BIO-INFORMATIQUE

Introduction à la programmation pour les biologistes

ENSEIGNANT

- M.Djoudi Brahim
- Docteur en Informatique
- Faculté des science de la nature et de la vie
- Département de Biochimie et Biologie cellulaire et moléculaire
- Email (meilleure façon de me contacter): DjoudiBrahim@hotmail.fr

OBJECTIFS

- Ce cours permet de vous familiariser avec les applications de la bio-informatique pour :
  - **Traitement ultérieur des données**,
  - **Emettre de nouvelles hypothèse**s et
  - **Générer de nouvelles données**.

  Il est nécessaire que vous « carrosser » ces méthodes **sous forme** de **logiciels ou serveurs Web avec ses interfaces graphiques** conviviales surtout en 3eme année afin de vous bénéficier de développement rapide et l'utilisation facile de la bio-informatique.

OBJECTIFS

- Comprendre le vocabulaire biologique et bio-informatique
- Comprendre les questions moléculaires et les techniques bio informatiques qui ont résolus ces questions.
- Être capable d'utiliser et de choisir entre des algorithmes connus qui peut résoudre une certaine question biomoléculaire.
- Être capable d'utiliser certaines techniques pour résoudre un problème moléculaire
- Augmentez votre intérêt pour des autres sciences et la recherche interdisciplinaire

PRÉREQUIS

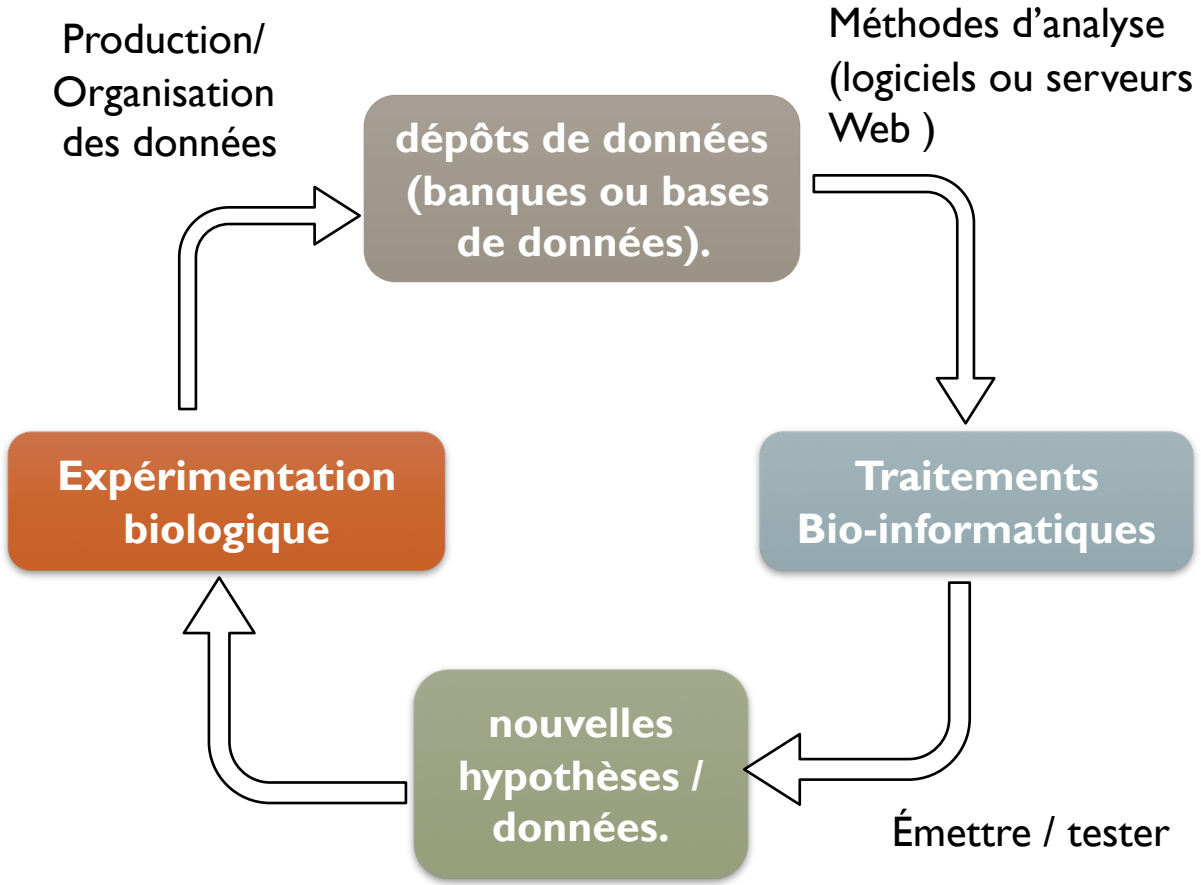
- Pour réussir ce module il faut au préalable :
- Avoir une bonne maîtrise des outils informatique (navigateurs web...etc.) et la recherche sur internet,
- Savoir les notions de base relatives aux Acides nucléique et protéique.

RESPONSABILITÉS DE L'ÉTUDIANT

- Les élèves doivent lire / visionner le matériel assigné avant la classe pour laquelle ils sont programmés, assister aux cours, participer en classe, remplir les devoirs et demander de l'aide tôt s'ils ont des problèmes.

BIO-INFORMATIQUE

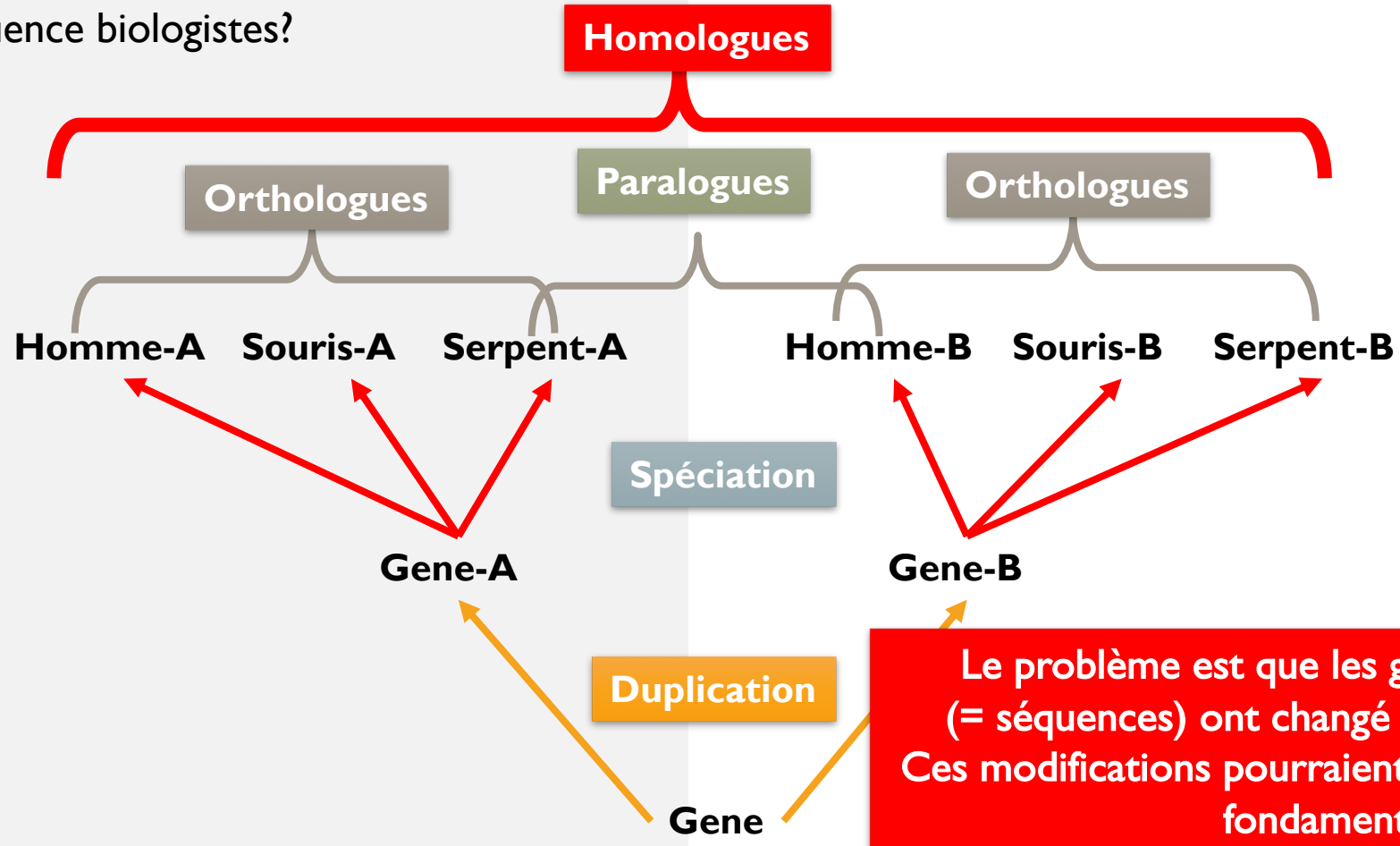
La bio-informatique est une « inter-discipline » à la frontière de la biologie, de l'informatique et des mathématiques qui permet d'analyser les informations biologiques **énorme** contenues dans les cellules vivantes sous forme de séquences nucléiques ou protéiques



COMPARER EN BIOLOGIE

- **Ce sont les mutations qui, au cours de l'évolution naturelle, causent des modification au moment de la réplication de l'ADN ;**

Evolution des séquence biologistes?



Le problème est que les gènes ou protéines (= séquences) ont changé pendant l'évolution
Ces modifications pourraient cacher une similarité fondamentale

POURQUOI COMPARER EN BIOLOGIE ?

- Trouver des **ressemblances** ou des **différences significatives** au niveau de **Fonctions**, et de Structures des séquences,

S'il y a similitude, cela signifie qu'il est possible **que les deux séquences présentent la même fonction biologique**, ou du moins les deux séquences présente une structure fortement similaire,

En pratique, l'homologie entre deux protéines est inférée avec confiance lorsque le pourcentage d'identité entre leurs deux séquences est supérieur à 75 % et que l'alignement couvre 70 % des deux séquences.

Identité vs. similarité ?

- Identité : Résidus identiques
- Similarité : Deux résidus sont similaires si la substitution de l'un par l'autre n'a aucun / peu d'effet sur la fonctionnalité
- Homologie: Deux protéines sont homologues si et seulement si elles résultent de l'évolution à partir d'un ancêtre commun. Une homologie peut indiquer une structure ou fonction similaire, Pourcentage d'identité > 30 %,

POURQUOI COMPARER EN BIOLOGIE ?

- Trouver des **ressemblances** ou des **différences significatives** au niveau de **Fonctions**, et de Structures des séquences,

S'il y a similitude, cela signifie qu'il est possible **que les deux séquences présentent la même fonction biologique**, ou du moins les deux séquences présente une structure fortement similaire,

En pratique, l'homologie entre deux protéines est inférée avec confiance lorsque le pourcentage d'identité entre leurs deux séquences est supérieur à 75 % et que l'alignement couvre 70 % des deux séquences.

Voila un exemple de comparaison

```
THISSEQUENCE
||   ||||||||
THATSEQUENCE
```

PB : si nous avons des tailles différents

```
THISISASEQUENCE
    THATSEQUENCE
```

Solution : Chercher **l'alignement** qui produit la similarité maximale

```
THISISASEQUENCE
|||||||||||||||
THATSEQUENCE---
```

```
THIS--A-SEQUENCE
||    | ||||||||
TH----ATSEQUENCE
```

ALIGNEMENT DE SÉQUENCES

- Alignement = Comparaison

- Ecrire une séquence au dessous de l'autre et les faire bouger pour faire apparaitre :

  ✓ Le max d'Identités
  ✓ Substitutions

- **Exemple :**

  Soit les deux séquences suivantes :

  Sequence 1 :A G V S I L N Y A
  Sequence 2 :V S I L Y A K R

1	2	3	4	5	6	7	8	9
A	G	V	S	I	L	N	Y	A
V	S	I	G	Y	A	K	R	

- Si on admet que la séquence 2 a perdu les deux acides aminés N terminaux au cours de l'évolution, l'alignement devient :

1	2	3	4	5	6	7	8	9	10
A	G	V	S	I	L	N	Y	A	----
---	----	V	S	I	G	Y	A	K	R

ALIGNEMENT DE SÉQUENCES

- En pratique, l'évolution a pu faire disparaître (**délétion**) ou apparaître (**insertion**) des acides aminés à l'intérieur des séquences. Une délétion dans une séquence correspond à une insertion dans l'autre.

- On parle alors d'« **indel** ».

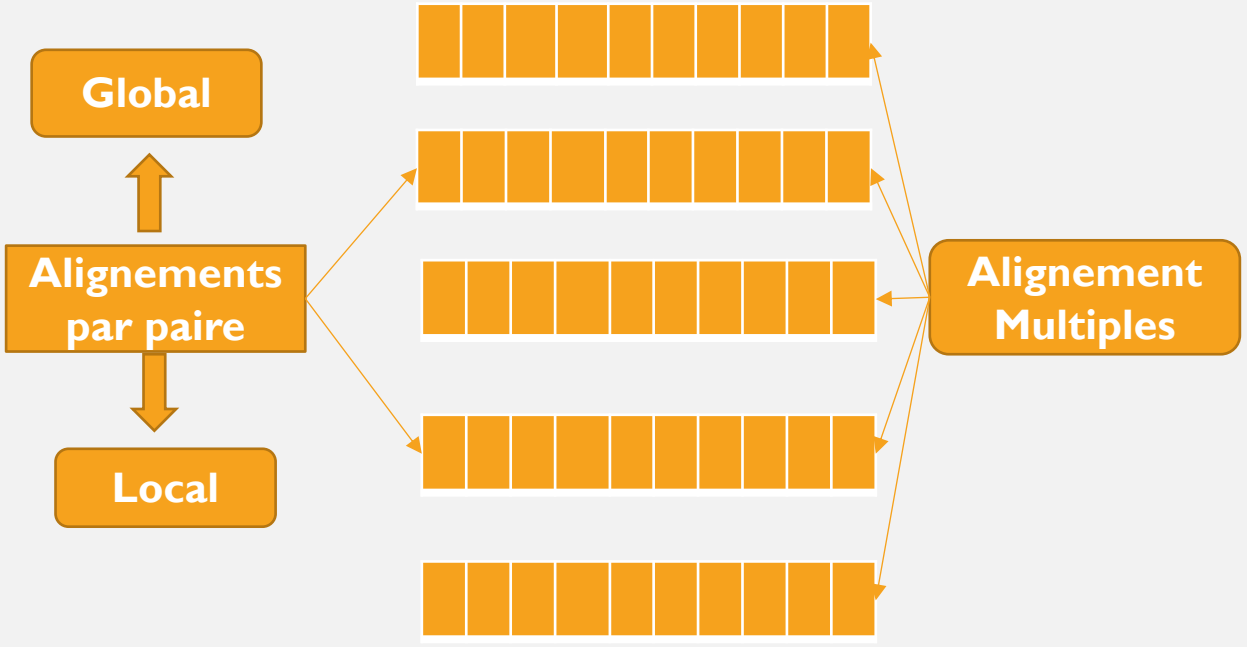
- Exemple :

1	2	3	4	5	6	7	8	9	10
A	G	V	S	I	L	N	Y	A	----
---	----	V	S	I	G	Y	A	K	R

Indel : est un mot-clé inventé par le mathématicien Joseph Kruskal et utilisé en génétique et en bio-informatique pour désigner une insertion ou une délétion dans une séquence biologique (acide nucléique ou protéine) par rapport à une séquence de référence.

ALIGNEMENT DE SÉQUENCES

- L'alignement de séquences concerne au minimum deux séquences.




```
metL          16 KFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTNQLINWLK-LS     64
                 ||||:|:||.....|.|.|:...:.. .::|:||:...||.|:...: |.
lysC           8 KFGGTSVADFDAMNRSADIVLSDANV-RLVVLSASAGITNLLVALAEGLE     56
```

Alignement par paire

```
                    10        20        30        40        50        60        70
                     |         |         |         |         |         |         |
 1 GNLSPA1A3  MAELTIDPTTIRKALDEFVESYKPSDTPTQEVGYVATAGDGIAHVTGLPGCMANELLTFEDG---TLGLAFNLDA
 2 GNLSPA0LS  MSELTIRPEEIRAALEEFVSSYTPDVASREEVGRVTEAGDGIARIEGLPSTMANELLRFEDG---TLGLALNLDV
 3 GNLSPA0R2  MAELTISAADIEGAIEDYVSSFSAD-TEREEIGTVIDAGDGIAHVEGLPSVMTQELLEFPGG---VLGVALNLDE
 4 GNLSPA0PU  MAELTISANDIQSAIEEYVGSFTSD-TSREEVGTVVDAGDGIAHVEGLPSVMTQELLEFPGG---VLGVALNLDE
 5 GNLSPA0QC  MAELTISADDIQSAIEEYVGSFTSD-TSREEVGTVVDAGDGIAHVEGLPSVMTQELLEFPGG---VLGVALNLDE
 6 GNLSPA1E9  --MATLRVDEINKILRERIEQYNRK-VGIENIGRVVQVGDGIARIIGLGEIMSGELVEFAEG---TRGIALNLES
 7 GNLSPA1EA  --MATLRVDEIHKILRERIEQYNRK-VGIENIGRVVQVGDGIARIIGLGEIMSGELVEFAEG---TRGIALNLES
 8 GNLSPA0A3  --MVTIRADEISNIIRERIEQYNRE-VKIVNTGTVLQVGDGIARIHGLDEVMAGELVEFEEG---TIGIALNLES
 9 GNLSPA0ZZ  --MVTIRADEISNIIRERIEQYNRE-VKIVNTGTVLQVGDGIARIHGLDEVMAGELVEFEEG---TIGIALNLES
10 GNLSPA0T0  --MINIRPDEISSIIREQIEKYDQD-VKVDNIGTVLQVGDGIARVYGLDQVMSGELLEFEDK---TIGIALNLEN
11 GNLSPA0T0  --MINIRPDEISSIIREQIEQYDQD-VKIDNIGTVLQVGDGIARVYGLDQVMSGELLEFEDK---TIGIALNLEN
12 GNLSPA1AX  ---MQLNAHEISDLIKKQIEGFDFD-AEVRTEGSVVSVSDGIVRIHGLADVQFGEMLEFPNN---TFGMALNLEQ
13 GNLSPA0L2  ---MQLNSTEISDLIKQRIEQFEVV-SESRNEGTIVAVSDGIIRIHGLADVMQGEMIELPGS---RFAIALNLER
14 GNLSPA1JT  ---MQLNSTEISELIKQRIAQFNVV-SEAHNEGTIVSVSDGIIRVHGLADVMQGEMIALPGN---RYAIALNLER
15 GNLSPA0Q8  ---MQLSPSEISGLIKQRIEKFDNS-VELKSEGTIVSVADGIVTIYGLNDVAAGEMIKLPGD---VYGLALNLNT
16 GNLSPA0K2  ---MQLNPSEISELIKSRIQGLEAS-ADVRNQGTVISVTDGIVRIHGLSDVMQGEMLEFPGN---TFGLALNLER
17 GNLSPA1K1  ---MQLNPSEISDLIKSRIQNLQLA-ATSRNEGTVVSVTDGITRIHGLTDVMQGEMLEFPGN---TFGLALNLER
18 GNLSPA0LL  ---MEIRAEEISQIIREQIKDYEKQ-VELSETGRVLSVGDGIARVYGVEKCMSMELLEFPTEHGVVYGLALNLEE
19 GNLSPA0Q2  ---MNVKPEEITSIIKKQIESYEHK-IQTVDSGTIIQIGDGIARVYGIEDCMEGELLEFPND---VYGMALNLEQ
20 GNLSPA0RL  ---MSIRAEEISALIKQQIENYQSE-IEVSDVGTVIQVGDGIARAHGLDNVMAGELVEFSNG---VMGLAQNLEE
21 GNLSPA0RR  -MSVKLKADEISSIIKERIENYNLS-VDIEETGKVISVADGVANVYGLKNVMAGEMVEFETG---EKGMALNLEE
22 GNLSPA1BJ  -MSTTVRPDEVSSILRKQLAGFESE-ADVYDVGTVLQVGDGIARVYGLSKAAAGELLEFPNK---VMGMALNLEE
23 GNLSPA0LD  ---MQVSVAEISGILKKQIAEYGKE-AEVSEVGEVIAVGDGIARAYGLDNVMAGEMVEFEDG---TQGMALNLEE
24 GNLSPA1B8  ---MGIQAAEISAILKDQIKNFGQD-AEVAEVGQVLSVGDGIARVYGLDKVQAGEMVEFPGG---IRGMVLNLET
```

Alignement Multiples

MATRICE DE SCORE

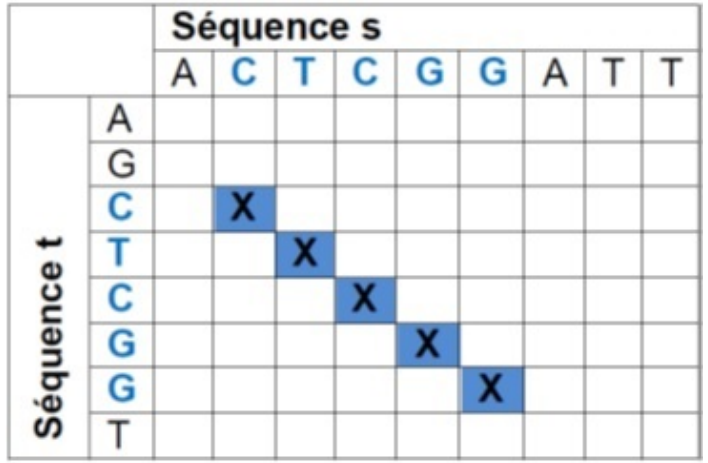
- les trois peptides suivants ne présentent pas d'identité entre eux.

Peptide 1 : A E I G L M A E I G L S E K I L
Peptide 2 : L D V A A I G D L A I T Q R L M
Peptide 3: W R G I Y S H H D E T W D C P C

- **N'importe quel biochimiste ou phylogénéticien est capable de dire que les peptides 1 et 2 se ressemblent plus entre eux que les peptides 1 et 3.**

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

Visibles que les résidus identiques ou les segments identiques.



Matrices nucléiques

MATRICE DE SUBSTITUTION

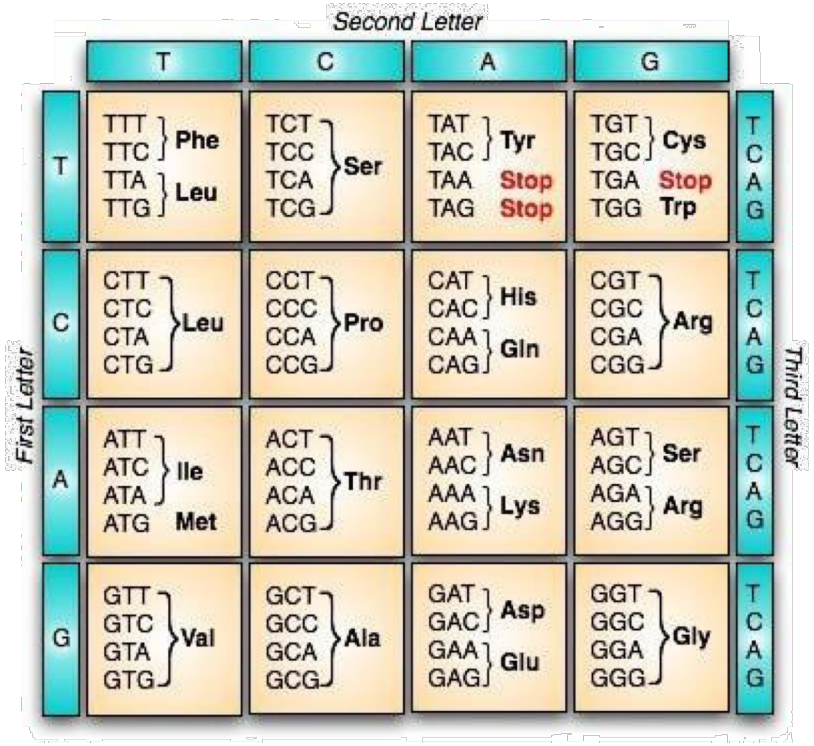
Besoin d'un systèmes de score qui prend on compte :

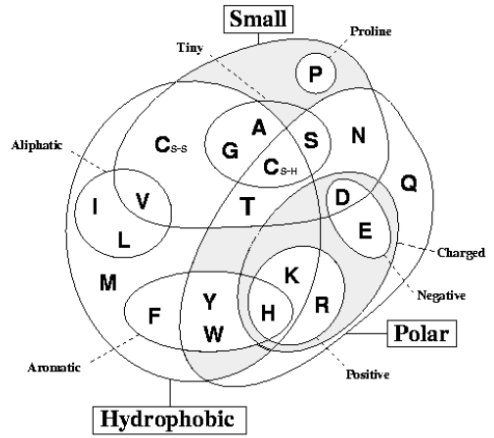
- Redondance
- Certains mutations/changements/ substitutions acceptables

Exemple:

Phénylalanine : TTT = TTC
Leucine : CTT = CTC = CTA
Histidine : CAC = CAA : Glutamine

le biologiste souhaite comparer avec **non pas des identités** mais avec **des similitudes** et pouvoir **quantifier la ressemblance** entre des peptides qui ne présentant pas d'identités.





MATRICE DE SUBSTITUTION

Besoin d'un systèmes de score qui :

- Modélise le changement des séquences par rapport au temps d'évolution.
- Favorise les acides aminés identiques ou apparentés
- Pénalise des acides aminés mal appariés ou des gap

Matrice De Substitution

une matrice de substitution permet, pour chaque acide aminé, de connaître sa capacité à être substitué par chaque autre acide aminé, y compris lui-même.

Les deux types de matrice utilisent des scores basés sur la comparaison entre la fréquence observée des substitutions et leur fréquence attendue

PAM

	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

BLOSUM

		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
A	Ala	4																			
R	Arg	-1	5																		
N	Asn	-2	0	6																	
D	Asp	-2	-2	1	6																
C	Cys	0	-3	-3	-3	9															
Q	Gln	-1	1	0	0	-3	5														
E	Glu	-1	0	0	2	-4	2	5													
G	Gly	0	-2	0	-1	-3	-2	-2	6												
H	His	-2	0	1	-1	-3	0	0	-2	8											
I	Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

"POINT ACCEPTED MUTATION"

- Elles ont été créées par **Magaret Dayhoff** et ses collaborateurs, après l'alignement d'environ 1300 séquences **phylogénétiquement proches** très semblables (> 85% d'identité) appartenant à 71 familles de protéines.

- Ce type de matrice donne la probabilité que, suite à une mutation par substitution au cours de l'évolution, n'importe quel acide aminé remplace n'importe quel autre acide aminé sans que la fonction de la protéine ne soit altérée, d'où la terminologie "mutation acceptée".

```
metL         16 KFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTNQLINWLK-LS     64
                ||||:|:||.....|.|.|:...:.. .::|:||:...||.|:...: |.
lysC          8 KFGGTSVADFDAMNRSADIVLSDANV-RLVVLSASAGITNLLVALAEGLE     56

metL         65 QTDRLSAHQVQQTLRRYQCDLISGL----LPAEEADSLISAFVSDLERLA    110
                ..:|.   :....:|..|..:..|    :..||.:.|:.. ::.|...|
lysC         57 PGERF---EKLDAIRNIQFAILERLRYPNVIREEIERLLEN-ITVLAEAA    102

metL        111 ALLDSGINDAVYAEVVGHGEVWSARLMSAVLNQQGLPAAWLDAREFLRA-    159
                ||..|   .|:..|:|.|||:.|..|...:|.::.:.|.|.|.|:.:|.
lysC        103 ALATS---PALTDELVSHGELMSTLLFVEILRERDVQAQWFDVRKVMRTN    149

metL        160 ERAAQPQVDEGLSYPLLQQLLVQHPGKRLVVT-GFISRNNAGETVLLGRN    208
                :|..:.:.|......|....|:...:.||:| |||...|.|.|..|||.
lysC        150 DRFGRAEPDIAALAELAALQLLPRLNEGLVITQGFIGSENKGRTTTLGRG    199

metL        209 GSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKVKDACLLPLLRLDEAS    258
                ||||:|..:.....|||.||:||.|:|:.|||.|..|..:..:...||:
lysC        200 GSDYTAALLAEALHASRVDIWTDVPGIYTTDPRVVSAAKRIDEIAFAEAA    249

metL        259 ELARLAAPVLHARTLQPVSGSEIDLQLRCSYTPDQGSTRI---------E    299
                |:|...|.|||..||.|...|:|.:.:..|..|..|.:         .
lysC        250 EMATFGAKVLHPATLLPAVRSDIPVFVGSSKDPRAGGTLVCNKTENPPLF    299

metL        300 RVLASGTGARIVTSHDDVCLIEFQVPASQDFKLAHKEIDQILKRAQVRPL    349
                |.||.....::|.|     ...:..|:.| ||  |:..||.|
lysC        300 RALALRRNQTLLTLH------SLNMLHSRGF-LA--EVFGILAR------    334
```

Margaret Dayhoff (1978) a mesuré les taux de substitutions entre chaque paire d'acides aminés, dans une collection de 71 alignements de paires de protéines.

MATRICE DE SUBSTITUTION

- Les **valeurs positives** indiquent des substitutions fréquentes ("acceptées"), càd des substitutions observées plus fréquemment que ce à quoi l'on s'attendrait par hasard. les scores élevés correspondent souvent à des acides aminés partageant des propriétés physico-chimiques

- Les **valeurs négatives** indiquent les mutations rares, càd celles qu'on observe moins fréquemment que ce à quoi l'on s'attendrait par hasard. Ce taux inférieur est interprété comme un indice du contre-sélection, suggérant que ces mutations sont généralement défavorables pour la fonction de la protéine.

- La **diagonale** correspond aux identités. **Il** constituée de scores positifs, qui reflètent le taux de conservation des résidus

		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
A	Ala	4																			
R	Arg	-1	5																		
N	Asn	-2	0	6																	
D	Asp	-2	-2	1	6																
C	Cys	0	-3	-3	-3	9															
Q	Gln	-1	1	0	0	-3	5														
E	Glu	-1	0	0	2	-4	2	5													
G	Gly	0	-2	0	-1	-3	-2	-2	6												
H	His	-2	0	1	-1	-3	0	0	-2	8											
I	Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

"POINT ACCEPTED MUTATION"

- Pour prendre en compte le taux de divergence, Margret Dayhoof a calculé une série de matrices de score, reflétant chacune un certain taux de substitutions.

- Les fréquences de substitutions dépendent du degré de divergence entre séquences, car leur nombre augmente avec le temps.

  - PAM001 taux de substitutions entre acides aminés au terme d'un temps évolutif donnant lieu à ~1% de substitutions par position.

  - PAM050 taux de substitutions entre acides aminés au terme d'un temps évolutif donnant lieu à ~50% de substitutions par position.

  - PAM250 idem avec 250% mutations/position (note: une même position peut faire l'objet de plusieurs mutations successives)

- Quand on fait un alignement, on doit choisir l'une des matrices de cette série, en tenant compte du taux de différences entre les deux séquences qu'on veut aligner.

"POINT ACCEPTED MUTATION"

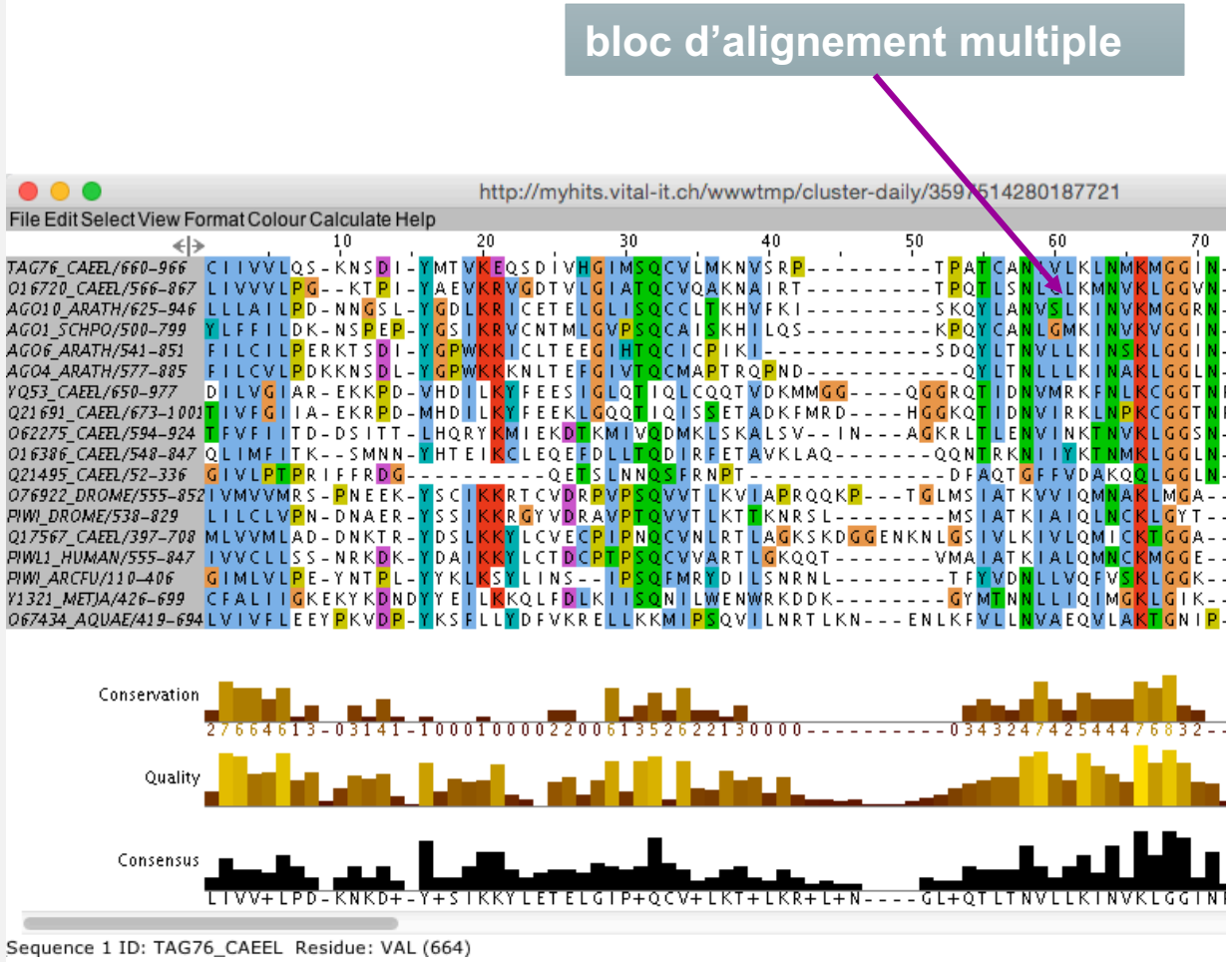
- La PAM 250 est la plus utiliser et donne la probabilité que 250 mutations soit acceptées pour 100 acides aminés.

	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

PAM 250

BLOCKS OF AMINO ACID SUBSTITUTION MATRIX

- S Henikoff and J Henikoff (1992) ont analysé les fréquences de substitutions dans 2000 blocs d'alignements multiples générés à partir d'un grand nombre de familles de protéines (500 familles).
  - Dans les matrices de type BLOSUM, les fréquences sont observées sur des alignements de séquences très divergentes.
  - Néanmoins, dans un tel alignement, les séquences sont moins bien aligner et les "trous" (gaps) sont plus fréquents.
  - Afin d'éviter ces trous, les matrices BLOSUM utilisent des blocs bien alignés et surtout sans trous provenant de la base BLOCKS.



Les BLOCKS sont des régions conservées de familles de protéines ne contenant pas d'insertions ou de délétions.

BLOCKS OF AMINO ACID SUBSTITUTION MATRIX

- Les séquences sont regroupées dans une matrice si leur identité dépasse un certain seuil. Dans une matrice BLOSUM62, les séquences présentant plus de 62% d'identité ont été regroupées ensemble.

- Quand on utilise les matrices BLOSUM pour aligner des séquences, on devrait systématiquement choisir la matrice la plus adéquate, en fonction du pourcentage de similarité.
- Exemples
  - La matrice BLOSUM62 a été calculée sur des blocs de >=62% d'identité
  - La matrice BLOSUM80 a été calculée sur des blocs de >=80% d'identité

		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
A	Ala	4																			
R	Arg	-1	5																		
N	Asn	-2	0	6																	
D	Asp	-2	-2	1	6																
C	Cys	0	-3	-3	-3	9															
Q	Gln	-1	1	0	0	-3	5														
E	Glu	-1	0	0	2	-4	2	5													
G	Gly	0	-2	0	-1	-3	-2	-2	6												
H	His	-2	0	1	-1	-3	0	0	-2	8											
I	Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

BLOSUM 62

RÉSUME

L'alignement, permet de mesurer **la similitude** entre les séquences. S'il y a similitude, cela signifie qu'il est possible que **les deux séquences présentent la même fonction biologique**, ou du moins les deux séquences présente une structure fortement similaire.

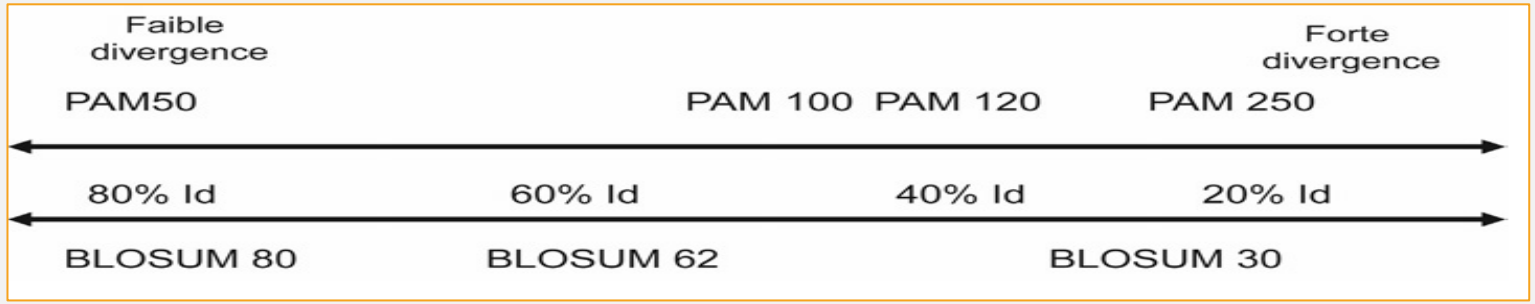
- Différentes matrices de substitution ont été établies
  - PAM (Dayhoff, 1979).
    - PAM signifie «Mutations acceptées en pourcentage»
  - BLOSUM (Henikoff et Henikoff, 1992).
    - BLOSUM signifie «somme forfaitaire».

- Les matrices de substitution permettent de détecter des similitudes entre des protéines plus distantes que ce qui serait détecté avec la simple identité des résidus.

RÉSUME

- La matrice doit être choisie avec soin, en fonction du taux de conservation attendu entre les séquences à aligner.

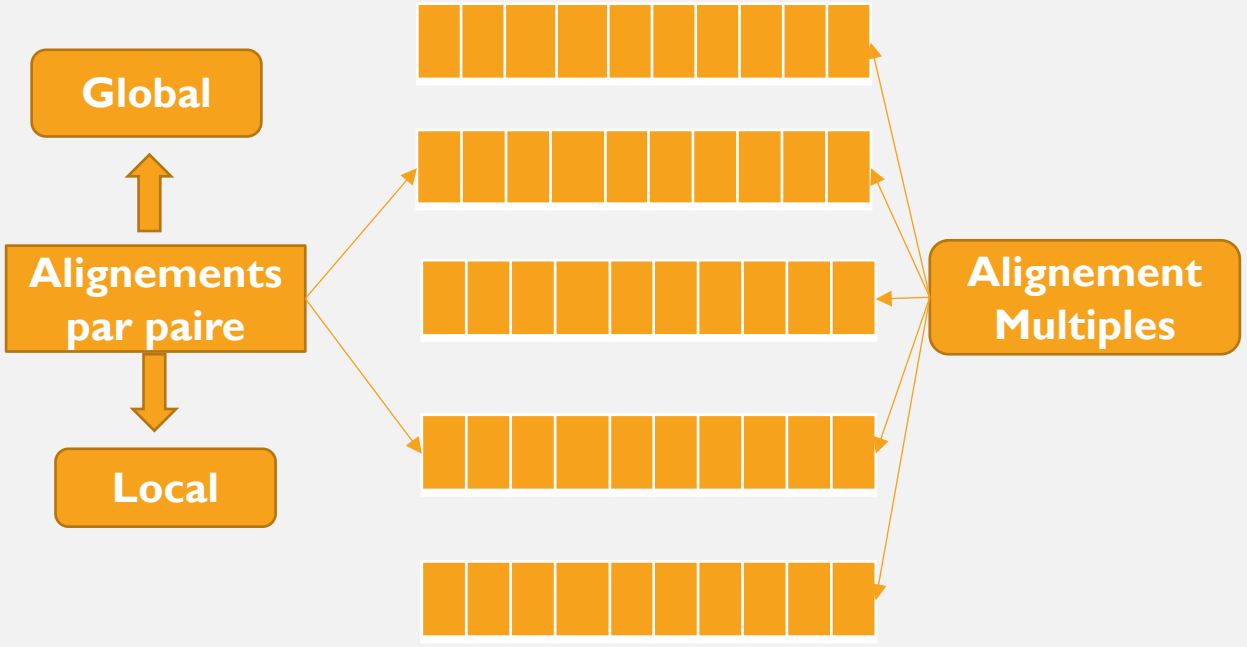
- Avec des matrices PAM
  - le score indique le pourcentage de substitution par position
  - des nombres plus élevés sont appropriés pour des protéines plus éloignées
- Avec les matrices BLOSUM
  - le score indique le pourcentage de conservation?
  - des nombres plus élevés sont appropriés pour des protéines plus conservées



- Le numéro de PAM indique la divergence entre les séquences et le numéro de BLOSUM indique le pourcentage d'identité

ALIGNEMENT DE SÉQUENCES

- L'alignement de séquences concerne au minimum deux séquences.




```
metL          16 KFGGSSLADVKCYLRVAGIMAEYSQPDDMMVVSAAGSTTNQLINWLK-LS     64
                 ||||:|:||.....|.|.|:...:.. .::|:||:...||.|:...: |.
lysC           8 KFGGTSVADFDAMNRSADIVLSDANV-RLVVLSASAGITNLLVALAEGLE     56
```

Alignement par paire

```
                    10        20        30        40        50        60        70
                     |         |         |         |         |         |         |
 1 GNLSPA1A3  MAELTIDPTTIRKALDEFVESYKPSDTPTQEVGYVATAGDGIAHVTGLPGCMANELLTFEDG---TLGLAFNLDA
 2 GNLSPA0LS  MSELTIRPEEIRAALEEFVSSYTPDVASREEVGRVTEAGDGIARIEGLPSTMANELLRFEDG---TLGLALNLDV
 3 GNLSPA0R2  MAELTISAADIEGAIEDYVSSFSAD-TEREEIGTVIDAGDGIAHVEGLPSVMTQELLEFPGG---VLGVALNLDE
 4 GNLSPA0PU  MAELTISANDIQSAIEEYVGSFTSD-TSREEVGTVVDAGDGIAHVEGLPSVMTQELLEFPGG---VLGVALNLDE
 5 GNLSPA0QC  MAELTISADDIQSAIEEYVGSFTSD-TSREEVGTVVDAGDGIAHVEGLPSVMTQELLEFPGG---VLGVALNLDE
 6 GNLSPA1E9  --MATLRVDEINKILRERIEQYNRK-VGIENIGRVVQVGDGIARIIGLGEIMSGELVEFAEG---TRGIALNLES
 7 GNLSPA1EA  --MATLRVDEIHKILRERIEQYNRK-VGIENIGRVVQVGDGIARIIGLGEIMSGELVEFAEG---TRGIALNLES
 8 GNLSPA0A3  --MVTIRADEISNIIRERIEQYNRE-VKIVNTGTVLQVGDGIARIHGLDEVMAGELVEFEEG---TIGIALNLES
 9 GNLSPA0ZZ  --MVTIRADEISNIIRERIEQYNRE-VKIVNTGTVLQVGDGIARIHGLDEVMAGELVEFEEG---TIGIALNLES
10 GNLSPA0T0  --MINIRPDEISSIIREQIEKYDQD-VKVDNIGTVLQVGDGIARVYGLDQVMSGELLEFEDK---TIGIALNLEN
11 GNLSPA0T0  --MINIRPDEISSIIREQIEQYDQD-VKIDNIGTVLQVGDGIARVYGLDQVMSGELLEFEDK---TIGIALNLEN
12 GNLSPA1AX  ---MQLNAHEISDLIKKQIEGFDFD-AEVRTEGSVVSVSDGIVRIHGLADVQFGEMLEFPNN---TFGMALNLEQ
13 GNLSPA0L2  ---MQLNSTEISDLIKQRIEQFEVV-SESRNEGTIVAVSDGIIRIHGLADVMQGEMIELPGS---RFAIALNLER
14 GNLSPA1JT  ---MQLNSTEISELIKQRIAQFNVV-SEAHNEGTIVSVSDGIIRVHGLADVMQGEMIALPGN---RYAIALNLER
15 GNLSPA0Q8  ---MQLSPSEISGLIKQRIEKFDNS-VELKSEGTIVSVADGIVTIYGLNDVAAGEMIKLPGD---VYGLALNLNT
16 GNLSPA0K2  ---MQLNPSEISELIKSRIQGLEAS-ADVRNQGTVISVTDGIVRIHGLSDVMQGEMLEFPGN---TFGLALNLER
17 GNLSPA1K1  ---MQLNPSEISDLIKSRIQNLQLA-ATSRNEGTVVSVTDGITRIHGLTDVMQGEMLEFPGN---TFGLALNLER
18 GNLSPA0LL  ---MEIRAEEISQIIREQIKDYEKQ-VELSETGRVLSVGDGIARVYGVEKCMSMELLEFPTEHGVVYGLALNLEE
19 GNLSPA0Q2  ---MNVKPEEITSIIKKQIESYEHK-IQTVDSGTIIQIGDGIARVYGIEDCMEGELLEFPND---VYGMALNLEQ
20 GNLSPA0RL  ---MSIRAEEISALIKQQIENYQSE-IEVSDVGTVIQVGDGIARAHGLDNVMAGELVEFSNG---VMGLAQNLEE
21 GNLSPA0RR  -MSVKLKADEISSIIKERIENYNLS-VDIEETGKVISVADGVANVYGLKNVMAGEMVEFETG---EKGMALNLEE
22 GNLSPA1BJ  -MSTTVRPDEVSSILRKQLAGFESE-ADVYDVGTVLQVGDGIARVYGLSKAAAGELLEFPNK---VMGMALNLEE
23 GNLSPA0LD  ---MQVSVAEISGILKKQIAEYGKE-AEVSEVGEVIAVGDGIARAYGLDNVMAGEMVEFEDG---TQGMALNLEE
24 GNLSPA1B8  ---MGIQAAEISAILKDQIKNFGQD-AEVAEVGQVLSVGDGIARVYGLDKVQAGEMVEFPGG---IRGMVLNLET
```

Alignement Multiples

ALIGNEMENT GLOBALE

Comparaison des Séquences Homologues

- L'algorithme de l'alignement Globale était proposé par Saul Needleman et Christian Wunsch en 1970

- L'algorithme de Needleman et Wunsch cherche l'alignement qui donne la plus grande similarité,

- Dans un algorithme d'alignement global, ce qui est recherché c'est l'ensemble des séquences avec un score significatif sur une longueur proche de la longueur des deux séquences.

- L'application majeure de ce type d'alignement est l'alignement de séquences de façon à préserver (voire optimiser) la fonction biologique au cours de l'évolution. Cela correspond typiquement à la recherche d'homologues.

ALIGNEMENT GLOBAL

Comparaison des Séquences Homologues

- Il s'agit d'un algorithme de programmation dynamique pour l'alignement global et optimal entre deux séquences.

- Trois étapes a suivre :

1. Remplir une matrice des scores
2. Retour arrière (Backtracing)
3. Génération de l'alignement

ALIGNEMENT GLOBAL

1. Remplir une matrice des scores

- Les deux séquences sont placées dans une matrice de score.
- Pour remplir la matrice nous avons besoin des trois paramètres sont des scores pour :

Identité Substitution Indel

- Dans ce exemple nous allons utiliser les valeur suivant:

Identité=3 Substitution= -1 Indel= -2

		M	P	R	C	L	C	Q	R
P									
Y									
R									
C									
K									
C									
R									

D

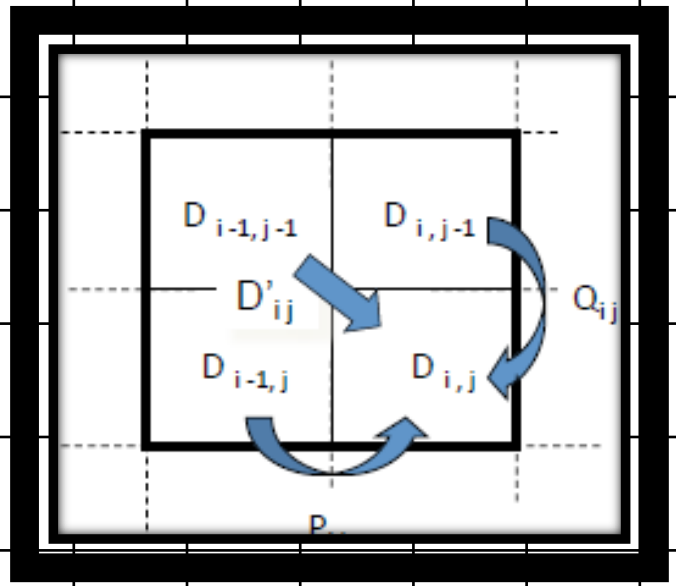
ALIGNEMENT GLOBAL

1. Remplir une matrice des scores

- On doit initialiser la matrice on utilisant le score des indels.
- Avant d'aller plus loin vous devez apprendre de lire les cases dans une matrice.

Identité=3 Substitution= -1 Indel= -2

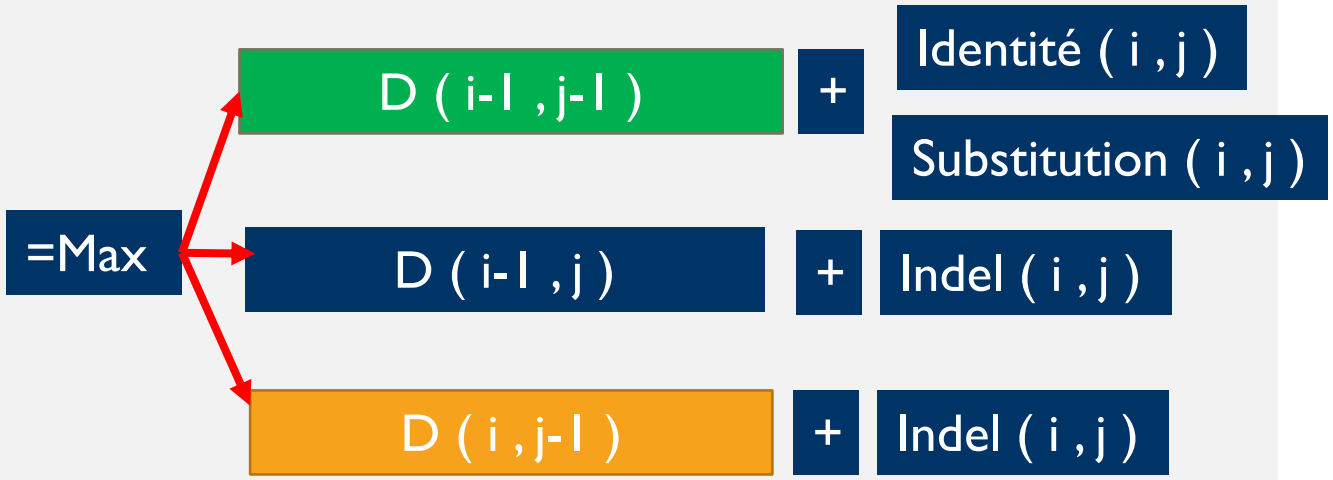
		M	P	R	C	L	C	Q	R
	0	-2	-4	-6	-8	-10	-12	-14	-16
P	-2	D(i,j)							
Y	-4								
R	-6								
C	-8								
K	-10								
C	-12								
R	-14								



ALIGNEMENT GLOBAL

I. Remplir une matrice des scores

- Pour le remplir une case D(i,j) en utilise les équations suivant :



Identité=3 Substitution= -1 Indel= -2

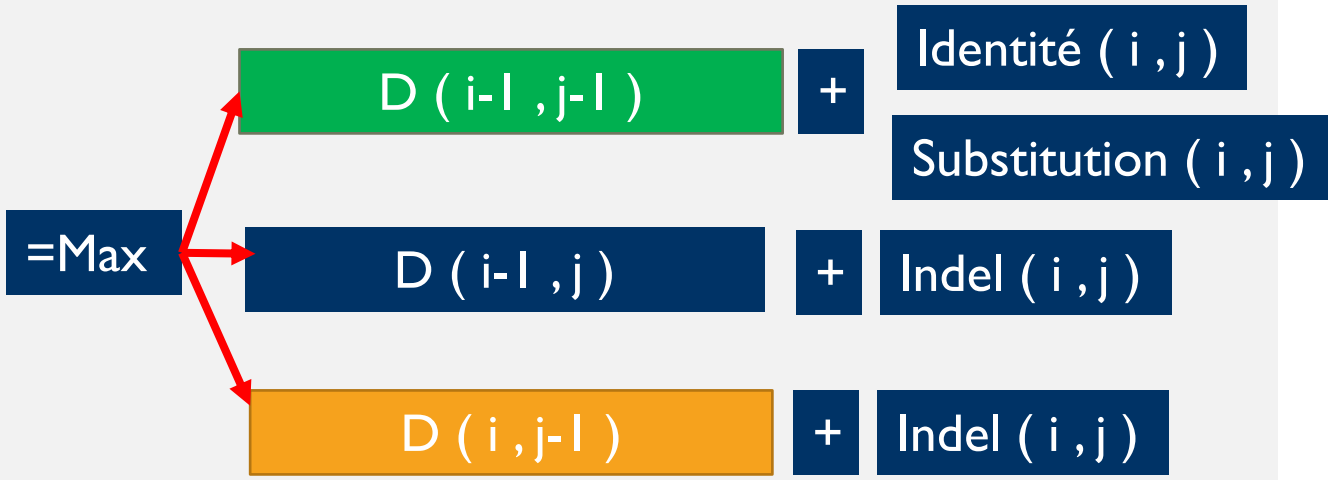
		M	P	R	C	L	C	Q	R
	0	-2	-4	-6	-8	-10	-12	-14	-16
P	-2	-1 -4 -4							
Y	-4								
R	-6								
C	-8								
K	-10								
C	-12								
R	-14								

D

ALIGNEMENT GLOBAL

1. Remplir une matrice des scores

- Pour le remplir une case D(i,j) en utilise les équations suivant :



Identité=3 Substitution= -1 Indel= -2

		M	P	R	C	L	C	Q	R
	0	-2	-4	-6	-8	-10	-12	-14	-16
P	-2	-1	-3						
Y	-4								
R	-6								
C	-8								
K	-10								
C	-12								
R	-14								

D

ALIGNEMENT GLOBAL

2. Retour arrière (Backtracing)

- Le retour arrière commence à la position dans la dernière ligne avec la plus haute valeur

3. Génération de l'alignement

-1	1	0	2	5	4	7	6	8
M	P	--	R	C	L	C	Q	R
--	P	Y	R	C	K	C	--	R

Identité=3 Substitution= -1 Indel= -2

		M	P	R	C	L	C	Q	R
	0	-2	-4	-6	-8	-10	-12	-14	-16
P	-2	-1	1	-1	-3	-5	-7	-9	-11
Y	-4	-3	-1	0	-2	-4	-6	-8	-10
R	-6	-5	-3	2	0	-2	-4	-6	-5
C	-8	-7	-5	0	5	3	1	-1	-3
K	-10	-9	-7	-2	3	4	2	0	-2
C	-12	-11	-9	-4	1	2	7	5	3
R	-14	-13	-11	-6	-1	0	5	6	8

D

ALIGNEMENT GLOBAL

2. Retour arrière (Backtracing)

- Problème:

Le retour arrière commence à la position dans la dernière ligne avec la plus haute valeur, si on a les mêmes valeur on fait quoi ?

		?	?	?	?	?	?	?	?
		-2	-4						
?		-1	1	-1					
?			-1	0					
?			-3	2	0	-2			
?				0	1	3			
?				-2	3	2	2		
?						2	7	5	3
?						0	5	6	8

D

ALIGNEMENT GLOBAL

2. Retour arrière (Backtracing)

- Le retour arrière commence à la position dans la dernière ligne avec la plus haute valeur

3. Génération de l'alignement

- Un autre Problème que on aura 2 alignement probablement correct.
  - Comment choisir ?
  - Comment dire que un alignement et mieux que l'autre.
  - Comment valoriser un alignement.

		?	?	?	?	?	?	?	?
		-2	-4						
?		-1	1	-1					
?			-1	0					
?			-3	2	0	-2			
?				0	1	3			
?				-2	3	2	2		
?						2	7	5	3
?						0	5	6	8

D

ALIGNEMENT GLOBAL

✓ Calculer un score pour chaque alignement

Score d'Identité :

la mesure dans laquelle deux séquences sont invariantes. Une mesure très médiocre car elle ne tient pas compte des subtilités des relations de séquence (par exemple une petite région d'un domaine hautement conservé au sein de deux séquences qui sont autrement très mal conservées).

%id = (NB d'identité/nombre total)*100

le score d'alignement (S):

Une mesure très précise qui est normalisée sur le système de score particulier utilisé.

S = Nb identités * Score identités +
Nb substitution* Score substitution +
Nb gaps * Score gaps

ALIGNEMENT GLOBAL

2. Retour arrière (Backtracing)

-1	1	0	2	5	4	7	6	8
M	P	--	R	C	L	C	Q	R
--	P	Y	R	C	K	C	--	R

✓ S = (5 * 3) + (1 * -1) + (3 * -2) = 8

Identité=3 Substitution= -1 Indel= -2

		M	P	R	C	L	C	Q	R
	0	-2	-4	-6	-8	-10	-12	-14	-16
P	-2	-1	1	-1	-3	-5	-7	-9	-11
Y	-4	-3	-1	0	-2	-4	-6	-8	-10
R	-6	-5	-3	2	0	-2	-4	-6	-5
C	-8	-7	-5	0	5	3	1	-1	-3
K	-10	-9	-7	-2	3	4	2	0	-2
C	-12	-11	-9	-4	1	2	7	5	3
R	-14	-13	-11	-6	-1	0	5	6	8

D

REMARQUE

M	**P**	**--**	**R**	**C**	**L**	**C**	**Q**	**R**
--	**P**	**Y**	**R**	**C**	**K**	**C**	**--**	**R**

- ✓ Dans certains documents, nous ne comptons pas les anges lorsqu'ils se trouvent dans le début ou la fin d'un alignement (dans le calcule de score d'alignement), ce qui peut s'être produit après l'extraction incorrecte de coupures d'ADN.

- ✓ S = (5 * 3) + (1 * -1) + (2 * -2) = 10

A FAIRE

1. Donner l'alignement des deux sequences suivant :
    1. Sequence 1 : A G V S I L N Y A
    2. Sequence 2 : V S I L Y A K

    En utilisant les scores suivant :

    Id=2, Sub=0, Gap=-1
2. Calculer le score d'alignement

A FAIRE

1. Donner l'alignement des deux sequences suivant :
    1. Sequence 1 : MPRCLCQRINCYA
    2. Sequence 2 : PYRCKCRNICIA

    En utilisant les scores suivant :

    Id=2, Sub=-1, Gap=-2
2. Calculer le score d'alignement

RÉFÉRENCES

- PAM series
- Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. (1978). A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure 5, 345--352.
- BLOSUM substitution matrices
- Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89, 10915-9.

ALIGNEMENT GLOBAL

Comparaison des Séquences Homologues

- Il s'agit d'un algorithme de programmation dynamique pour l'alignement global et optimal entre deux séquences.

- Trois étapes a suivre :

1. Remplir une matrice des scores
2. Retour arrière (Backtracing)
3. Génération de l'alignement

ALIGNEMENT GLOBAL

1. Remplir une matrice des scores

- On doit initialiser la matrice (se) (matrice initiale) à partir d'une des matrice de substitution (PAM_,BLOSUM_).

se

	V	T	E	E	R	D	A	F
L								
T								
S								
H								
E								
A								
L								

I. Remplir une matrice des scores

	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

se

	V	T	E	E	R	D	A	F
L								
T								
S								
H								
E								
A								
L								

I. Remplir une matrice des scores

	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

se

	V	T	E	E	R	D	A	F
L	2	-2	-3	-3	-3	-4	-2	2
T	0	3	0	0	-1	0	1	-3
S	-1	1	0	0	0	0	1	-3
H	-2	-1	1	1	2	1	-1	-2
E	-2	0	4	4	-1	3	0	-5
A	0	1	0	0	-2	0	2	-4
L	2	-2	-3	-3	-3	-4	-2	2

ALIGNEMENT GLOBAL

1. Remplir une matrice des scores

- Garder que les valeurs des dernières ligne et colonne, et supprimer les autres valeurs
- La nouvelle matrice (matrice transformé) est appeler S

S

	V	T	E	E	R	D	A	F
L								**2**
T								**-3**
S								**-3**
H								**-2**
E								**-5**
A								**-4**
L	**2**	**-2**	**-3**	**-3**	**-3**	**-4**	**-2**	**2**

ALIGNEMENT GLOBAL

1. Remplir une matrice des scores

$$S(i,j) = se(i,j) + \max \begin{cases} S(i+1,j+1) \\ S(X,j+1) - P \\ S(i+1,Y) - P \end{cases}$$

- Par défaut la valeur de pénalité P = 0 .

S

	V	T	E	E	R	D	A	F
L								2
T								-3
S								-3
H								-2
E								-5
A							S(i,j)	-4
L	2	-2	-3	-3	-3	-4	-2	2

ALIGNEMENT GLOBAL

1. Remplir une matrice des scores

$S(i,j) = se(i,j) + MAX$ { $S(i+1,j+1)$; $S(X,j+1)$; $S(i+1,Y)$ }

$S(6,7) = se(6,7) + MAX$ { $S(7,8)$; $S(X,8)$; $S(7,Y)$ }

S

	J=0	1	2	3	4	5	6	7	8
i=0		V	T	E	E	R	D	A	F
1	L								**2**
2	T								**-3**
3	S								**-3**
4	H								**-2**
5	E								**-5**
6	A								**-4**
7	L	**2**	**-2**	**-3**	**-3**	**-3**	**-4**	**-2**	**2**

ALIGNEMENT GLOBAL

1. Remplir une matrice des scores

S (7,8)

S(6,7) = se(6,7) + MAX S (X,8)

S (7,Y)

2

S(6,7) = 2 + MAX S (X,8)

S (7,Y)

se

	V	T	E	E	R	D	A	F
L	2	-2	-3	-3	-3	-4	-2	2
T	0	3	0	0	-1	0	1	-3
S	-1	1	0	0	0	0	1	-3
H	-2	-1	1	1	2	1	-1	-2
E	-2	0	4	4	-1	3	0	-5
A	0	1	0	0	-2	0	2	-4
L	2	-2	-3	-3	-3	-4	-2	2

S

7	8
A	F
	2
	-3
	-3
	-2
	-5
	-4
-2	2

		6	A						
		7	L	2	-2	-3	-3	-3	-4

ALIGNEMENT GLOBAL

1. Remplir une matrice des scores

2

S(6,7) = 2 + MAX S (X,8)

S (7,Y)

S(6,7) = 2 + 2 = 4

	J=0	I	2	3	4	5	6	7	8
i=0		V	T	E	E	R	D	A	F
I	L								2
2	T								-3
3	S								-3
4	H								-2
5	E								-5
6	A								
7	L	2	-2	-3	-3	-3	-4		

	-4		
-2	2	X	X
	Y		
	Y		

Pas de valeur pour les X et Y

ALIGNEMENT GLOBAL

I. Remplir une matrice des scores

S(6,6) = se(6,6) + MAX

S (7,7)

S (X,7)

S (7,Y)

S(6,6) = 0 + MAX

-2

S (X,7)

S (7,Y)

se

	V	T	E	E	R	D	A	F
L	2	-2	-3	-3	-3	-4	-2	2
T	0	3	0	0	-1	0	1	-3
S	-1	1	0	0	0	0	1	-3
H	-2	-1	1	1	2	1	-1	-2
E	-2	0	4	4	-1	3	0	-5
A	0	1	0	0	-2	0	2	-4
L	2	-2	-3	-3	-3	-4	-2	2

S

6	7	8
D	A	F
		2
		-3
		-3
		-2
		-5
	4	-4
	-2	2

6	A								
7	L	2	-2	-3	-3	-3	-4	-2	2

ALIGNEMENT GLOBAL

1. Remplir une matrice des scores

-2

S(6,6) = 0 + MAX 2

S (7,Y)

S(6,6) = 0 + 2 = 2

	J=0	I	2	3	4	5	6	7	8
i=0		V	T	E	E	R	D	A	F
I	L								2
2	T								-3
3	S								-3
4	H								-2
5	E								-5
6	A								
7	L	2	-2	-3	-3	-3			

	4	-4	
-4	-2	2	X
	Y		
	Y		

ALIGNEMENT GLOBAL

1. Remplir une matrice des scores

-4

S(5,7) = 0 + MAX S(X,8)

2

S(6,7) = 0 + 2 = 2

	J=0	1	2	3	4	5	6	7	8
i=0		V	T	E	E	R	D	A	F
1	L								2
2	T								-3
3	S								-3
4	H								-2
5	E								
6	A	2	3	2	2	0	2		
7	L	2	-2	-3	-3	-3	-4		

	-5		
4	-4	X	X
-2	2		
	Y		

ALIGNEMENT GLOBAL

1. Remplir une matrice des scores

3

S(4,4) = | 1 | + | MAX | 7

3

S(4,4) = | 1 | + | 7 | = 8

	J=0	1	2	3	4	5	6	7	8
i=0		V	T	E	E	R	D	A	F
1	L								2
2	T								-3
3	S								-3
4	H					9	5	1	-2
5	E	2	4	8	8	3	7	2	-5
6	A	2	3	2	2	0	2	4	-4
7	L	2	-2	-3	-3	3	-4	-2	2

ALIGNEMENT GLOBAL

I. Remplir une matrice des scores

- **Compléter la matrice**

	J=0	I	2	3	4	5	6	7	8
i=0		V	T	E	E	R	D	A	F
I	L								**2**
2	T								**-3**
3	S								**-3**
4	H					**9**	**5**	**I**	**-2**
5	E	**2**	**4**	**8**	**8**	**3**	**7**	**2**	**-5**
6	A	**2**	**3**	**2**	**2**	**0**	**2**	**4**	**-4**
7	L	**2**	**-2**	**-3**	**-3**	**-3**	**-4**	**-2**	**2**

ALIGNEMENT GLOBAL

2. Retour arrière (Backtracing)

- Le retour arrière commence à la position dans la dernière ligne avec la plus haute valeur

S

	J=0	1	2	3	4	5	6	7	8
i=0		V	T	E	E	R	D	A	F
1	L	**14**	**7**	**6**	**6**	**4**	**4**	**0**	**2**
2	T	**10**	**12**	**9**	**9**	**6**	**4**	**3**	**-3**
3	S	**8**	**10**	**9**	**9**	**7**	**4**	**3**	**-3**
4	H	**6**	**7**	**9**	**8**	**9**	**5**	**1**	**-2**
5	E	**2**	**4**	**8**	**8**	**3**	**7**	**2**	**-5**
6	A	**2**	**3**	**2**	**2**	**0**	**2**	**4**	**-4**
7	L	**2**	**-2**	**-3**	**-3**	**-3**	**-4**	**-2**	**2**

ALIGNEMENT GLOBAL

2. Retour arrière (Backtracing)

- Le retour arrière commence à la position dans la dernière ligne avec la plus haute valeur

3. Génération de l'alignement

1	2	3	4	5	6	7	8	9
V	T	---	E	E	R	D	A	F
L	T	S	H	E	---	---	A	L
14	12		9	8			4	2

S = 49

S

	J=0	1	2	3	4	5	6	7	8
i=0		V	T	E	E	R	D	A	F
1	L	**14**	7	6	6	4	4	0	2
2	T	10	**12**	9	9	6	4	3	-3
3	S	8	**10**	9	9	7	4	3	-3
4	H	6	7	**9**	8	9	5	1	-2
5	E	2	4	8	**8**	**3**	**7**	2	-5
6	A	2	3	2	2	0	2	**4**	-4
7	L	2	-2	-3	-3	-3	-4	-2	**2**

A FAIRE

	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

1. Donner l'alignement des deux sequences suivant :
    1. Sequence 1 : A G V S I L N Y A
    2. Sequence 2 : V S I L Y A K
2. Calculer le score d'alignement

A FAIRE

A	Ala	4																		
R	Arg	-1	5																	
N	Asn	-2	0	6																
D	Asp	-2	-2	1	6															
C	Cys	0	-3	-3	-3	9														
Q	Gln	-1	1	0	0	-3	5													
E	Glu	-1	0	0	2	-4	2	5												
G	Gly	0	-2	0	-1	-3	-2	-2	6											
H	His	-2	0	1	-1	-3	0	0	-2	8										
I	Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4									
L	Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4								
K	Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5							
M	Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5						
F	Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6					
P	Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7				
S	Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4			
T	Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5		
W	Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	
Y	Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7
V	Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
		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

1. Donner l'alignement des deux sequences suivant :
    1. Sequence 1 : MPRCLCQRINCYA
    2. Sequence 2 : PYRCKCRNICIA
2. Calculer le score d'alignement

ALIGNEMENT LOCAL

Meilleur Chevauchement Entre Séquences

- Dans l'algorithme d'alignement local, ce qui est recherché ce sont des zones de similitudes dans des séquences quelconques (homologues ou pas).

- Cet alignement local sera privilégié si le plus long chevauchement entre deux séquences est recherché.

- L'algorithme de Algorithme Smith & Waterman (1981) cherche l'alignement qui donne le Meilleur Chevauchement Entre Séquences

- Trois étapes a suivre :

1. Remplir une matrice des scores
2. Retour arrière (Backtracing)
3. Génération de l'alignement

ALIGNEMENT LOCAL

1. Remplir une matrice des scores

- Les deux séquences sont placées dans un tableau de score.
- La première ligne et la première colonne sont initialisées à 0.

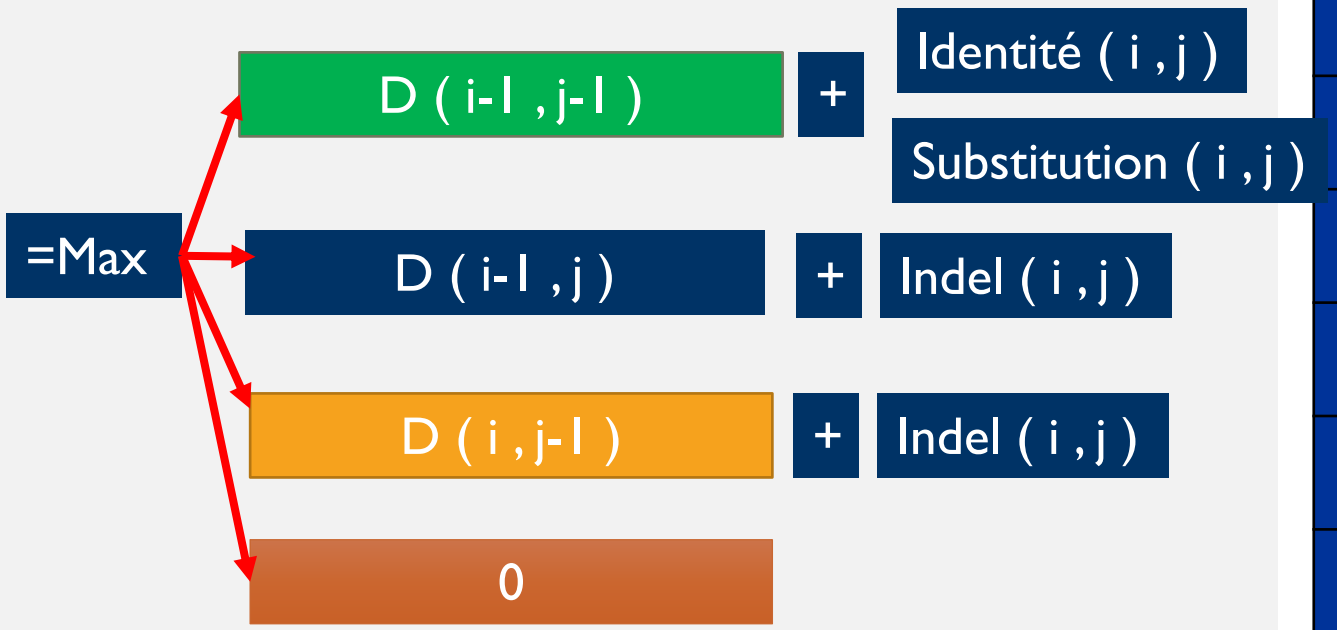
		L	I	B	R	E	S	E	Q	E	N
	0	0	0	0	0	0	0	0	0	0	0
S	0										
E	0										
Q	0										
A	0										
N	0										
C	0										
E	0										

D

ALIGNEMENT LOCAL

I. Remplir une matrice des scores

- Pour le remplir une case D(i,j) en utilise les équations suivant :



Identité=2 Substitution= 0 Indel= -I

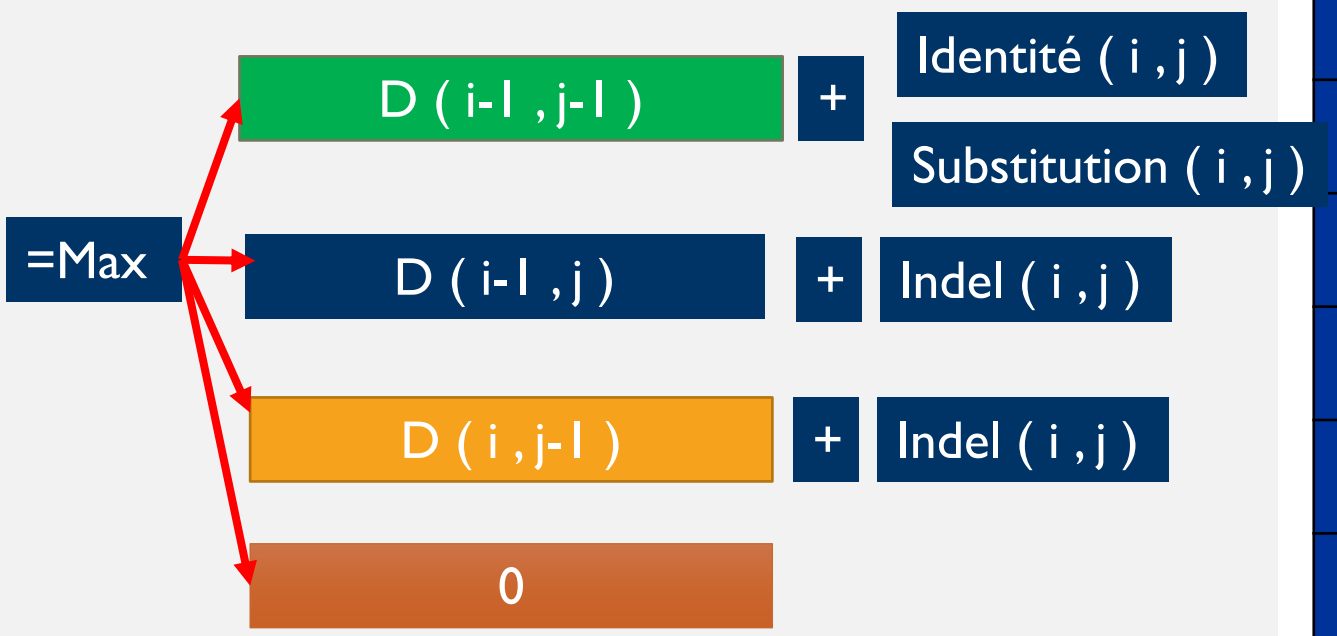
		L	I	B	R	E	S	E	Q	E	N
	0	0	0	0	0	0	0	0	0	0	0
S	0	0 / 0 / -I / -I									
E	0										
Q	0										
A	0										
N	0										
C	0										
E	0										

D

ALIGNEMENT LOCAL

I. Remplir une matrice des scores

- Pour le remplir une case D(i,j) en utilise les équations suivant :



Identité=2 Substitution= 0 Indel= -1

		L	I	B	R	E	S	E	Q	E	N
	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	2	1	0	0	0
E	0	0	0	0	0	2	1	4	3	2	1
Q	0	0	0	0	0	1	2	3	6	5	4
A	0	0	0	0	0	0	1	2	5	6	5
N	0	0	0	0	0	0	0	1	4	5	8
C	0	0	0	0	0	0	0	0	3	4	7
E	0	0	0	0	0	2	1	2	2	5	6

D

ALIGNEMENT LOCAL

2. Retour arrière (Backtracing)

- Le retour arrière commence à la position dans la dernière ligne avec la plus haute valeur

3. Génération de l'alignement

8	6	6	4	2
N	E	Q	E	S
N	A	Q	E	S

Identité=2 Substitution= 0 Indel= -1

		L	I	B	R	E	S	E	Q	E	N
	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	2	1	0	0	0
E	0	0	0	0	0	2	1	4	3	2	1
Q	0	0	0	0	0	1	2	3	6	5	4
A	0	0	0	0	0	0	1	2	5	6	5
N	0	0	0	0	0	0	0	1	4	5	8
C	0	0	0	0	0	0	0	0	3	4	7
E	0	0	0	0	0	2	1	2	2	5	6

D

ALIGNEMENT LOCAL

3. Génération de l'alignement

8	6	6	4	2
N	E	Q	E	S
N	A	Q	E	S

SEQ1: LIB R E SEQEN _ _
                  | | | |
SEQ2: _ _ _ _ _ SEQ AN C E

Identité=2 Substitution= 0 Indel= -1

		L	I	B	R	E	S	E	Q	E	N
	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	0	0	2	1	0	0	0
E	0	0	0	0	0	2	1	4	3	2	1
Q	0	0	0	0	0	1	2	3	6	5	4
A	0	0	0	0	0	0	1	2	5	6	5
N	0	0	0	0	0	0	0	1	4	5	8
C	0	0	0	0	0	0	0	0	3	4	7
E	0	0	0	0	0	2	1	2	2	5	6

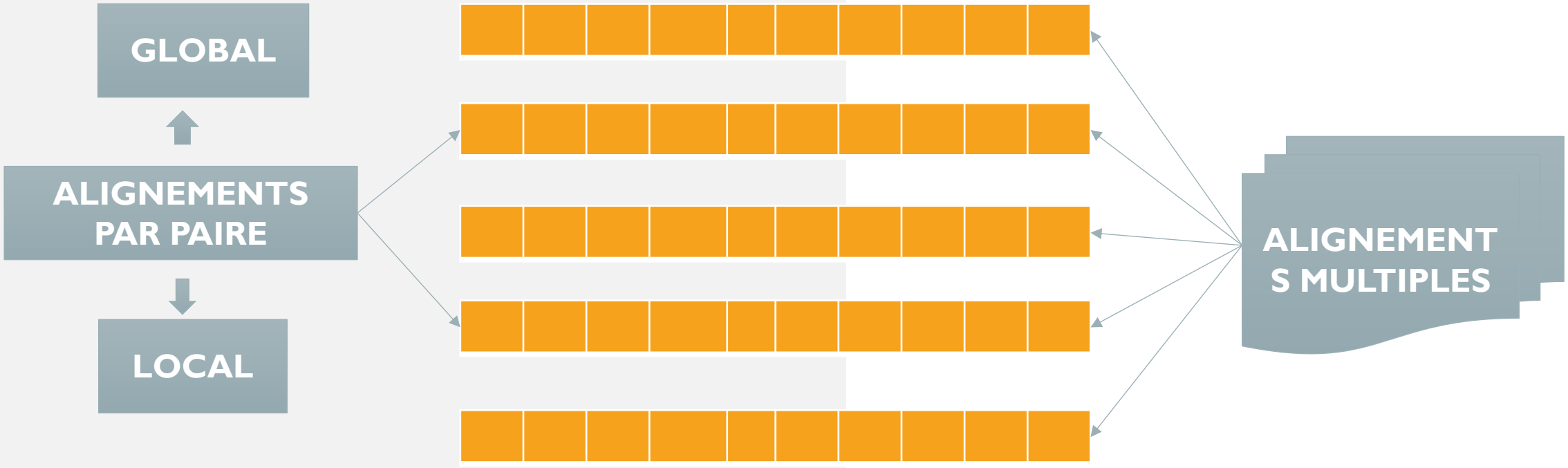
D

QUEL TYPE D'ALIGNEMENT NOUS DEVRIONS CHOISIR?

- Le choix de l'alignement global ou local revient donc à l'utilisateur biologiste en fonction des objectifs poursuivis.

- D'un point de vue biologique, l'alignement le plus pertinent est celui qui retrace le déroulement évolutif le plus probable.

ALIGNEMENTS MULTIPLES



ALIGNEMENTS MULTIPLES

Alignement multiple de séquences de sous-unités d'ATP synthase.

ALIGNEMENTS MULTIPLES

- Vous avez appris que l'alignement de séquence par paire permet l'inférence de la fonction basée sur l'hypothèse qu'un bon alignement de séquence = fonction connexe. Cependant, que se passe-t-il si un bon alignement ne se produit pas entre les résidus conservés?

Vous pourriez faire une mauvaise hypothèse et une inférence incorrecte.

- Une grande raison pour laquelle nous voulons générer un MSA est de vérifier les alignements avant de faire une telle supposition et une inférence - parce qu'il y a plus d'une autre séquence à comparer. Par conséquent, MSA peut réellement fournir un alignement plus précis que les méthodes par paires,

ALIGNEMENTS MULTIPLES

L'alignement multiple est effectué à l'une des utilisations suivantes :

- Identifier les régions de forte similarité et de différence : Les régions conservées peuvent avoir des fonctionnalités importantes.
- Créer une séquence consensus ,
- Contribuer à la prédiction des structures secondaires et tertiaires de nouvelles séquences,
- L'alignement multiple est une étape préliminaire pour la construction d'arbres phylogénétiques (dendrogrammes),

RÉFÉRENCES

- Notes de cours de Bio-informatique. Pr DJEKOUN A. Pr HAMIDECHI M. A.
- Bioinformatic Methods I N. Provart. 20016
- Bio-informatique: Cours et applications. G.Deléage, M.Gouy. Dunod. 2015 - 2ème édition.
- Notes de cours de Bio-informatique. Notes de Philippe Gambette. novembre 2004.
- Biologie Moléculaire. C. Housset. A. Raisonnier2009.