds
Length=303

Score = 315 bits (348), Expect = 1e-82
Identities = 248/303 (82%), Gaps = 27/303 (9%)
Strand=Plus/Plus

```
Query 1      TTTCTTCTACCTGCCCGCTCATGCTGCTGCTTACTGGGCCACGTTCCGGGGCCTGCGG 60
Sbjct 1      TTTCTTCTACCTGCCCGCTCATGCTGCTGCTTACTGGGCCACTTTTCGGGGCCTGCAG 60

Query 61     CGCTGGGAGCGCGCTCGCCAGGCCAAGCTGCAC TGCCGGGCGCCTCGTcgagccagcggc 120
Sbjct 61     CGCTGGGAGGCTGCCGCTCGCCAGGCTGCACGCGCGGCGCGCGCGCGCCAGCGC 120

Query 121    cccggcccaaccgccccccgaggtcgggcgag-----ccccccgagcgc 162
Sbjct 121    CTTGGCCCGCGCGCCCCCGAGCGCGTCCGAGAGACCCCGAGGGCCTTCTGCCCCCTGAGAGCC 180

Query 163    gtgcgcccccccgag-----cgccgtccagagcgagccggcggcGGCAGGCACCCAGG 213
Sbjct 163    GTTCCGCCCCCGAGTCTCTGACGCCATCCCGCCGAGCCGCCCGCTGCAGGCAAGCAAG 240

Query 214    AGGAGCGCGCCCAAGATCACCGGCGCGGAGGCCCAAGGCCATGAGGGTCTTCCCGCTGGTG 273
Sbjct 241    AGGAGGCCACCAAGATCACCGCGCGGAGGCCCAAGGCCATGAGGGTCTTCCCGCTGGTG 300

Query 274    GTC 276
Sbjct 301    GTC 303
```

Exemple : Résultats

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 (bits)	E 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

Exemple : Résultats

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 (bits)	E 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

Exemple : Résultats

```
>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 ttcttctaccctgcccgcctcatgctgctgctctactgggccacgttcc 48  
|||||  
Sbjct 1 ttcttctaccctgcccgcctcatgctgctgctctactgggccacgttcc 48
```

```
Query 49 ggggcctgcggcgctgggaggcgctcgccaggccaagctgcactgccgg 99  
|||||  
Sbjct 49 ggggcctgcggcgctgggaggcgcgctcgggccaagctgcacggccgg 99
```

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

```
Query 215 ggaggcgcgccaagatcacggcgccgggagcgcaaggccatgagggtcct 252  
|||||  
Sbjct 332 ggagacgcgccaagatcacggcgccgggagcgcaaggccatgagggtcct 379
```

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

BLAST : ... vs Alignement Réel

Felis Catus/ Nyctereute

1	ttcttcctaccctgcccgtcatgtgctgctctactgggccacg	145	ggcgagc.....
1	ttcttcctaccctgcccgtcatgtgctgctctactgggccacg	181	ggc.agccccggagggcacccccggcccgccgcccccgacggcac
46	ttccggggcctgcggcgctgggaggcggctcgccaggccaagctg	152
46	ttccggggcctgcggcgctgggaggcggcgctcgggccaagctg	225	ccccgatgacacccccgacgccacccccctgcccccgcccccgcc
91	cactgccggggcgctcgtcgggccagcgggccccggcccaccgccc	153	ccccgacggcgctcgcccccccgacggcggtcccagccgagccgcc
91	cacggccgggacacggcgagaccagcgggccccggcccgcaccc	270	ccccgacggcgccgccccccccgcccggaccctgcggagcccag
136	cccga.ggt.....c	198	gcggcaggcaccccaggaggaggcgcgccaagatcacggccggga
136	cccgacggtacccccggccccccgcccccgacggcagccccgac	315	gtggcagccacgcaagcggagacgcgccaagatcacggccggga
		243	gcgcaaggccatgagggtcctgccggtggtggtc
		360	gcgcaaggccatgagggtcctgccggtggtggtc

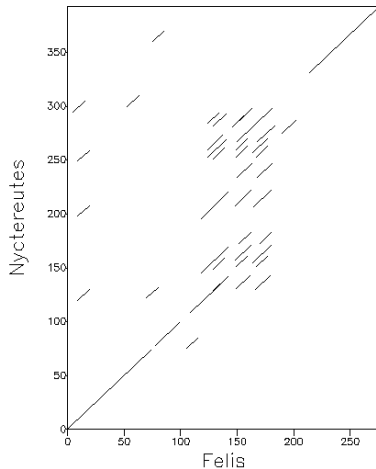
BLAST : ... vs Alignement Réel

Felis Catus/ Nyctereute

1	ttcttcctaccctgccgctcatgctgctgctctactgggccacg	145	ggcgagc.....
1	ttcttcctaccctgccgctcatgctgctgctctactgggccacg	181	ggc.agccccggagcggcacccccggcccgccgcccccgacggcac
46	ttccggggcctgcggcgtgggaggcggctcgccaggccaagctg	152
46	ttccggggcctgcggcgtgggaggcggcgtcgggccaagctg	225	ccccgatgacacccccgacgccacccccctgcccccgcccccgcc
91	actgcccggcgccctcgtcgccacagcgccccggccacccgcc	153	ccccgacgcgctcgcccccccgacgccgtcccagccgagccgcc
91	cacggccggacacgcgcgacaccagcgccccggcccgccaccc	270	ccccgacgcgccccgcccccgcccgaccctgcggagcccag
136	cccga.ggt.....c	198	gcggcaggcacccagga
136	cccgacggtacccccggcccccgcccccgacggcagccccgac	315	gtggcagccacgcaagc
		243	gcgcaaggccatgagggtcctgccggtggtggtc
		360	gcgcaaggccatgagggtcctgccggtggtggtc

BLAST : ... vs Alignement Réel

Felis Catus/ Nyctereute



- deux séquences peuvent toujours être alignées
- il existe toujours un (au moins) alignement de meilleur score S entre deux séquences (un MSP)

question : ce score est-il suffisamment élevé pour prouver une homologie ?

problème : peut-on trouver un MSP de meilleur score dans deux séquences aléatoires ?

- la **p-valeur (p-value)**

mesure la *Probabilité* que 2 séquences aléatoires de même longueur et de même composition possèdent un MSP de score $\geq S$

- la **p-valeur (p-value)**

mesure la *Probabilité* que 2 séquences aléatoires de même longueur et de même composition possèdent un MSP de score $\geq S$

- la **E-valeur (E-value)**

mesure l'*Esperance* E du nombre n de MSPs de score $\geq S$ dans 2 séquences aléatoires de même longueur et de même composition

- la **p-valeur (p-value)**

mesure la *Probabilité* que 2 séquences aléatoires de même longueur et de même composition possèdent un MSP de score $\geq S$

- la **E-valeur (E-value)**

mesure l'*Esperance* E du nombre n de MSPs de score $\geq S$ dans 2 séquences aléatoires de même longueur et de même composition

$$E = \sum_n p(n) \times n$$

- soient deux séquences a et b aléatoires suivant une distribution de probabilité connue
- on suppose que les MSPs sont données par les diagonales du dotplot
- plutôt que de décrire un alignement par des paires de lettres tirées aléatoirement, on peut le décrire par une suite de scores tirés aléatoirement
- on veut calculer l'esperance du nombre de MSPs de score $\geq S$

- Selon *Karlin et Altschul 1991* :

$$\mathbf{E\text{-}value} = Kmne^{-\lambda S} \qquad \mathbf{p\text{-}value} = 1 - e^{-\mathbf{e\text{-}value}}$$

avec m la taille de la séquence requête, n la taille de la banque de données, S le score du hit (K et λ dépendent de la matrice de score, K peut être ajusté en fonction du coût des gaps)

- si S est le score d'un hit
- le bit-score (score normalisé) est :

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

- l'expression de la E-value devient :

$$E(S) = mn2^{-S'}$$

Variation de la E-value

- si la taille de la séquence query augmente : la E-value
.....
- si la taille de la banque est divisée par deux : la E-value
.....
- si le score augmente : la E-value
.....
- quel bit-score pour obtenir une E-value de 0.05 pour une séquence de longueur 250 et une bd de longueur 50000000 ?
.....
- si on passe la E-value à 0.01, quel sera le bit-score ?
.....

Variation de la E-value

- si la taille de la séquence query augmente : la E-value **augmente**
- si la taille de la banque est divisée par deux : la E-value
.....
- si le score augmente : la E-value
- quel bit-score pour obtenir une E-value de 0.05 pour une
séquence de longueur 250 et une bd de longueur 50000000 ?
.....
- si on passe la E-value à 0.01, quel sera le bit-score ?

Variation de la E-value

- si la taille de la séquence query augmente : la E-value **augmente**
- si la taille de la banque est divisée par deux : la E-value **diminue**
- si le score augmente : la E-value
- quel bit-score pour obtenir une E-value de 0.05 pour une séquence de longueur 250 et une bd de longueur 50000000 ?
.....
- si on passe la E-value à 0.01, quel sera le bit-score ?

Variation de la E-value

- si la taille de la séquence query augmente : la E-value **augmente**
- si la taille de la banque est divisée par deux : la E-value **diminue**
- si le score augmente : la E-value **diminue**
- quel bit-score pour obtenir une E-value de 0.05 pour une séquence de longueur 250 et une bd de longueur 50000000 ?
.....
- si on passe la E-value à 0.01, quel sera le bit-score ?

Variation de la E-value

- si la taille de la séquence query augmente : la E-value **augmente**
- si la taille de la banque est divisée par deux : la E-value **diminue**
- si le score augmente : la E-value **diminue**
- quel bit-score pour obtenir une E-value de 0.05 pour une séquence de longueur 250 et une bd de longueur 50000000 ? **38 bits**
- si on passe la E-value à 0.01, quel sera le bit-score ?

Variation de la E-value

- si la taille de la séquence query augmente : la E-value **augmente**
- si la taille de la banque est divisée par deux : la E-value **diminue**
- si le score augmente : la E-value **diminue**
- quel bit-score pour obtenir une E-value de 0.05 pour une séquence de longueur 250 et une bd de longueur 50000000 ? **38 bits**
- si on passe la E-value à 0.01, quel sera le bit-score ? **40 bits**

Variations de la E-value

Mus musculus chromosome 5, clone RP23-301L9, complete sequence

Length = 212246

Score = 32.2 bits (16), Expect = 2.1

Identities = 19/20 (95%)

Strand = Plus / Plus

Query: 2 ttcattatgaagcacgagga 21

|||||

Sbjct: 136843 ttcattatgatgcacgagga 136862

Mus musculus BAC clone RP23-13L19 from chromosome 9, complete sequence

Length = 224108

Score = 30.2 bits (15), Expect = 8.1

Identities = 15/15 (100%)

Strand = Plus / Plus

Query: 6 ttatgaagcacgagg 20

|||||

Sbjct: 93798 ttatgaagcacgagg 93812

Lambda K H

1.37 0.711 1.31

Number of Hits to DB: 99,084,306

Number of Sequences: 2130505

Number of extensions: 2852

Number of successful extensions: 19

Number of sequences better than 10.0: 0

Number of HSP's better than 10.0 without gapping: 0

Number of HSP's successfully gapped in prelim test: 16

length of query: 21

length of database: 10,249,863,584

Variations de la E-value

Mus musculus chromosome 5, clone RP23-301L9, complete sequence
Length = 212246

Score = 32.2 bits (16), Expect = 2.1
Identities = 19/20 (95%)
Strand = Plus / Plus

Query: 2 ttcattatgaagcacgagga 21
 ||||| |||||
Sbjct: 136843 ttcattatgatgcacgagga 136862

Mus musculus, clone RP23-277C24, complete sequence
Length = 199946

Score = 32.2 bits (16), Expect = 2.1
Identities = 16/16 (100%)
Strand = Plus / Minus

Query: 1 attcattatgaagcac 16
 |||||
Sbjct: 69080 attcattatgaagcac 69065

Variations de la E-value

Query length : 21
Mus musculus, clone RP23-277C24, complete sequence
Length = 199946

Score = 32.2 bits (16), Expect = 7.6
Identities = 16/16 (100%)
Strand = Plus / Minus

Query: 1 attcattatgaagcac 16
 |||||
Sbjct: 69080 attcattatgaagcac 69065

Query length : 20
Mus musculus, clone RP23-277C24, complete sequence
Length = 199946

Score = 32.2 bits (16), Expect = 5.1
Identities = 16/16 (100%)
Strand = Plus / Minus

Query: 1 attcattatgaagcac 16
 |||||
Sbjct: 69080 attcattatgaagcac 69065

Les différents programmes BLAST

Query \ Database	nucléique	protéique	nucléique traduit
nucléique	blastn	x	x
protéique	x	blastp	tblastn
nucléique traduit	x	blastx	tblastx

Le bon programme pour la bonne requête

extrait de "BLAST Program Selection Guide"

- MEGABLAST is the tool of choice to identify a nucleotide sequence
- Discontiguous MEGABLAST is better at finding nucleotide sequences similar, but not identical, to your nucleotide query
- les pages "Search for short nearly exact matches"
 - nucleotide : useful for primer or short nucleotide searches
 - proteins : optimized to find matches to a short peptide
 - principales différences :
 - taille de mots plus petite
 - suppression des filtres
 - relâchement de la E-value
 - matrice de score PAM30 (au lieu de BLOSUM62) pour les protéines

Evolutions de Blast : PSI-Blast

PSI-Blast is designed for more sensitive protein-protein similarity searches

Position Specific Iterated BLAST

- 1 recherche initiale avec BLASTp
- 2 construction d'un alignement multiple, puis d'un **profil**
 - à partir d'un alignement multiple des meilleurs hits
 - construit une matrice position-spécifique :
 - chaque colonne représente un AA
 - chaque ligne une position dans l'alignement
- 3 nouvelle recherche avec le profil et modification

réitère le processus un certain nombre de fois ou jusqu'à convergence

Profil - exemple

```
Alignement multiple:  # Pure Frequency Matrix
                        # Columns are amino acid counts A->Z
                        # Rows are alignment positions 1->n
                        Simple
T-VAAPSVFIFPPSDEQ      Name          mymatrix
A-DAAPTVSIFPPSSEQ      Length       17
A-NAAPTVSIFPPSTZZ      Maximum score 60
D-PVAPTVLIFPPAADQ      Thresh       75
DPPIAPTVLLFPPSADQ      Consensus    APPAAPTVLIFPPSADQ
2002000000000000000100000000
00000000000000000100000000000
000100000000000102000001000000
3000000010000000000001000000
50000000000000000000000000000
00000000000000000500000000000
0000000000000000000140000000
00000000000000000000000500000
0000010000020000002000000000
0000000040010000000000000000
0000050000000000000000000000
00000000000000000500000000000
00000000000000000500000000000
1000000000000000000400000000
2001000000000000000110000000
0002200000000000000000000010
0000000000000000040000000010
```


Pattern Hit Initiated BLAST

- pour les séquences protéiques
- entrée : une séquence et un motif (expression régulière à la Prosite)
- restriction de la banque aux séquences pour lesquelles le motif est retrouvé
- puis application de BLAST
- couplage possible avec PSI-Blast

On peut classer les résultats d'une méthode en 4 catégories :

On peut classer les résultats d'une méthode en 4 catégories :

VP les vrais positifs (*classé positif et bien positif*)

On peut classer les résultats d'une méthode en 4 catégories :

VP les vrais positifs (*classé positif et bien positif*)

FP les faux positifs (*classé positif mais réellement négatif*)

ex. alignements parasites, prédiction pour une fonction que la
séquence n'a pas en réalité. . .

On peut classer les résultats d'une méthode en 4 catégories :

VP les vrais positifs (*classé positif et bien positif*)

FP les faux positifs (*classé positif mais réellement négatif*)

ex. alignements parasites, prédiction pour une fonction que la
séquence n'a pas en réalité. . .

FN les faux négatifs (*classé négatif mais réellement positif*)

ex. alignements perdus, prédiction négative pour une fonction
que la séquence a en réalité. . .

On peut classer les résultats d'une méthode en 4 catégories :

VP les vrais positifs (*classé positif et bien positif*)

FP les faux positifs (*classé positif mais réellement négatif*)

ex. alignements parasites, prédiction pour une fonction que la
séquence n'a pas en réalité. . .

FN les faux négatifs (*classé négatif mais réellement positif*)

ex. alignements perdus, prédiction négative pour une fonction
que la séquence a en réalité. . .

VN les vrais négatifs (*classé négatif et bien négatif*)

Sensibilité et spécificité

On peut classer les résultats d'une méthode en 4 catégories :

VP les vrais positifs (*classé positif et bien positif*)

FP les faux positifs (*classé positif mais réellement négatif*)

ex. alignements parasites, prédiction pour une fonction que la séquence n'a pas en réalité. . .

FN les faux négatifs (*classé négatif mais réellement positif*)

ex. alignements perdus, prédiction négative pour une fonction que la séquence a en réalité. . .

VN les vrais négatifs (*classé négatif et bien négatif*)

$$\text{sensibilité} = \frac{VP}{VP+FN} \quad \text{spécificité} = \frac{VN}{VN+FP}$$

■ sensibilité : capacité de la méthode à ne pas "louper" de Positifs

■ spécificité : capacité de la méthode à ne pas "ramener" de Négatifs