Module Projet Bioinformatique

Module Projet Bioinformatique
L3 GBI-DL
Carène Rizzon
2023-2024

Programme du module

- Principes de base pour l'analyse des séquences: alignement de séquences 2 à 2 (scores, matrices, Needleman et Wunch, Smith –Waterman, présentation simple de BLAST)

- Introduction à l'environnement Unix/Linux (rappels): filtres et utilitaires de base : présentation des notions avec un support cours et exercices
(pour s'entraîner: https://cocalc.com/ ou https://www.cygwin.com/)

- Exercices théoriques et sur machine autour de la notion d'alignements 2 à 2 :
    - application sur site web du package EMBOSS
    - utilisation sous terminal du package EMBOSS
    - BLAST sur site web (Galaxy, NCBI)
    - BLAST sous terminal

- Projet (par binômes) analyse fine de séquences de copies d'éléments transposables (avec cours sur les ET) via BLAST et EMBOSS et établissement d'une base de données SQL.

- autoinscription pour le module: clé= **PROJETBIOINFO24**

Module Projet Bioinformatique

Comparaison de séquences 2 à 2
Principes de base

L3 GBI/DL
Carène Rizzon

Université Evry Val d'Essonne
2023-2024

- **Aligner des séquences pour les comparer**
- **Score d’alignement**
- **Algorithmes d’alignement de 2 séquences (ADN)**

Documents et sources

- **Sites web**

  - http://bioinfo-fr.net/

  - Jeu Phylo: bioinformatique participative http://phylo.cs.mcgill.ca

  - Interstices (sciences du numérique) https://interstices.info/

  - Site de la Société Française de Bioinformatique (SFBI): http://www.sfbi.fr/

  - JeBiF: le site des jeunes bioinformaticiens de France: https://jebif.fr/fr/

- **Ouvrages:**

  - Bioinformatique, cours et cas pratique, G. Deléage, M. Gouy, 2013, ed. Dunod

  - Bio-informatique - Principes d'utilisation des outils, D. Tagu, JL Risler, coord., 2010, ed. Quae.

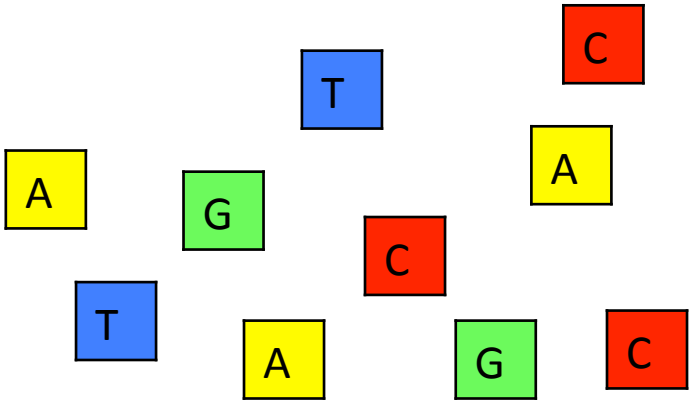
Objectifs

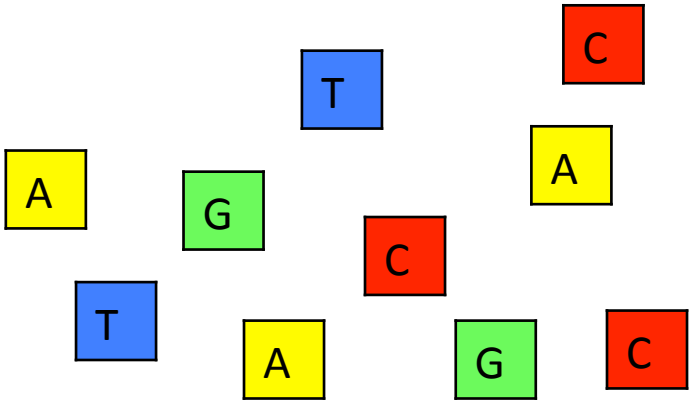
- **Quand compare-t-on des séquences ?**
  Aujourd'hui en Biologie, tout le temps!

- **Pourquoi ? 2 types principaux de comparaisons:**

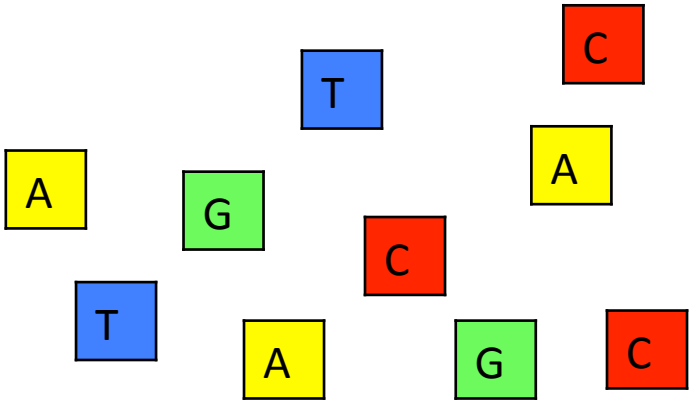
  Recherche de séquences identiques
  - Assemblage des séquences en contigs
  - Localisation d'ARNm sur le génomique

  Recherche de séquences homologues
  - Détection de gènes
  - Prédiction de fonction et de structure
  - Etudes évolutives (phylogénie, recherche de synténies)
  - Etudes dynamiques des génomes
  - Génomique comparée
  - Réseaux métaboliques


C
T
A
A
G
C
T
A
G
C

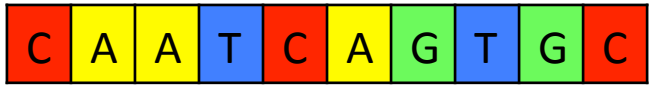


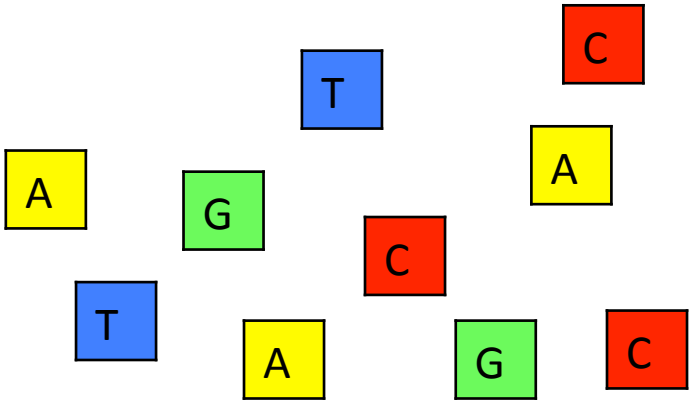
Une séquence:



Une séquence:

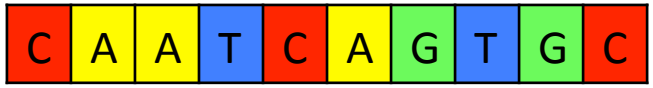
Une autre séquence:





Une séquence:

Une autre séquence:



Sens de lecture

Est-ce que ces séquences sont identiques?

Séquence 1

Séquence 2

Est-ce que ces séquences sont identiques?

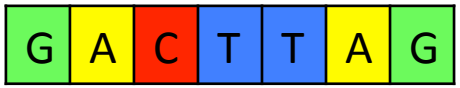
Séquence 1

Séquence 3

Est-ce que ces séquences sont identiques?

Séquence 1

Séquence 4



Est-ce que ces séquences sont identiques?

Séquence 1 

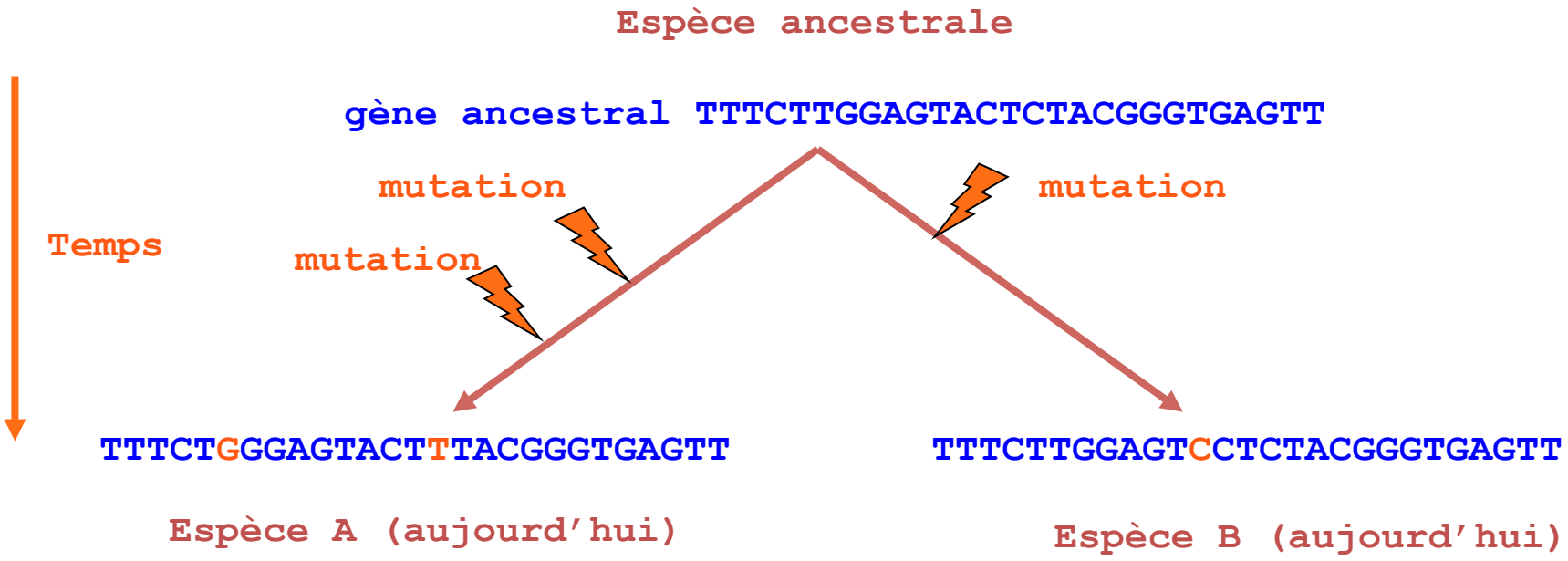
Séquence 5 

Non, mais elles se ressemblent beaucoup, elles sont similaires

Homologie ≠ Similitude

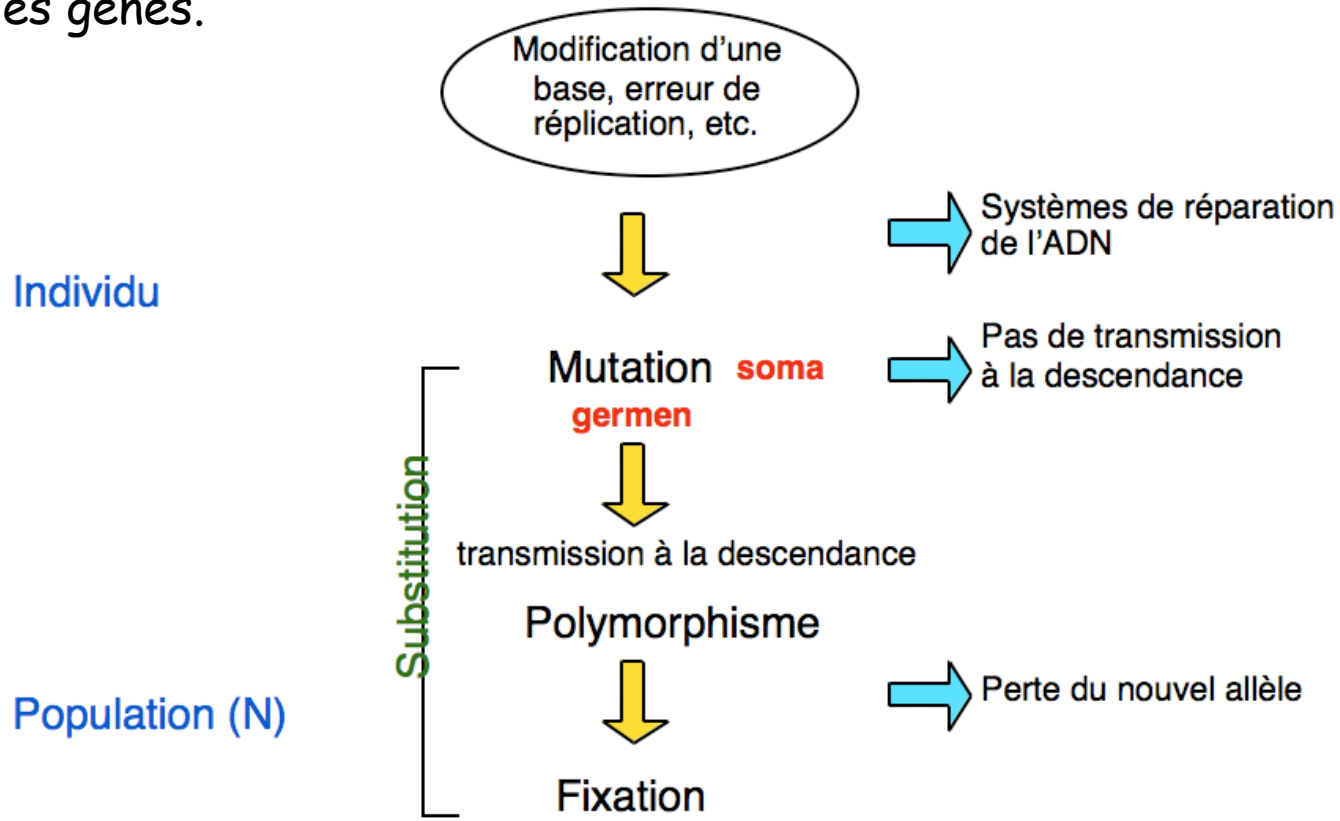
- homologie = inférence (existe un ancêtre commun)
- similarité = pas d'inférence (qualitatif ou quantitatif)

- **Sur quel(s) principe(s) compare-t-on des séquences?**
  - Les séquences de 2 molécules de fonctions apparentées vont en général présenter des ressemblances
  - Réciproquement, deux molécules dont les séquences présentent des ressemblances ont probablement des fonctions apparentées.



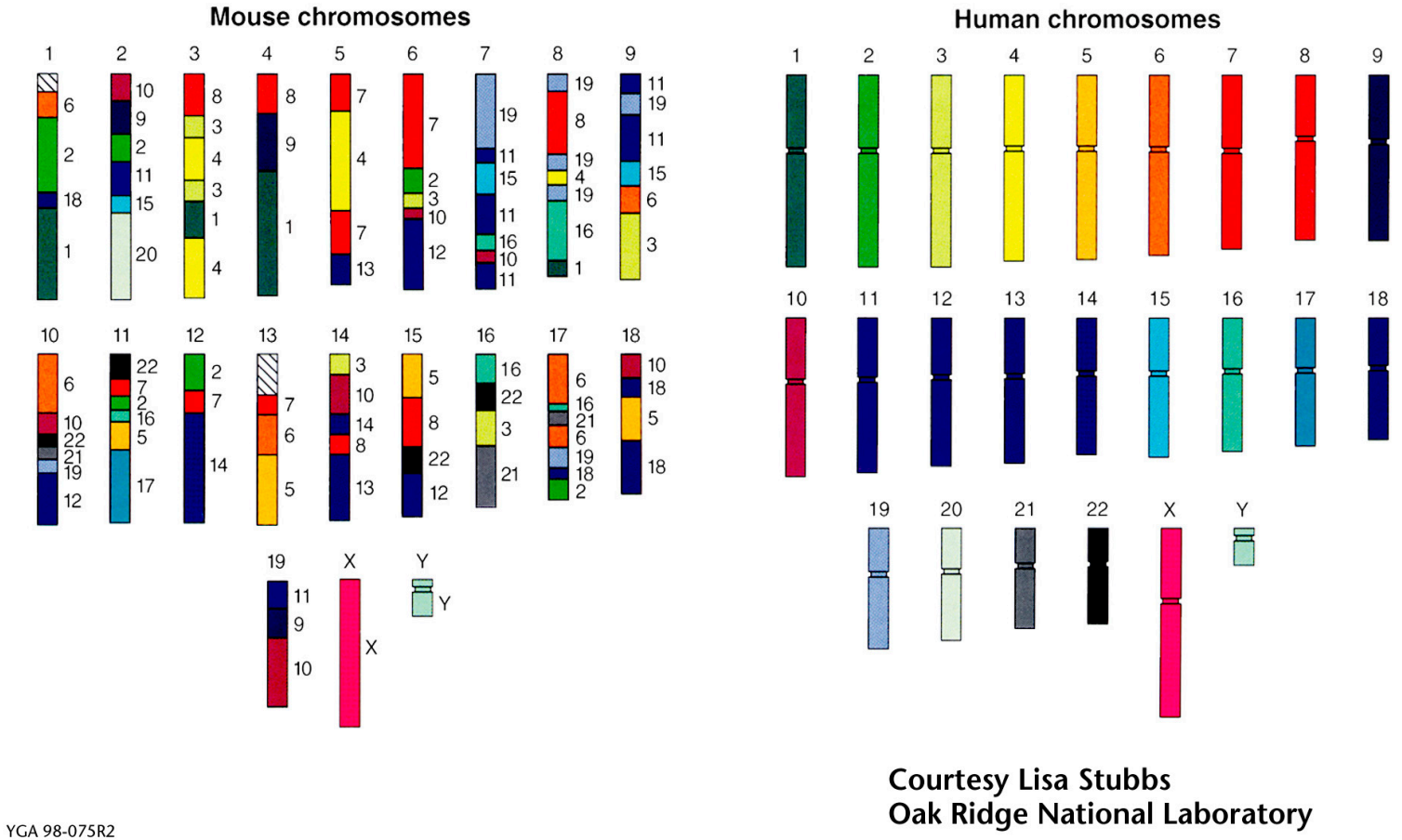
Aujourd'hui: le néo-darwinisme

- Lois de Mendel -> les variations génétiques sont générées par mutation
- **théorie synthétique de l'évolution: néo-Darwinisme** : les mutations sont la source de la variation génétique mais la sélection naturelle façonne le contenu génétique des populations et les processus de substitution chez les gènes.



Duret, HDR, 2001

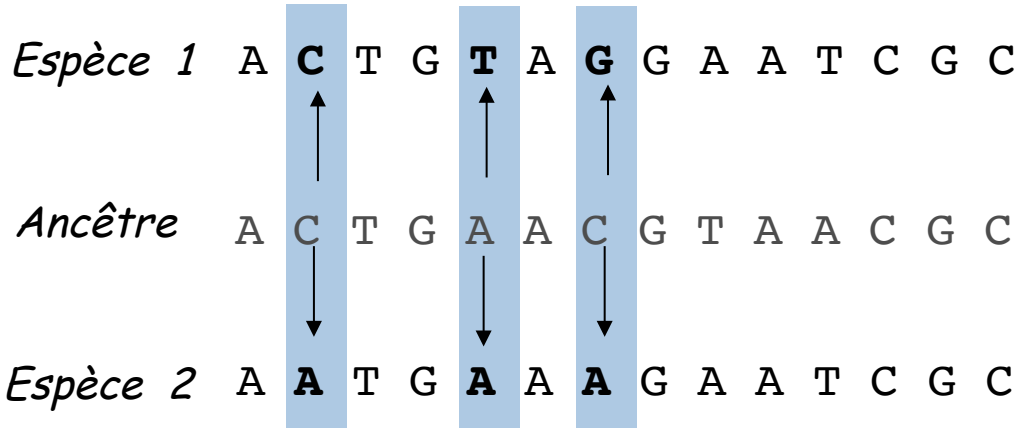
Mouse and Human Genetic Similarities



Source: Lisa Stubbs, Oak Ridge National Laboratory.

Méthodes phylogénétiques de distance:

On travaille avec des **distances évolutives** = Une fonction des différences observées entre 2 séquences

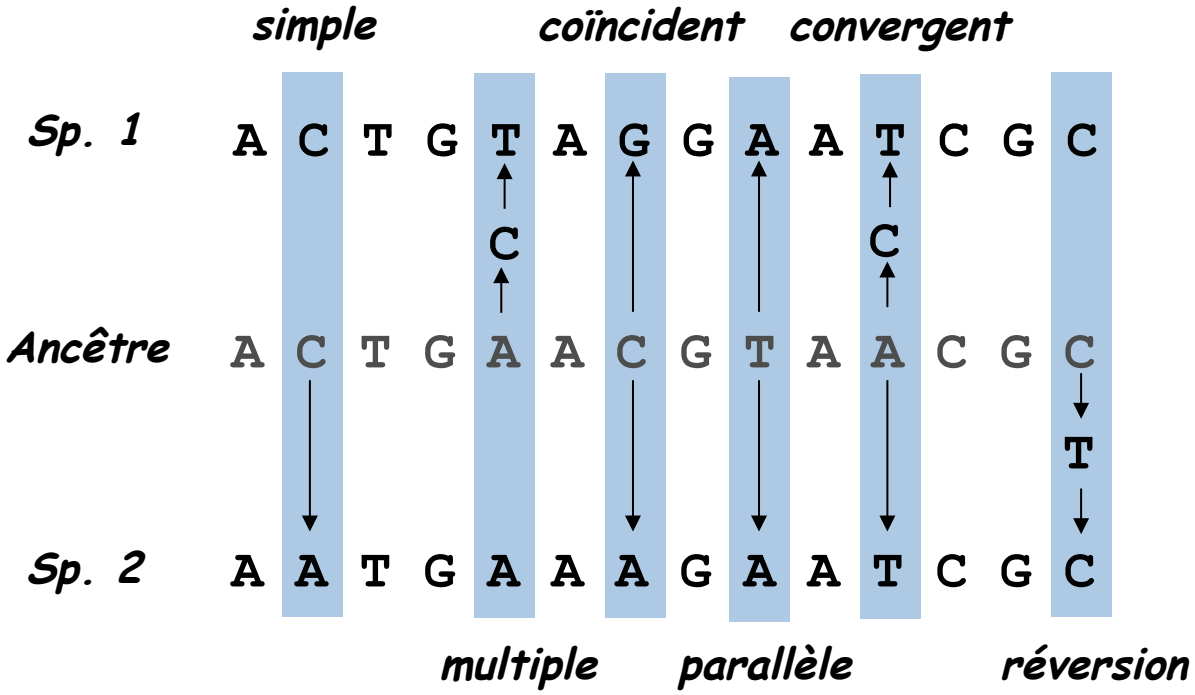


Distance observée (Sp1, Sp2) = 3 / 14

Mais substitutions cachées: distance vraie ?

Méthodes phylogénétiques de distance:

On travaille avec des **distances évolutives** = Une fonction des différences observées entre 2 séquences



D_{Obs} (Sp1, Sp2) = 3 / 14

D_{vrai} (Sp1, Sp2) = 12 / 14

Alignement : manière de placer une séquence sur une autre pour rendre clair les correspondances entre des caractères ou des sous-chaînes semblables.

Séquence 1: A T G C G T C G T T Séquence 2: A T C C G A C G T T

Alignement possible

```
A T G C G T C G T T
: :   : :   : : : :
A T C C G A C G T T
```

Longueur: 10 nucléotides
8 appariements identiques sur 10 sites
ici les séquences sont identiques à 80%

```
INS_HUMAN          1 MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFY     50
                     |||.:..|||||||||||.|.|..|||.|||||.||||||||||||||||
INS1_MOUSE         1 MALLVHFLPLLALLALWEPKPTQAFVKQHLCGPHLVEALYLVCGERGFFY     50

INS_HUMAN         51 TPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSIC    100
                     |||:|||.||.||.|:||||.|  |.||.||||.:.||||||:|||||||
INS1_MOUSE        51 TPKSRREVEDPQVEQLELGGSP--GDLQTLALEVARQKRGIVDQCCTSIC     98

INS_HUMAN        101 SLYQLENYCN    110
                     ||||||||||
INS1_MOUSE        99 SLYQLENYCN    108
```

```
BT006808.1 (HS)    1 ----------------------------ATGGCCCTGTGGATGCGCCTCC     22
                                                 ||||||||||.|.|||.|.|||
BC145868.1 (MM)    1 CCATCAGCAAGCAGGTCATTGTTTCAACATGGCCCTGTTGGTGCACTTCC     50

BT006808.1        23 TGCCCCTGCTGGCGCTGCTGGCCCTCTGGG-GACCTGACCCAGCCGCA-G     70
                     |.||||||||||.|||||.|||||||||| |.||..|||||  |.|| |
BC145868.1        51 TACCCCTGCTGGCCCTGCTTGCCCTCTGGGAGCCCAAACCCA--CCCAGG     98

BT006808.1        71 CCTTTGTGAACCAACACCTGTGCGGCTCACACCTGGTGGAAGCTCTCTAC    120
                     |.|||||.||.||.||.||.||.||..|.||||||||.||.|||||||||
BC145868.1        99 CTTTTGTCAAACAGCATCTTTGTGGTCCCCACCTGGTAGAGGCTCTCTAC    148

BT006808.1       121 CTAGTGTGCGGGGAACGAGGCTTCTTCTACACACCCAAGACCCGCCGGGA    170
                     ||.|||||.|||||.||.|||||||||||||||||||||.|||||||.||
BC145868.1       149 CTGGTGTGTGGGGAGCGTGGCTTCTTCTACACACCCAAGTCCCGCCGTGA    198

BT006808.1       171 GGCAGAGGACCTGCAGGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTG    220
                     .|..|||||||..||.||||..||..||||||||||.||..||||.||.|
BC145868.1       199 AGTGGAGGACCCACAAGTGGAACAACTGGAGCTGGGAGGAAGCCCCGGGG    248

BT006808.1       221 CAGGCAGCCTGCAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGC    270
                           .|||.|||.|||||||..||||||.|.|||.||||||||||||||
BC145868.1       249 ------ACCTTCAGACCTTGGCGTTGGAGGTGGCCCGGCAGAAGCGTGGC    292

BT006808.1       271 ATTGTGGAACAATGCTGTACCAGCATCTGCTCCCTCTACCAGCTGGAGAA    320
                     ||||||||.||.|||||.||||||||||||||||||||||||||||||||
BC145868.1       293 ATTGTGGATCAGTGCTGCACCAGCATCTGCTCCCTCTACCAGCTGGAGAA    342

BT006808.1       321 CTACTGCAACTAG-----------------------------------    333
                     ||||||||||||.
BC145868.1       343 CTACTGCAACTAAGGCCCACCTCGACCCGCCCCACCCCTTTGCAATGAAT    392
```

http://mobyle.pasteur.fr/cgi-bin/portal.py

```
#=======================================
#
# Aligned_sequences: 2
# 1: SYE_ARATH
# 2: SYE_TOBAC
# Matrix: EPAM250
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 49
# Identity:      16/49 (32.7%)
# Similarity:    26/49 (53.1%)
# Gaps:          16/49 (32.7%)
# Score: 83.5
#
#
#=======================================

SYE_ARATH          1 MASLVYGTPWLRVRSLPELAPAFLRRRQSSLFYC-SRRSFA--------     40
                     ||:|. ::||:|||.:|||      :...||:|| :.:|:.
SYE_TOBAC          1 MATLA-AAPWFRVRLIPEL------KNSQSLLYCRGNHSYRQSLCSRRR     42
```

SYE_ARATH:Glutamyl-tRNA synthetase chez l'arabette
SYE_TOBAC:Glutamyl-tRNA synthetase chez le tabac
(pour acide aminé E, acide glutamique

Comparaison avec score

2 séquences
de longueur
16

```
AATTGGAGCAGCCGTA
|   |   ||||   |||
ATCTCTAGCACGCGTG
```

```
1001001111001110  score=9
```

```
2002002222002220  score=18
```

Matrices de substitution ADN

- **Comparaisons ADN**

La matrice identité

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

Smith et Waterman

	A	T	G	C
A	1.0	-0.9	-0.9	-0.9
T	-0.9	1.0	-0.9	-0.9
G	-0.9	-0.9	1.0	-0.9
C	-0.9	-0.9	-0.9	1.0

Prise en compte des transitions/transversions

	A	T	C	G
A	1.0	-0.9	-0.9	-0.5
T	-0.9	1.0	-0.5	-0.9
C	-0.9	-0.5	1.0	-0.9
G	-0.5	-0.9	-0.9	1.0

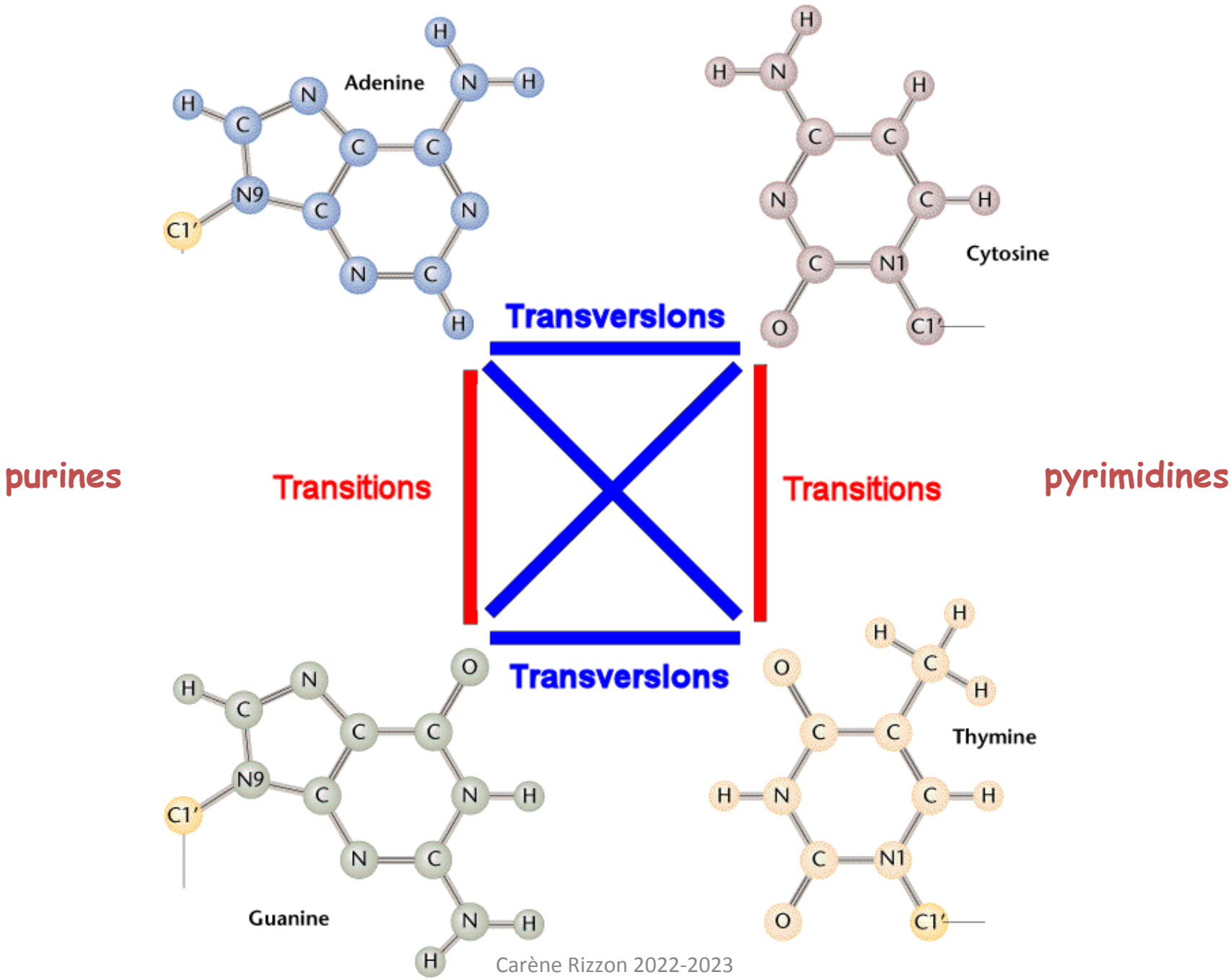
Score maximum recherché

1 si X = Y
-0.5 si X : Y = transition
-0.9 si X : Y = transversion

	A	T	C	G
A	0	2	2	1
T	2	0	1	2
C	2	1	0	2
G	1	2	2	0

Score minimum recherché

0 si X = Y
1 si X : Y = transition
2 si X : Y = transversion


Adenine
Cytosine
Transversions
purines
Transitions
Transitions
pyrimidines
Transversions
Guanine
Thymine

Comparaison avec score

2 séquences
de longueur
16

```
AATTGGAGCAGCCGTA
|   |   ||||   |||
ATCTCTAGCACGCGTG
```

```
1 -0.9 -0.5 1 -0.9 -0.9 1 1 1 1 -0.9 -0.9 1 1 1 -0.5
Score= 1*9 -0.9*5 -0.5*2= 3,5
```

	A	T	C	G
A	1.0	-0.9	-0.9	-0.5
T	-0.9	1.0	-0.5	-0.9
C	-0.9	-0.5	1.0	-0.9
G	-0.5	-0.9	-0.9	1.0

Matrices de substitution protéique

Code génétique

ACIDE AMINE	
phénylalanine	F
leucine	L
isoleucine	I
méthionine	M
valine	V
sérine	S
proline	P
thréonine	T
alanine	A
tyrosine	Y
histidine	H
glutamine	Q
asparagine	N
lysine	K
acide aspartique	D
acide glutamique	E
cystéine	C
tryptophane	W
arginine	R
glycine	G

nucléotide n°1	nucléotide en n°2 U		nucléotide en n°2 C		nucléotide en n°2 A		nucléotide en n°2 G		nucléotide en n°3
U	UUU	F	UCU	S	UAU	Y	UGU	C	U
	UUC		UCC		UAC		UGC		C
	UUA	L	UCA		UAA	*	UGA	*	A
	UUG		UCG		UAG		UGG	W	G
C	CUU	L	CCU	P	CAU	H	CGU	R	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Q	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	I	ACU	T	AAU	N	AGU	S	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	K	AGA	R	A
	AUG	M	ACG		AAG		AGG		G
G	GUU	V	GCU	A	GAU	D	GGU	G	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	E	GGA		A
	GUG		GCG		GAG		GGG		G

http://svt.ac-dijon.fr/schemassvt/article.php3?id_article=1409

Matrices de substitution protéique

ACIDE AMINE	
phénylalanine	F
leucine	L
isoleucine	I
méthionine	M
valine	V
sérine	S
proline	P
thréonine	T
alanine	A
tyrosine	Y
histidine	H
glutamine	Q
asparagine	N
lysine	K
acide aspartique	D
acide glutamique	E
cystéine	C
tryptophane	W
arginine	R
glycine	G

PAM250

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12	0	-2	-3	-2	-3	-4	-5	-5	-5	-3	-4	-5	-5	-2	-6	-2	-4	0	-8
S	0	2	1	1	1	1	1	0	0	-1	-1	0	0	-2	-1	-3	-1	-3	-3	-2
T	-2	1	3	0	1	0	0	0	0	-1	-1	-1	0	-1	0	-2	0	-3	-3	-5
P	-3	1	0	6	1	-1	-1	-1	-1	0	0	0	-1	-2	-2	-3	-1	-5	-5	-6
A	-2	1	1	1	2	1	0	0	0	0	-1	-2	-1	-1	-1	-2	0	-4	-3	-6
G	-3	1	0	-1	1	5	0	1	0	-1	-2	-3	-2	-3	-3	-4	-1	-5	-5	-7
N	-4	1	0	-1	0	0	2	2	1	1	2	0	1	-2	-2	-3	-2	-4	-2	-4
D	-5	0	0	-1	0	1	2	4	3	2	1	-1	0	-3	-2	-4	-2	-6	-4	-7
E	-5	0	0	-1	0	0	1	3	4	2	1	-1	0	-2	-2	-3	-2	-5	-4	-7
Q	-5	-1	-1	0	0	-1	1	2	2	4	3	1	1	-1	-2	-2	-2	-5	-4	-5
H	-3	-1	-1	0	-1	-2	2	1	1	3	6	2	0	-2	-2	-2	-2	-2	0	-3
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6	3	0	-2	-3	-2	-4	-4	2
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5	0	-2	-3	-2	-5	-4	-3
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6	2	4	2	0	-2	-4
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5	2	4	1	-1	-5
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6	2	2	-1	-2
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4	-1	-2	-6
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9	7	0
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	0
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

Matrices de substitution protéique

ACIDE AMINE	
phénylalanine	F
leucine	L
isoleucine	I
méthionine	M
valine	V
sérine	S
proline	P
thréonine	T
alanine	A
tyrosine	Y
histidine	H
glutamine	Q
asparagine	N
lysine	K
acide aspartique	D
acide glutamique	E
cystéine	C
tryptophane	W
arginine	R
glycine	G

PAM250

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12	0	-2	-3	-2	-3	-4	-5	-5	-5	-3	-4	-5	-5	-2	-6	-2	-4	0	-8
S	0	2	1	1	1	1	1	0	0	-1	-1	0	0	-2	-1	-3	-1	-3	-3	-2
T	-2	1	3	0	1	0	0	0	0	-1	-1	-1	0	-1	0	-2	0	-3	-3	-5
P	-3	1	0	6	1	-1	-1	-1	-1	0	0	0	-1	-2	-2	-3	-1	-5	-5	-6
A	-2	1	1	1	2	1	0	0	0	0	-1	-2	-1	-1	-1	-2	0	-4	-3	-6
G	-3	1	0	-1	1	5	0	1	0	-1	-2	-3	-2	-3	-3	-4	-1	-5	-5	-7
N	-4	1	0	-1	0	0	2	2	1	1	2	0	1	-2	-2	-3	-2	-4	-2	-4
D	-5	0	0	-1	0	1	2	4	3	2	1	-1	0	-3	-2	-4	-2	-6	-4	-7
E	-5	0	0	-1	0	0	1	3	4	2	1	-1	0	-2	-2	-3	-2	-5	-4	-7
Q	-5	-1	-1	0	0	-1	1	2	2	4	3	1	1	-1	-2	-2	-2	-5	-4	-5
H	-3	-1	-1	0	-1	-2	2	1	1	3	6	2	0	-2	-2	-2	-2	-2	0	-3
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6	3	0	-2	-3	-2	-4	-4	2
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5	0	-2	-3	-2	-5	-4	-3
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6	2	4	2	0	-2	-4
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5	2	4	1	-1	-5
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6	2	2	-1	-2
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4	-1	-2	-6
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9	7	0
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	0
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

Matrices de substitution protéiques

Matrice PAM 250

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	2	4

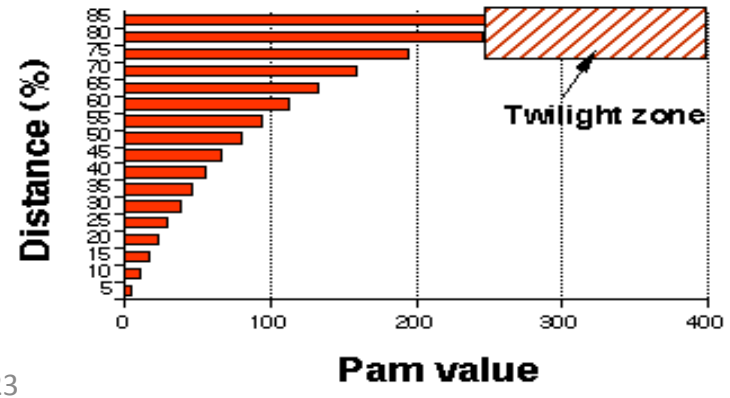
Matrices PAM:

- Conviennent pour les séquences avec un ancêtre commun

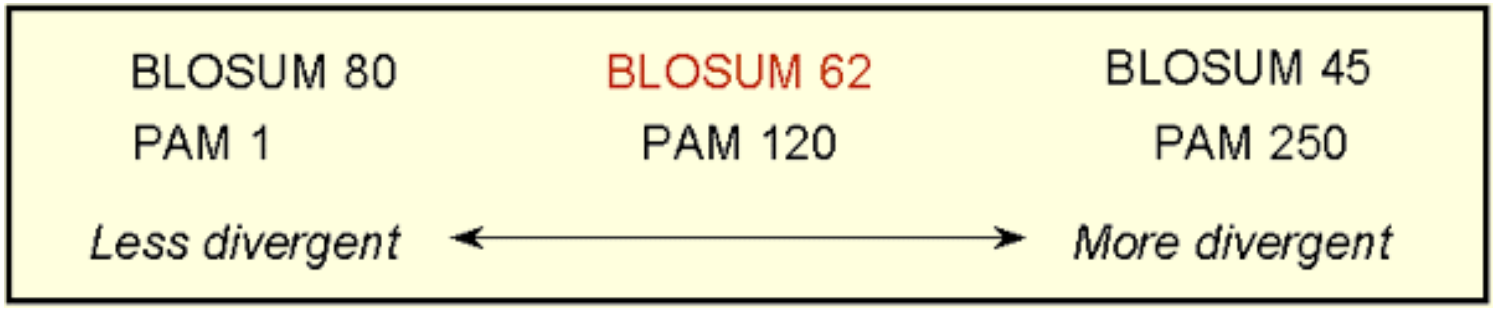
- Choix de la matrice en fonction de l'évolution supposée

- Si distance mutationnelle inconnue, essayer plusieurs PAM

- Alignements sur la totalité de la longueur des séquences, donc incluant les régions très semblables mais aussi les régions divergentes.



Correspondances entre matrices BLOSUM et matrices PAM



Comparaison avec score

2 séquences
de longueur
16

```
AATTGGAGCAGCCGTA
|   |   ||||   |||
ATCTCTAGCACGCGTG
```

```
1 -0.9 -0.5 1 -0.9 -0.9 1 1 1 1 -0.9 -0.9 1 1 1 -0.5
Score= 1*9 -0.9*5 -0.5*2= 3,5
```

	A	T	C	G
A	1.0	-0.9	-0.9	-0.5
T	-0.9	1.0	-0.5	-0.9
C	-0.9	-0.5	1.0	-0.9
G	-0.5	-0.9	-0.9	1.0

Comparaison avec score

2 séquences protéiques de longueur 6

M A G N M K
:
M G P D L C

Avec la matrice PAM250:

6 1 -1 2 4 -5
Score= 7

Question 6: Si on utilise la matrice de comparaison suivante et un score de gap égal à 3, quel est le « meilleur » alignement entre les alignements 1 et 2 ?

	A	T	G	C
A	0	2	2	2
T	2	0	2	2
G	2	2	0	2
C	2	2	2	0

```
Alignement 1 :

-ATTGC-GTGC
 | ||| ||||
AA-TGCGGTGC
```

```
Alignement 2 :

ATTGC-GTGC
| ||| ||||
AATGCGGTGC
```

Question 6: Si on utilise la matrice de comparaison suivante et un score de gap égal à 3, quel est le « meilleur » alignement entre les alignements 1 et 2 ?

	A	T	G	C
A	0	2	2	2
T	2	0	2	2
G	2	2	0	2
C	2	2	2	0

Alignement 1 :

```
-ATTGC-GTGC
 | ||| ||||
AA-TGCGGTGC
```

Alignement 2 :

```
ATTGC-GTGC
| ||| ||||
AATGCGGTGC
```

Correction :

Alignement 1 :

Score : 3+0+3+3*0+3+4*0=9

Alignement 2 :

Score : 0+2+3*0+3+4*0=5

Avec cette matrice de comparaison il faut **minimiser** les scores donc le meilleur alignement entre l'alignement 1 et l'alignement 2 est le 2.

Alignement de 2 séquences

Séquence 1: A T G C G T C G T T

Séquence 2: A T C C G C G T C

Alignement 1:

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

7 appariements
5 brèches

Alignement 2:

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

7 appariements
2 mésappariements
1 brèche

Alignement de 2 séquences

- **Situations d'alignement**

Situation		Événements biologiques
Appariements (matches)	→	Conservation
Mésappariements (mismatches)	→	Mutation / Substitution
Brèches (gaps)	→	Insertion / Délétion (indel)

```
INS_HUMAN     ... G G G P G A G S L Q P L A L E ...
INS1_MOUSE    ... G G S P - - G D L Q T L A L E ...
INS2_MOUSE    ... G G G P G A G D L Q T L A L E ...
                  * *   *     *   * *   * * * *
```

- **Alignement optimal**

  Minimiser le nombre de mésappariements et brèches selon les critères définis pour le calcul d'un score

Alignement global/alignement local

alignement global

Séquence 1 CCAATAAGCCATCTAAAGCGAAATGCCCCTTTCAAGCACACCTTATGACAATGGACTGCCGACACCTCTGTCATCACTGCCCAATAAGCCATC

Séquence 2 CCAATAAGCCATCTAAAGCGAAATGCCCCTTATGGTAGTAGTCAAGCACACCTTATGACAATGGACTGCCGACACCTCTGTCATCACTGCCCAATAAGCCATC

Séquence 3 CCCTTTCAAGCACACCTTATGACAATGGACTGCCGACACCTCTGTCATCACTGCCCAATAAGCCATC

Alignement des séquences 1 et 2

```
CCAATAAGCCATCTAAAGCGAAATGCCCCTTTCAAGCACACCTTATGA----------CAATGGACTGCCGACACCTCTGTCATCACTGCCCAATAAGCCATC
CCAATAAGCCATCTAAAGCGAAATGCCCCTTATGGTAGTAGTCAAGCACACCTTATGACAATGGACTGCCGACACCTCTGTCATCACTGCCCAATAAGCCATC
```

Alignement des séquences 2 et 3

```
CCAATAAGCCATCTAAAGCGAAATGCCCCTTATGGTAGTAGTCAAGCACACCTTATGACAATGGACTGCCGACACCTCTGTCATCACTGCCCAATAAGCCATC
------------CCC-----TTTCAAGC-----ACACC-----TTATGACAATGGACTGCC—GACAC—CTCTGTCATCACTGCCCAATAAGCCATC-------
```

Alignement global/alignement local

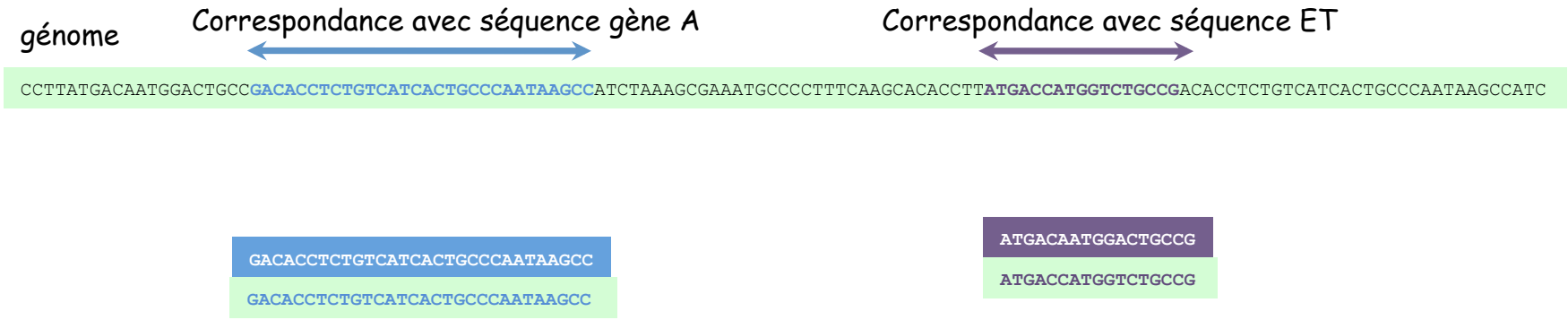
alignement local

Séquence gène A

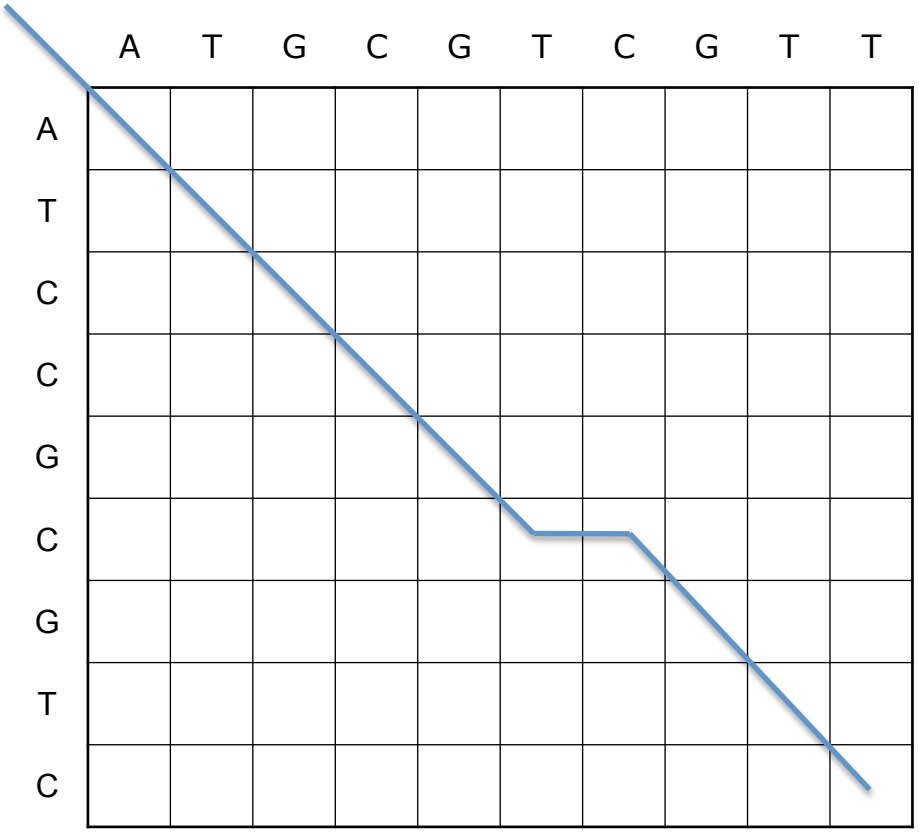
GACACCTCTGTCATCACTGCCCAATAAGCC

Séquence ET

ATGACAATGGACTGCCG

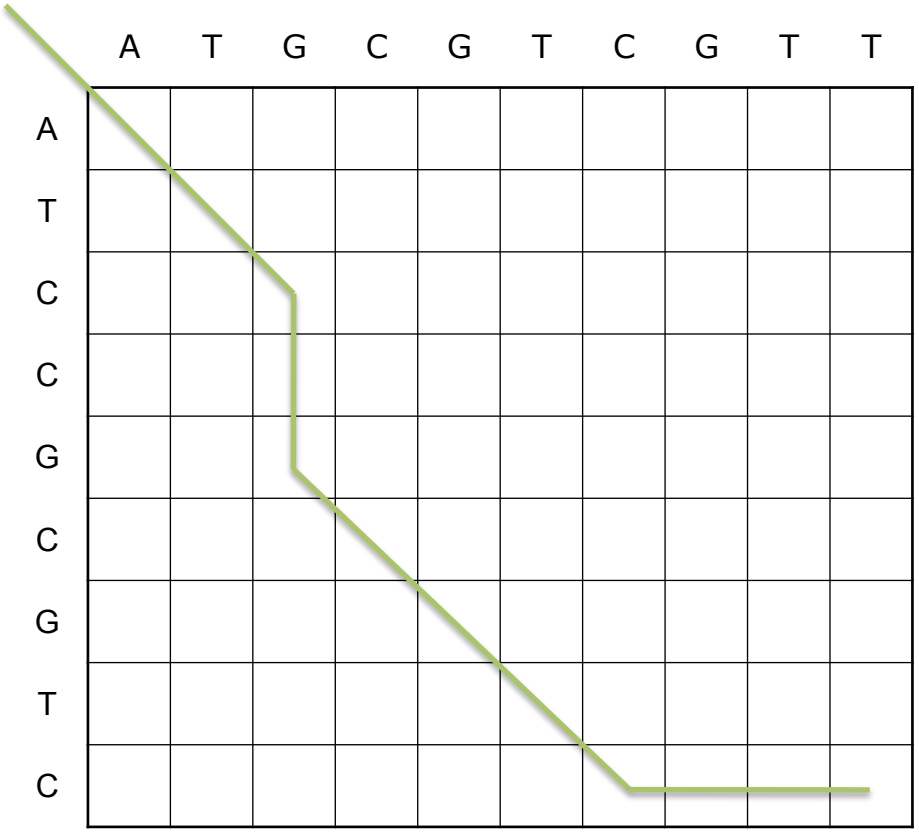


Alignement global de 2 séquences à partir d'un chemin




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

Alignement global de 2 séquences à partir d'un chemin




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

- **Alignement optimal : démarche**
  - Définition d'un modèle de calcul du score d'alignement
    - événements autorisés

    (appariements, substitutions, brèches)
    - Coût des événements autorisés

    (ouverture de brèches, extension de brèche…)
  - Définition d'un algorithme pour le calcul de l'alignement
    - Alignements de 2 séquences ou plus
    - Alignements globaux, locaux

    Algorithme de programmation dynamique
  - Recherche du chemin qui coûte le moins cher, *i.e.* minimum de mutations (ponctuelles ou indels) ou de score maximal.

Avantage : règles objectives et exhaustivité

Inconvénient : perte d'informations biologiques (cas des alignements globaux/locaux)

- **Modèle de calcul d'un score (ou d'un coût)**
  - Matrices de score / matrices de substitution
    - ADN

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

Score à maximiser

	A	T	C	G
A	0	2	2	1
T	2	0	1	2
C	2	1	0	2
G	1	2	2	0

Coût à minimiser

  - Protéines : PAM (score), BLOSUM (score) …
  - Contribution des indels

$\omega_k = \alpha + k \times \beta$

α = pénalité d'ouverture d'une brèche

β = pénalité d'extension d'une brèche

k = extension de la brèche

Inconvénient: pas vraiment de modèle biologique

- **Algorithme d'alignement de 2 séquences**
  - Programmation dynamique
    - Adapté aux problèmes d'optimisation
    - Résout un problème en combinant les solutions de sous-problèmes non indépendants (sous problèmes communs)
    - Principe : recenser, stocker des sous-problèmes pour éviter de les recalculer à chaque fois
  - Alignements par programmation dynamique
    - But : Enumérer et calculer le score de **tous les chemins possibles** puis « extraire » le(s) chemin(s) présentant un score optimal
    - Ajout d'une paire de résidus à un alignement
    - Calcul cumulatif du score par ajout du chemin entre les cases dans la matrice
    - Alignement optimal obtenu par cheminement vers l'arrière

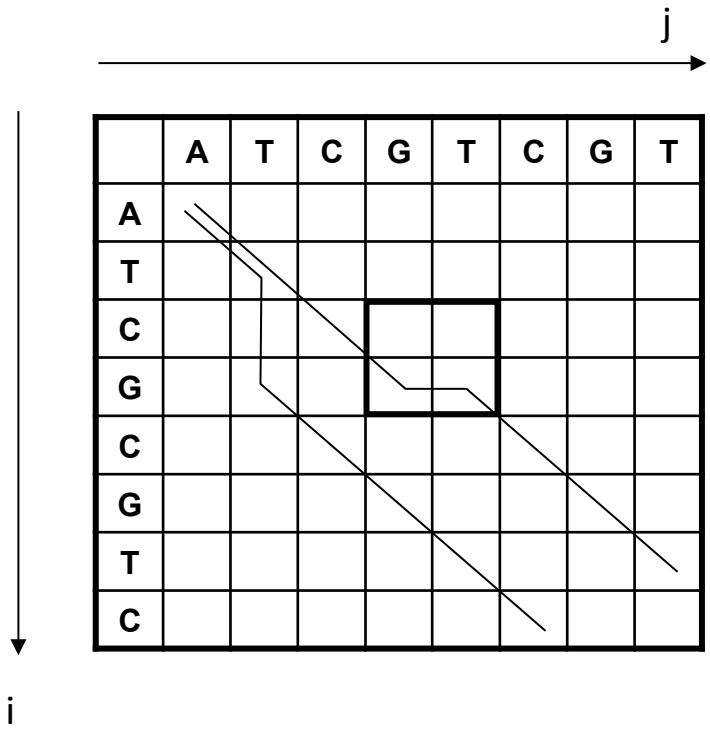
- **Alignement global: algorithme de Needleman et Wunsch (1970)**

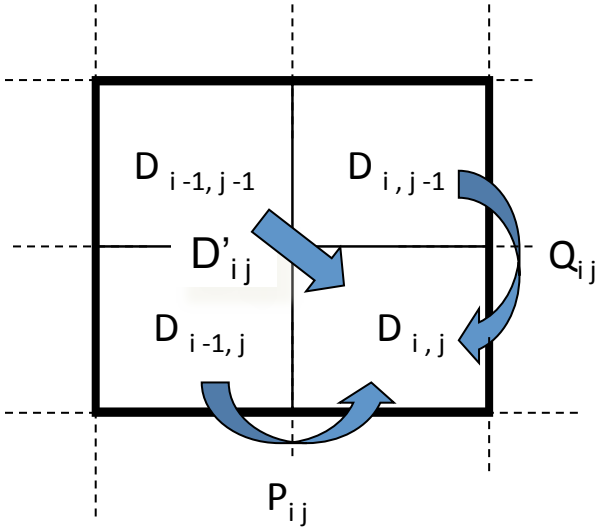
  - On cherche à aligner les séquences dans leur entière longueur

  - exemple de modèle de calcul du score

    - Matrice de substitution $\begin{cases} \gamma_{x,y} = 0 \text{ si } X = Y \\ \gamma_{x,y} = 3 \text{ si } X \neq Y \end{cases}$
    - Contribution aux brèches $\omega_k = 4$
    - Condition initiales : $D_{00} = P_{00} = Q_{00} = 0$

•Alignement global: algorithme de Needleman et Wunsch (1970)





$$D_{i\,j} = \text{Min}\,(D'_{i\,j}, P_{i\,j}, Q_{i\,j})$$

$$D'_{i\,j} = D_{i-1,j-1} + \gamma\,(a_i, b_j)$$

$$Q_{i\,j} = D_{i,j-1} + \omega_k$$

$$P_{i\,j} = D_{i-1,j} + \omega_k$$

Score (ou coût) d'une substitution de a_i en b_j

- Alignement global: algorithme de Needleman et Wunsch (1970)

		0 D' P Q -	1 D' P Q A	2 D' P Q T	3 D' P Q C	4 D' P Q C	5 D' P Q G	6 D' P Q C
0	D' P Q	-						
1	D' P Q	A						
2	D' P Q	T						
3	D' P Q	G						
4	D' P Q	C						

- **Alignement global: algorithme de Needleman et Wunsch (1970)**

	0 D' P Q -	1 D' P Q A	2 D' P Q T	3 D' P Q C	4 D' P Q C	5 D' P Q G	6 D' P Q C
0 D' P - Q	**0** 0 0	**4** 4 4	**8** 8 8	**12** 12 12	**16** 16 16	**20** 20 20	**24** 24 24
1 D' P A Q	**4** 4 4						
2 D' P T Q	**8** 8 8						
3 D' P G Q	**12** 12 12						
4 D' P C Q	**16** 16 16						

- **Alignement global: algorithme de Needleman et Wunsch (1970)**

	0 D' P Q -	1 D' P Q A	2 D' P Q T	3 D' P Q C	4 D' P Q C	5 D' P Q G	6 D' P Q C
0 D' P - Q	**0** 0 0	**4** 4 4	**8** 8 8	**12** 12 12	**16** 16 16	**20** 20 20	**24** 24 24
1 D' P A Q	**4** 4 4	**0** 8 8					
2 D' P T Q	**8** 8 8						
3 D' P G Q	**12** 12 12						
4 D' P C Q	**16** 16 16						

- **Alignement global: algorithme de Needleman et Wunsch (1970)**

		0 D' P Q -	1 D' P Q A	2 D' P Q T	3 D' P Q C	4 D' P Q C	5 D' P Q G	6 D' P Q C
0	D' P - Q	**0** 0 0	**4** 4 4	**8** 8 8	**12** 12 12	**16** 16 16	**20** 20 20	**24** 24 24
1	D' P A Q	**4** 4 4	**0** 8 8	7 **4** 12	11 **8** 16	15 **12** 20	19 **16** 24	23 **20** 28
2	D' P T Q	**8** 8 8						
3	D' P G Q	**12** 12 12						
4	D' P C Q	**16** 16 16						

- **Alignement global: algorithme de Needleman et Wunsch (1970)**

	0 D' P Q -	1 D' P Q A	2 D' P Q T	3 D' P Q C	4 D' P Q C	5 D' P Q G	6 D' P Q C
0 D' P - Q	**0** 0 0	**4** 4 4	**8** 8 8	**12** 12 12	**16** 16 16	**20** 20 20	**24** 24 24
1 D' P A Q	**4** 4 4	**0** 8 8	7 **4** 12	11 **8** 16	15 **12** 20	19 **16** 24	23 **20** 28
2 D' P T Q	**8** 8 8	7 12 **4**					
3 D' P G Q	**12** 12 12	11 16 **8**					
4 D' P C Q	**16** 16 16	15 20 **12**					

- **Alignement global: algorithme de Needleman et Wunsch (1970)**

		0 D' P Q -	1 D' P Q A	2 D' P Q T	3 D' P Q C	4 D' P Q C	5 D' P Q G	6 D' P Q C
0	D' P - Q	**0** 0 0	**4** 4 4	**8** 8 8	**12** 12 12	**16** 16 16	**20** 20 20	**24** 24 24
1	D' P A Q	**4** 4 4	**0** 8 8	7 **4** 12	11 **8** 16	15 **12** 20	19 **16** 24	23 **20** 28
2	D' P T Q	**8** 8 8	7 12 **4**	**0** 8 8	7 **4** 12	11 **8** 16	15 **12** 20	19 **16** 24
3	D' P G Q	**12** 12 12	11 16 **8**	7 12 **4**	**3** 8 8	**7** **7** 12	**8** 11 16	15 **12** 20
4	D' P C Q	**16** 16 16	15 20 **12**	11 16 **8**	**4** 12 7	**3** 8 11	10 **7** 12	**8** 11 16

- **Alignement global: algorithme de Needleman et Wunsch (1970)**

		0 D' P Q -	1 D' P Q A	2 D' P Q T	3 D' P Q C	4 D' P Q C	5 D' P Q G	6 D' P Q C
0	D' P - Q	**0** 0 0	**4** 4 4	**8** 8 8	**12** 12 12	**16** 16 16	**20** 20 20	**24** 24 24
1	D' P A Q	**4** 4 4	0 8 8	7 4 12	11 8 16	15 12 20	19 16 24	23 20 28
2	D' P T Q	**8** 8 8	7 12 4	0 8 8	7 4 12	11 8 16	15 12 20	19 16 24
3	D' P G Q	**12** 12 12	11 16 8	7 12 4	3 8 8	7 7 12	8 11 16	15 12 20
4	D' P C Q	**16** 16 16	15 20 12	11 16 8	4 12 7	3 8 11	10 7 12	8 11 16

- **Alignement global: algorithme de Needleman et Wunsch (1970)**

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

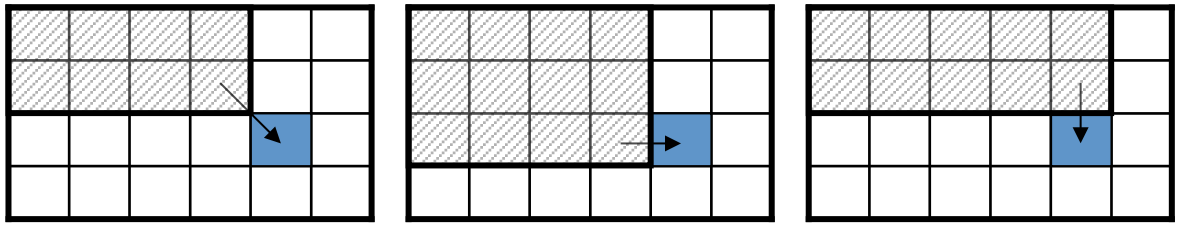
	0 D' P Q -	1 D' P Q A	2 D' P Q T	3 D' P Q C	4 D' P Q C	5 D' P Q G	6 D' P Q C
D' 0 P - Q	**0** 0 0	**4** 4 4	**8** 8 8	**12** 12 12	**16** 16 16	**20** 20 20	**24** 24 24
D' 1 P A Q	**4** 4 4	**0** 8 8	7 **4** 12	11 **8** 16	15 **12** 20	19 **16** 24	23 **20** 28
D' 2 P T Q	**8** 8 8	7 12 **4**	**0** 8 8	7 **4** 12	11 **8** 16	15 **12** 20	19 **16** 24
D' 3 P G Q	**12** 12 12	11 16 **8**	7 12 **4**	**3** 8 8	**7** **7** 12	**8** 11 16	15 **12** 20
D' 4 P C Q	**16** 16 16	15 20 **12**	11 16 **8**	**4** 12 7	**3** 8 11	10 **7** 12	**8** 11 16

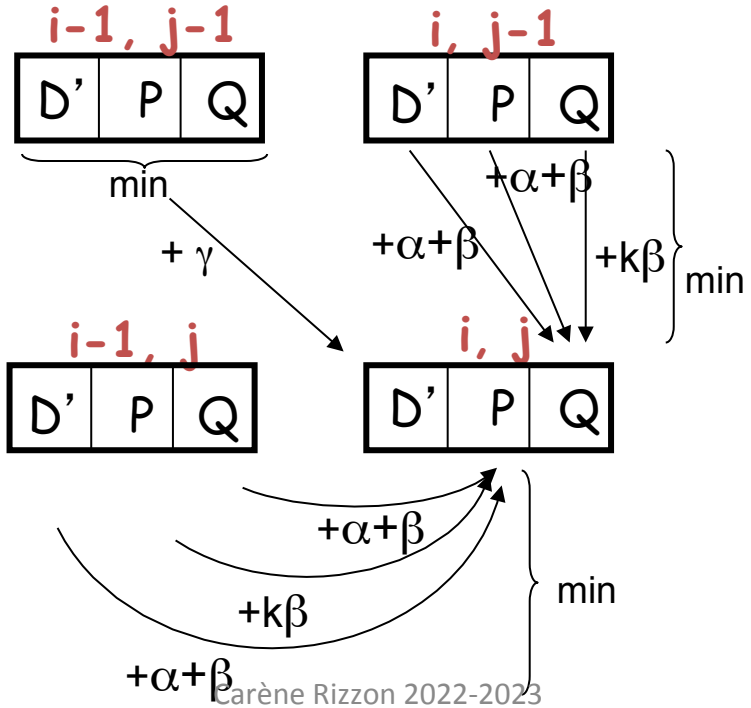
Remarque: ici la méthode est simplifiée par le fait que k=0 dans $\omega_k = \alpha + k \times \beta$

- **Alignement global: algorithme de Needleman et Wunsch (1970)**

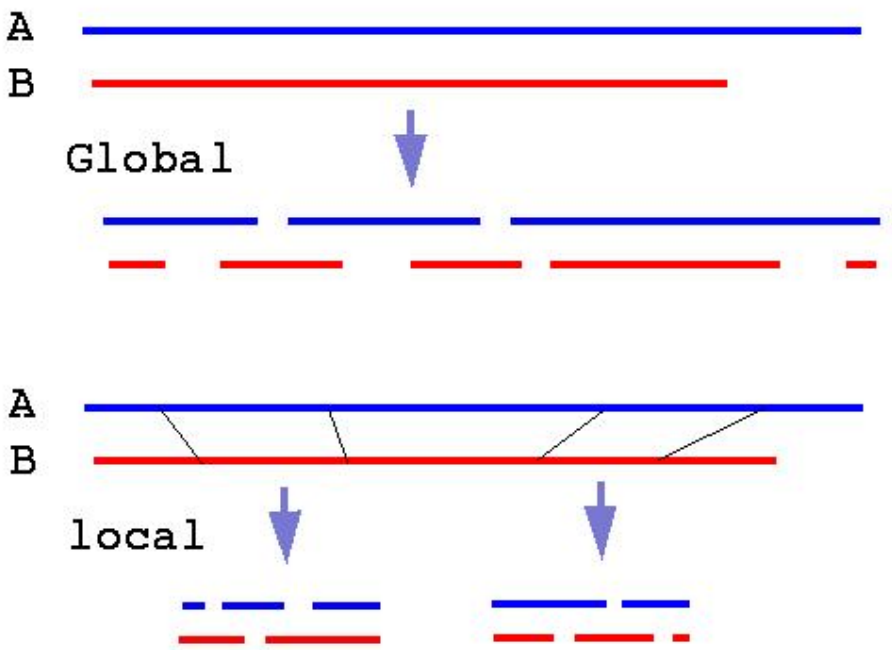
Prise en compte de la fonction affine pour la pénalité des brèches:

$\omega_k = \alpha + k \times \beta$





- **Alignement local: algorithme de Smith et Waterman (1981)**



- Alignement d' une séquence courte sur une séquence plus longue
- Alignement d' un CDS sur une séquence génomique
- Recherche de motifs conservés entre séquences divergentes
- Recherche d' une séquence dans une banque

- **Alignement local: algorithme de Smith et Waterman (1981)**

  Principe : Etant donné 2 séquences S et T et 2 indices i, j, trouver un suffixe (sous-séquence) s de S1..i et un suffixe (sous-séquence) t de T1..j telles que la valeur de leur alignement soit optimale sur l'ensemble des alignements des suffixes (sous-séquences) de S1..i et T1..j

- **Alignement local: algorithme de Smith et Waterman (1981)**

- Algorithme :
  - Équivalent à l'algorithme de Needleman et Wunsch (1970)
  - Sur scores de similarité uniquement
  - Restrictions sur les valeurs relatives des poids

$\gamma(x, y) \geq 0$ si $x = y$ $\gamma(x, y) \leq 0$ si $x \neq y$ $\omega k \leq 0$

- Equation de récurrence
  - $\sigma(i, j)$ = score de l'alignement optimal du suffixe local pour une paire d'indices i,j
  - Pour tout i, j, $\sigma(i, 0) = 0$ $\sigma(0, j) = 0$
  - Relation de récurrence

$$\sigma(i, j) = \text{Max} \begin{cases} 0 \\ \sigma(i-1, j-1) + \gamma(S[i], T[j]) \\ \sigma(i-1, j) + \omega \\ \sigma(i, j-1) + \omega \end{cases}$$

  - Calculer i* et j* tel que $\sigma(i^*, j^*) = \max \sigma(i, j)$

Alignement de 2 séquences

$$\begin{cases} \gamma(x, y) = 1 \text{ si } X = Y \\ \gamma(x, y) = -0{,}5 \text{ si } X \neq Y \\ \omega = -1{,}5 \end{cases}$$

		0 D' P Q -	1 D' P Q A	2 D' P Q T	3 D' P Q G	4 D' P Q C	5 D' P Q G	6 D' P Q T
0	D' P - Q	**0** 0 0	**0** 0 0	**0** 0 0	**0** 0 0	**0** 0 0	**0** 0 0	**0** 0 0
1	D' P A Q	**0** 0 0	1 -1,5 -1,5	-0,5 -0,5 -1,5	-0,5 -1,5 -1,5	-0,5 -1,5 -1,5	-0,5 -1,5 -1,5	-0,5 -1,5 -1,5
2	D' P G Q	**0** 0 0	-0,5 -1,5 -0,5	0,5 -1,5 -1	1 -1 -1,5	-0,5 -0,5 -1,5	1 -1,5 -1,5	-0,5 -0,5 -1,5
3	D' P C Q	**0** 0 0	-0,5 -1,5 -1,5	-0,5 -1,5 -1	0 -1,5 0	2 -1,5 -1,5	-0,5 0,5 -0,5	0,5 -1,5 -1,5
4	D' P C Q	**0** 0 0	-0,5 -1,5 -1,5	-0,5 -1,5 -1,5	-0,5 -1,5 0	1 -1,5 0,5	1,5 -0,5 -1,5	-0,5 0 -1

Alignement de 2 séquences

$$\begin{cases} \gamma\,(x, y) = 1 \text{ si } X = Y \\ \gamma\,(x, y) = -0{,}5 \text{ si } X \neq Y \\ \omega = -1{,}5 \end{cases}$$

Seq1 A T G C G T C G T T Seq2 A G C C C G T C

		A	T	G	C	G	T	C	G	T	T
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	0	0	0	0	0	0	0
G	0	0	0.5	1	0	1	0	0	1	0	0
C	0	0	0	0	2	0.5	0.5	1	0	0.5	0
C	0	0	0	0	1	1.5	0	1.5	0.5	0	0
C	0	0	0	0	1	0.5	1	1	1	0	0
G	0	0	0	1	0	2	0.5	0.5	2	0.5	0
T	0	0	1	0	0.5	0.5	3	1.5	0.5	3	1.5
C	0	0	0	0	1	0	1.5	4	2.5	1.5	2.5

Alignement de 2 séquences

$$\begin{cases} \gamma(x, y) = 1 \text{ si } X = Y \\ \gamma(x, y) = -0{,}5 \text{ si } X \neq Y \\ \omega = -1{,}5 \end{cases}$$

Seq1 A T G C G T C G T T Seq2 A G C C C G T C

		A	T	G	C	G	T	C	G	T	T
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	0	0	0	0	0	0	0
G	0	0	0.5	1	0	1	0	0	1	0	0
C	0	0	0	0	2	0.5	0.5	1	0	0.5	0
C	0	0	0	0	1	1.5	0	1.5	0.5	0	0
C	0	0	0	0	1	0.5	1	1	1	0	0
G	0	0	0	1	0	2	0.5	0.5	2	0.5	0
T	0	0	1	0	0.5	0.5	3	1.5	0.5	3	1.5
C	0	0	0	0	1	0	1.5	4	2.5	1.5	2.5

Alignement de 2 séquences

```
Seq1     A T G C G T C G T T
                 : : : :
Seq2   A G C C C G T C
```

```
Seq1   A T G C G T C G T T
                   : : :
Seq2       A G C C C G T C
```

		A	T	G	C	G	T	C	G	T	T
	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	0	0	0	0	0	0	0
G	0	0	0.5	1	0	1	0	0	1	0	0
C	0	0	0	0	2	0.5	0.5	1	0	0.5	0
C	0	0	0	0	1	1.5	0	1.5	0.5	0	0
C	0	0	0	0	1	0.5	1	1	1	0	0
G	0	0	0	1	0	2	0.5	0.5	2	0.5	0
T	0	0	1	0	0.5	0.5	3	1.5	0.5	3	1.5
C	0	0	0	0	1	0	1.5	4	2.5	1.5	2.5

About EMBOSS

Contents

Overview

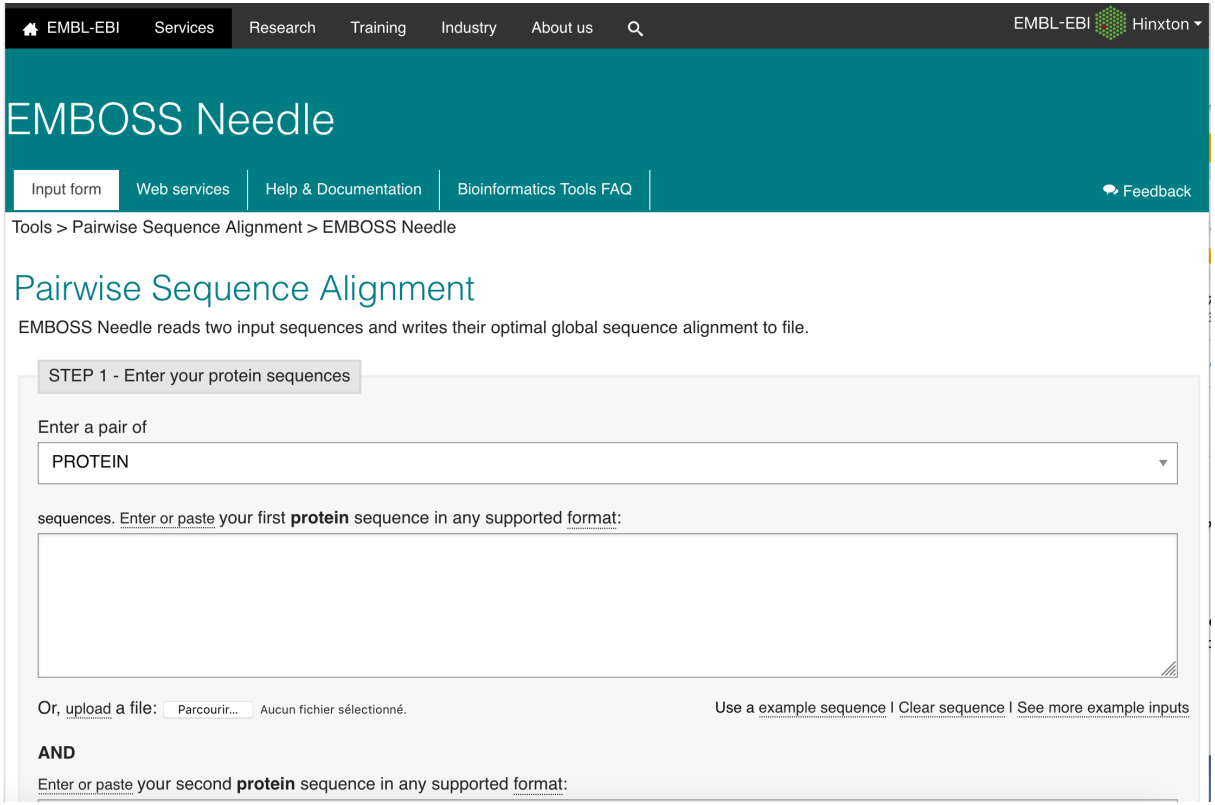
EMBOSS is "The European Molecular Biology Open Software Suite". EMBOSS is a free Open Source software analysis package specially developed for the needs of the molecular biology (e.g. EMBnet) user community. The software automatically copes with data in a variety of formats and even allows transparent retrieval of sequence data from the web. Also, as extensive libraries are provided with the package, it is a platform to allow other scientists to develop and release software in true open source spirit. EMBOSS also integrates a range of currently available packages and tools for sequence analysis into a seamless whole. EMBOSS breaks the historical trend towards commercial software packages.

EMBOSS key features

There have been tens of thousands of unique downloads in the short time it has been available including site-wide installations by administrators catering for hundreds or even thousands of users.

The uses and interfaces to EMBOSS have long grown beyond our ability to keep track of them. EMBOSS is used extensively in production environments rather than being the sort of "research project" code that gets presented at conferences, but never actually deployed.

EMBOSS has several important advantages:



+ en ligne de commande sous linux