Cours de Bioinformatique

Licence

Dr Benchouieb I
Département de Biologie moléculaire et cellulaire
Université MSB Jijel, Algérie.

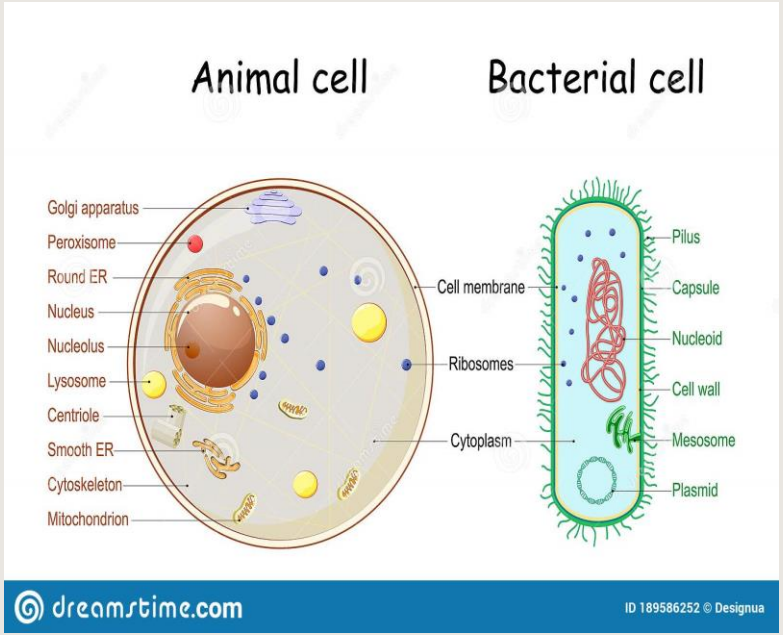
Contenu de la matière

03 Méthodes d’exploitation et alignement

Comparaison des séquences, comme méthode de prédiction

Depuis ses origines la Biologie se nourrit des « comparaisons » entre les différentes manifestations du vivants (cellules, organes, organismes).

Ainsi l'étude des séquences biologiques d'acides nucléiques ou de protéines n'échappe pas à cette logique, **par le biais de l'alignement et la comparaison des séquences : outils essentiels du bio-informaticien.**



L’exploitation des données issues des programmes de séquençage, notamment des séquences d’ADN, permet de déduire la séquence des acides aminés des protéines codées par ces gènes.

Relation entre les informations génétiques portées sur l'ADN et les protéines

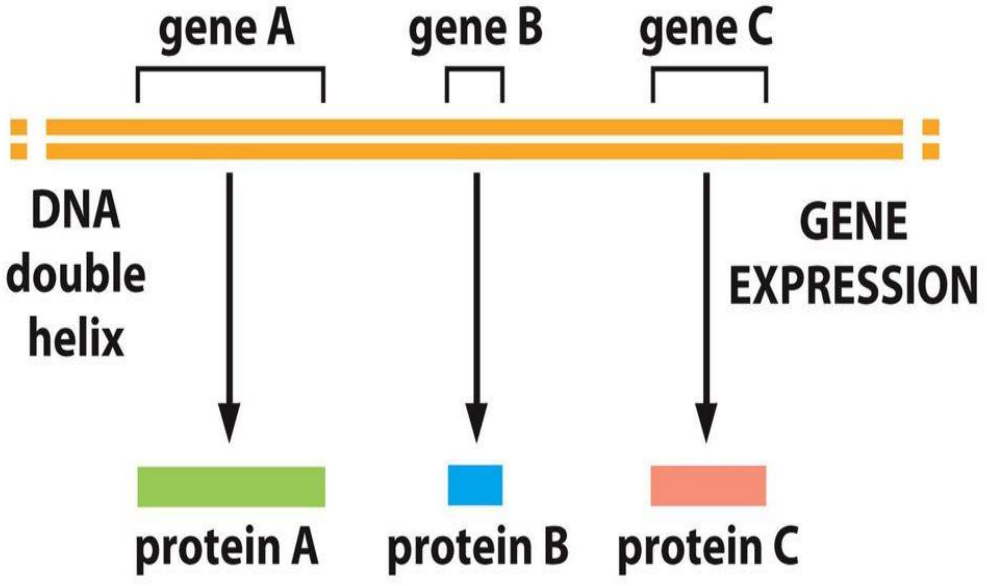


Figure 4-7 Molecular Biology of the Cell 6e (© Garland Science 2015)

Le génome d'une bactérie code pour 1000 à 10 000 protéines différentes (plus de 4000 pour Escherichia Coli), Celui d'un eucaryote pour 5000 à 50 000 (plus de 12 000 pour la drosophile).
Et Il faut Annoter les séquences, c'est-à-dire identifier leur fonction dans la cellule.

Dans certains cas, on trouve dans les banques, la séquence d'un gène apparenté chez une espèce voisine.
Exemple: certaines protéines du chimpanzé sont presque identiques aux protéines correspondantes chez l'homme.

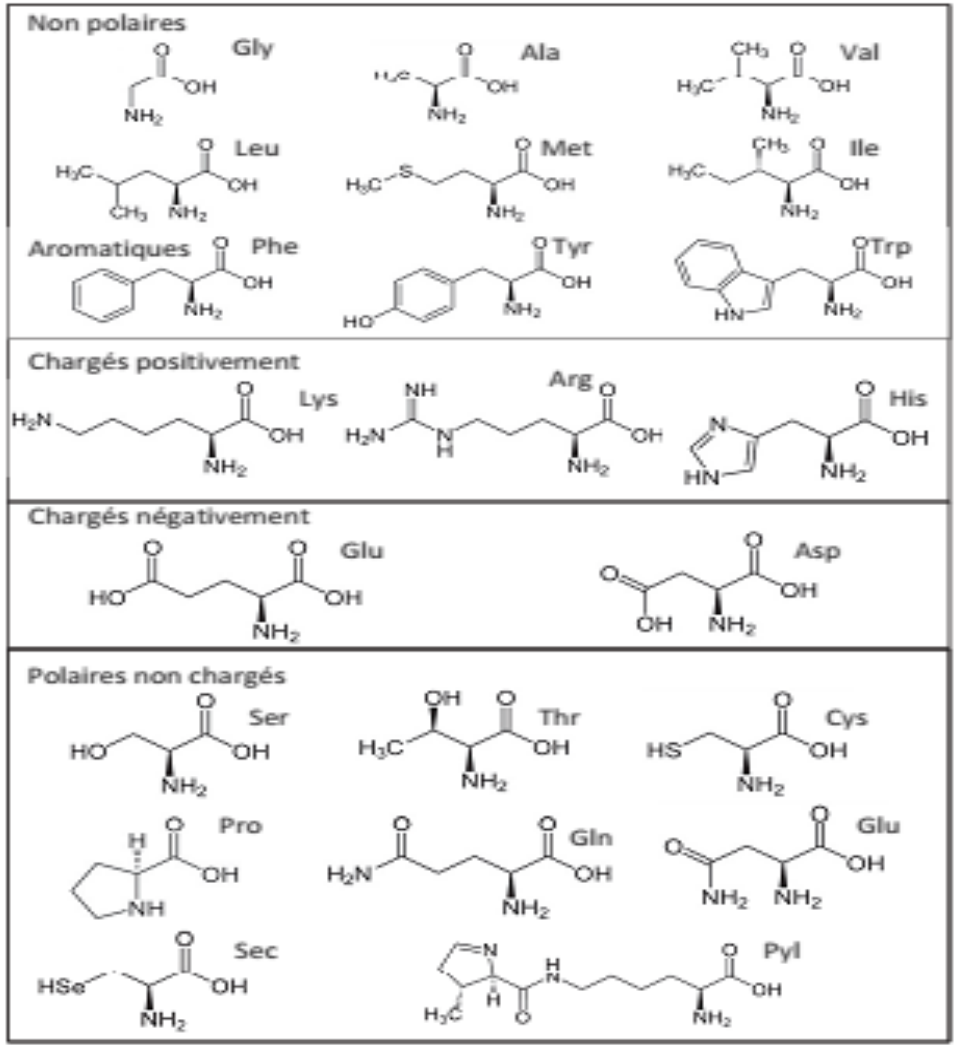
Dans le cas où les protéines nouvellement identifiées (grâce au séquençage) ne possèdent pas d'homologue parmi celles dèja connues, il faut livrer une analyse plus fine pour mettre au jour des ressemblances lointaines et tenter d'attribuer une fonction à la protéine étudiée.

Les annotations génomiques visent à associer le séquençage d'un génome ou d'un transcriptome à **une information biologique exploitable** qui sont indispensables à **la compréhension des fonctions biologiques des organismes.**

Les séquences protéiques

Les protéines naturelles sont constituées d'**acides aminés,** Il existe 20 acides aminés principaux dans les protéines naturelles. La correspondance entre les acides aminés, leur abréviations et leur structure chimique est donnée dans la figure:

A	Alanine	Ala
C	Cysteine	Cys
D	Aspartic Acid	Asp
E	Glutamic Acid	Glu
F	Phenylalanine	Phe
G	Glycine	Gly
H	Histidine	His
I	Isoleucine	Ile
K	Lysine	Lys
L	Leucine	Leu
M	Methionine	Met
N	Asparagine	Asn
O	Pyrrolysine	Pyl
P	Proline	Pro
Q	Glutamine	Gln
R	Arginine	Arg
S	Serine	Ser
T	Threonine	Thr
U	Sélénocystéine	Sec
V	Valine	Val
W	Tryptophane	Trp
Y	Tyrosine	Tyr
B		
Z		
X	Inconnu	



Encart 1.1 Combien de séquences protéiques différentes peut-on générer en théorie ?

Le nombre de séquences différentes de longueur N qu'il est possible de générer en prenant les 20 acides aminés principaux est 20^N.

Exemples :

Peptide (5 acides aminés) : 20^5

Protéine de taille standard moyenne de 400 acides aminés : 20^{400}

Protéome humain (soit ~20 000 protéines de longueur moyenne 400) : $20^{8\,000\,000}$

Notre molécule test: Le récepteur humain de l'androstérone

Pour illustrer la problématique et les méthodes utilisées, nous allons utiliser une séquence protéique de référence: **le récepteur nucléaire d'une hormone** stéroïde sexuelle, l'androstérone.
Cette protéine est localisée dans le noyau de nos cellules et sa fonction est de stimuler la transcription d'un certain nombre de gènes, en réponse à un signal hormonal. Elle pénètre dans la cellule-cible, se lie à son récepteur dans le noyau, et celui-ci se fixe alors à une séquence spécifique sur l'ADN en amont des promoteurs de transcription des gènes qu'il faut stimuler.
Ce qui permet le recrutement de l'ARN polymérase et l'activation de la trascription de la séquence protéique de la partie responsable de la liaison à l'ADN du récepteur humain de l'androstérone est indiquée ci-dessous (acides amin 550 à 620 sur un total de 919 dans la protéine complète):

550
DYYFPPORTCLICGDEASGCHYGALTCGSCKVFFKRAAEGKQKYLCASRNDCTID
KRRENCESCRIRECTL
La structure 3D du domaine correspondant lié à sa cible ADN est indiquée 620

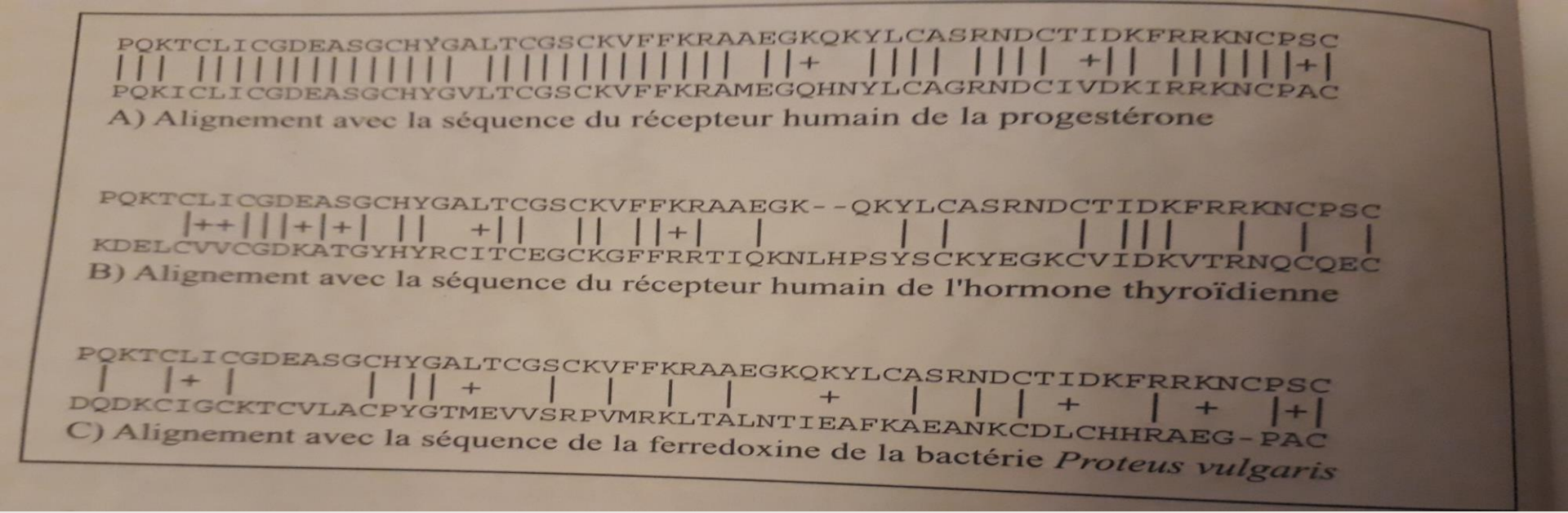


Figure 2. Comparaison de la séquence du domaine de liaison de l'ADN du récepteur de l'androstérone (ligne du haut) avec celles de trois autres protéines. Les conservations strictes sont indiquées par des barres verticales. Les remplacements par des acides aminés de fonction chimique similaires sont indiqués par + . Les tirets indiquent les positions où il a fallu faire des insertions pour poursuivre l'alignement.

Homologies de séquence homologies fonctionnelles

Le postulat fondamental de toute l'analyse des séquences est le suivant:

Les séquences de deux molécules de fonctions apparentées vont en général présenter des ressemblances.
Réciproquement, deux molécules dont les séquences présentent des ressemblances ont probablement des fonctions apparentées

En résumé,

Ceci est vrai pour les protéines où la nature des acides aminés et les fonctions chimiques portées par les chaînes latérales des acides aminés vont être conservées, en particulier au niveau du site actif. Ça l'est également dans une moindre mesure pour l'ADN, où des régions comme des promoteurs de transcription où des sites de liaisons pour une protéine donnée vont en général présenter des similitudes importantes.
L'objectif du Bioinformaticien est de **détecter par des méthodes informatiques ces ressemblances** pour en tirer des conclusions biologiques:
-Si deux molécules de fonctions connues se ressemblent, on peut en conclure qu'une partie de leur mécanisme d'action doit être commun.
- Si la séquence d'une protéine inconnue ressemble à celle d'une protéine Connue, cela donne une indication sur la fonction de la première.
-Plus la ressemblance est forte et étendue et plus l'homologie de fonction est probable.
L'identification des parties les plus conservées ou ressemblantes renseigne sur les parties importantes de la protéine et donc sur la localisation probable de son site actif.

Comparaison de séquences, objectifs

- Détermination la fonction et la structure d'une séquence
- Détection de régions fonctionnelles au sein des séquences
- Etude des processus de l'évolution à l'échelle moléculaire
- Construire la **phylogénie** des espèces.

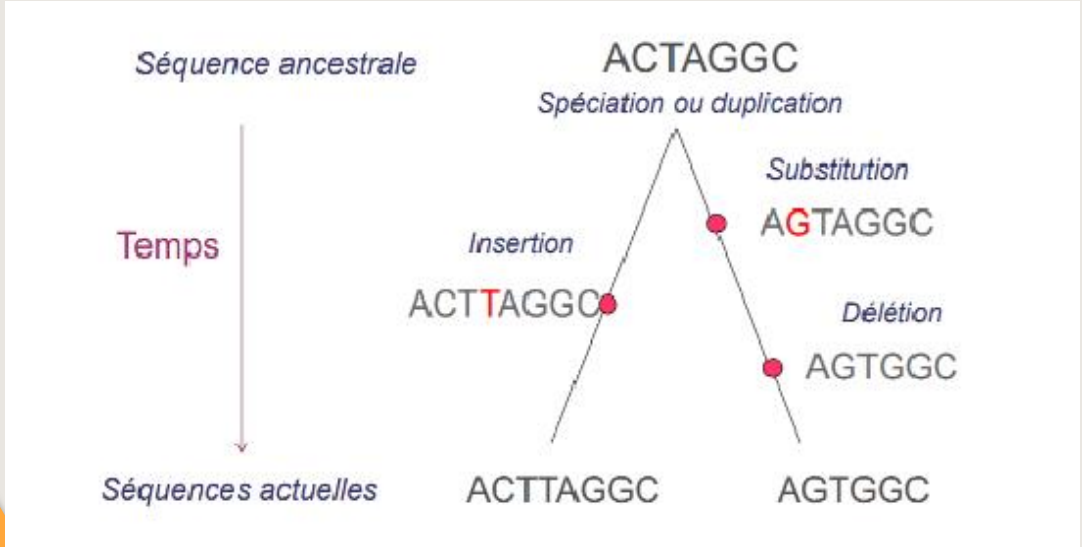
Deux séquences sont homologues si elles sont issues d'une séquence ancestrale commune.

Comparaison de séquences, objectifs

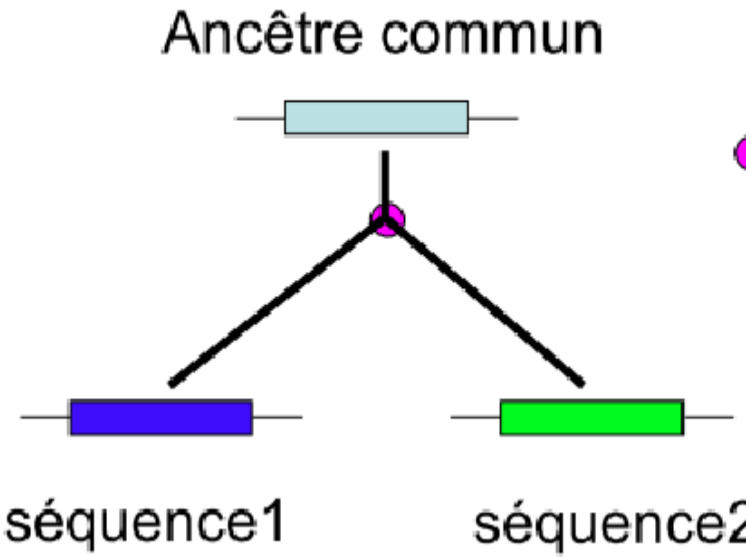
Deux séquences homologues sont:

➢ Orthologues si elles sont issues d'une séquence ancestrale commune et ont divergé suite à un évènement de spéciation.

➢ Paralogues si elles sont issues d'une séquence ancestrale commune et ont divergé suite à un évènement de duplication.



Comparaison de séquences, objectifs



- Événement évolutif (spéciation, duplication)

Comparaison de séquences, objectifs

Alignement de séquences : représentation de deux ou plusieurs séquences biologiques (ADN, ARN ou protéines) les unes sous les autres, de manière à en faire ressortir les régions homologues.

• Ceci nécessite en général l'introduction de "trous" (gaps) à certaines positions dans les séquences, de manière à aligner les résidus communs sur des colonnes successives.

Indel est un mot-valise 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.

Comparaison de séquences, objectifs

Alignez les séquences suivantes:

Seq1 : GTTACGA

Seq2 : GTTGGA

Comparaison de séquences, objectifs

Seq1 : GTTACGA

Seq2 : GTTGGA

Alignement 1

Seq1 GTTACGA

Seq2 GTTG- GA

*** **

Alignement 2

Seq1 GTTACGA

Seq2 GTT- GGA

*** **

Comparaison de séquences, objectifs

Alignement de
2 séquences
Alignement par paire:

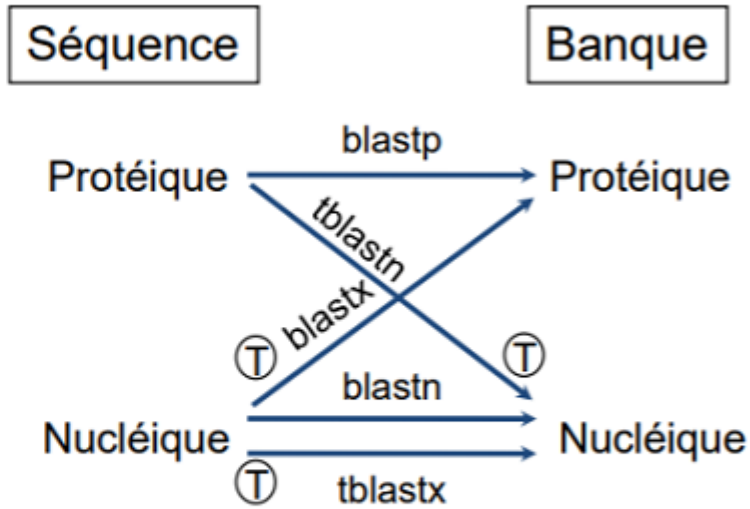
```
AACTGCATTGTA
AA-TGCAT-GTA
** ***** ***
```

Alignement multiple:

```
AACTGCATTGTA
AA-TGCAT-GTA
AACTCCATTGTA
AA-TGAATT-TA
** *  **  **
```

Comparaison de séquences, objectifs

- blastp : comparaison de séquences protéiques.
- blastn : comparaison de séquences nucléiques non codantes.
- blastx : identification de séquences codantes dans une séquence non annotée.
- tblastn : identification d'un gène codant pour une protéine d'intérêt dans des séquences non annotées.



Alignement multiple: Alignements de plus de deux séquences

L'alignement multiple consiste à **aligner collectivement un ensemble de séquences homologues,** comme des séquences de protéines assurant des fonctions similaires dans différentes espèces vivantes. L'alignement multiple permet d'identifier les régions très conservées qui sont en général associées à des fonctions biologiques importantes.

Principe: alignement portant sur **plusieurs séquences à la fois** et dans leur intégralité. Il permet de mettre en évidence les relations entre séquences que l'on ne peut pas visualiser en comparant les séquences 2 à 2.

il faut <u>généraliser la notion de fonction de score</u> définie plus haut pour deux séquences.
On ne compare les acides aminés/les nucléotides deux à deux, mais **à l'intérieur d'une colonne de l'alignement multiple.**

Score colonne= $\sum M(Xi, Xj)$

M : Matrice de similarité , X_i et X_j : acides aminés (ou nucléotides) dans la colonne, au niveau des lignes *i* et *j* de l'alignement.
Ce score inclut toutes les combinaisons deux à deux d'acides aminés (ou de nucléotides) dans la colonne, d'où la dénomination de « score de la somme des paires »

Applications

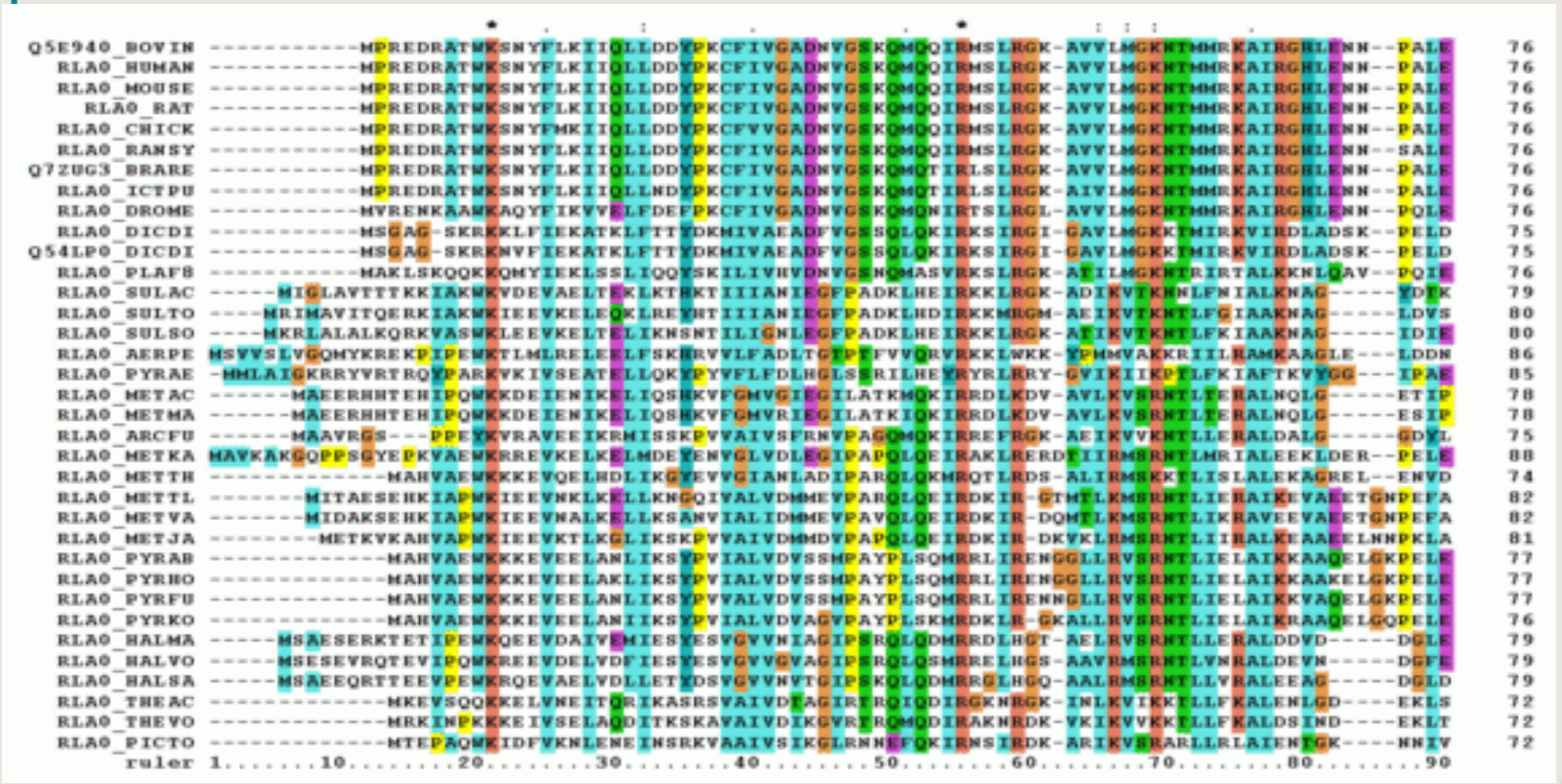
1. Identification de sites fonctionnels importants (conservés): L'alignement multiple de séquences entre espèces éloignées permet l'identification rapide des sites dont la conservation est requise pour la fonction.
2. Prédiction de fonction: Détecter des résidus identiques ou similaires ayant un rôle fonctionnel ou structural.
3. Prédiction de structure: La structure 3D des protéines étant plus conservée que la séquence primaire, on aura une prédiction de structure.

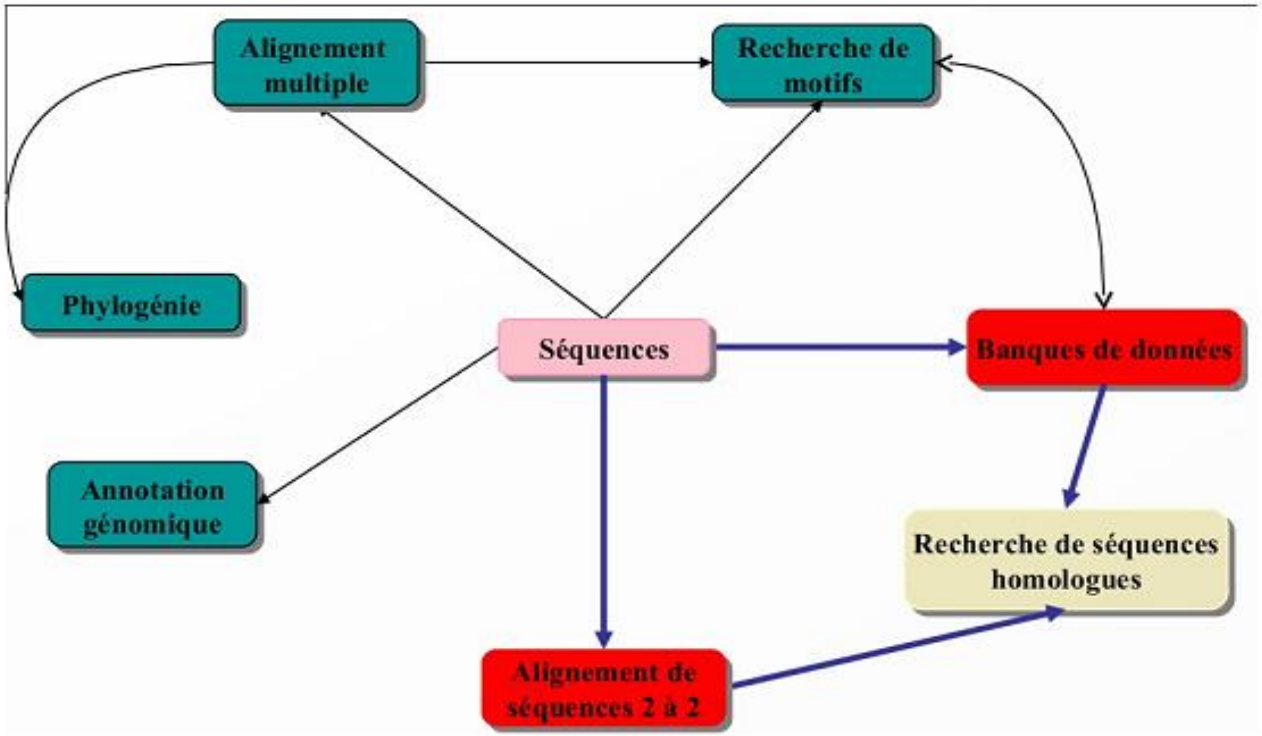
4. Recherche d'amorces de PCR: Pour dessiner des amorces (primers) de PCR pour amplifier le gène dans une nouvelle espèce.

5. Caractériser une nouvelle famille de protéines

6. Détecter une homologie entre différentes séquences
7. Etablir une phylogénie

ClustalW est un des programmes d'alignements multiples les plus utilisés, relativement performant et présent sur de très nombreux serveurs.


Q5E940_BOVIN
RLA0_HUMAN
RLA0_MOUSE
RLA0_RAT
RLA0_CHICK
RLA0_RAMSY
Q7ZUG3_BRARE
RLA0_ICTPU
RLA0_DROME
RLA0_DICDI
Q54LP0_DICDI
RLA0_PLAF8
RLA0_SULAC
RLA0_SULTO
RLA0_SULSO
RLA0_AERPE
RLA0_PYRAE
RLA0_METAC
RLA0_METMA
RLA0_ARCFU
RLA0_METKA
RLA0_METTH
RLA0_METTL
RLA0_METVA
RLA0_METJA
RLA0_PYRAB
RLA0_PYRHO
RLA0_PYRFU
RLA0_PYRKO
RLA0_HALMA
RLA0_HALVO
RLA0_HALSA
RLA0_THEAC
RLA0_THEVO
RLA0_PICTO
ruler



Le « dogme central » de la bioinformatique: La déduction par homologie

En bio-informatique, l'**alignement de séquences**

est une manière de représenter deux ou plusieurs séquences de macromolécules biologiques (ADN, ARN ou protéines) les unes sous les autres, de manière à en faire ressortir les régions homologues ou similaires.

Son objectif est de disposer les composants (nucléotides ou acides aminés) pour identifier les zones de concordance. Ces alignements sont réalisés par des programmes informatiques dont l'objectif est de **maximiser le nombre de coïncidences entre nucléotides ou acides aminés dans les différentes séquences.**

Ceci nécessite en général **l'introduction de « trous »** à certaines positions **dans les séquences, de manière à aligner les caractères communs** sur des colonnes successives.

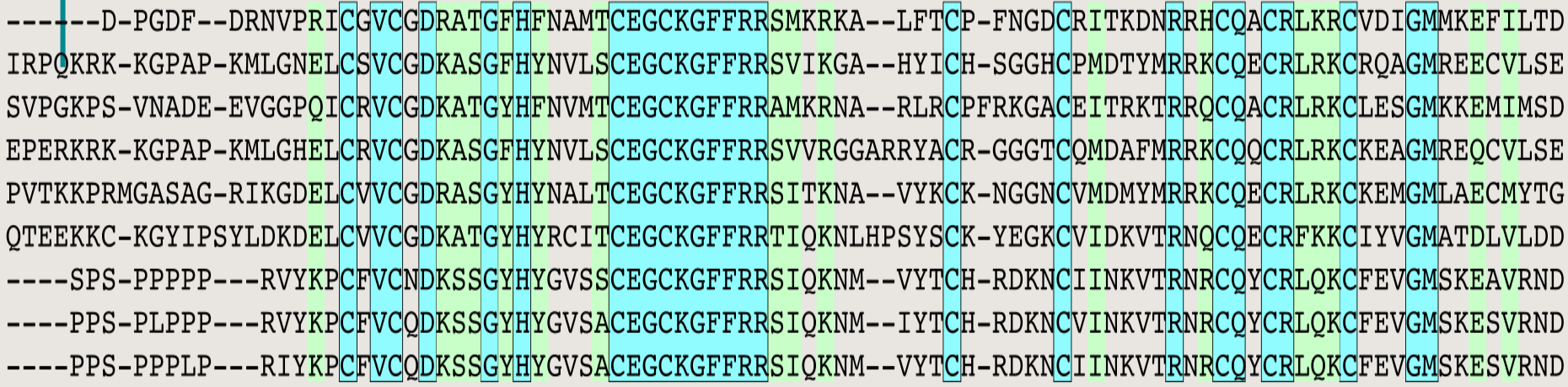


Figure . Alignement de séquences de récepteurs nucléaires. Les acides aminés conservés sont surlignés en bleu et en vert. Par endroits, on a inséré des trous, symbolisés par des tirets « - », pour permettre un alignement optimal.

Homologue ≠ Similaire

Similaire (%) = Présence d'un ensemble de position identiques et conservatives dans deux séquences

Homologues = fait référence à une parenté évolutive entre séquences

Principes de bases de l'alignement de séquences:

La diversité génétique est due à des mutations ponctuelles et à des insertions / délétions apparue au cours de l'évolution. Comparer les séquences biologiques par alignement permet de déterminer le degré de similarité et d'émettre une hypothèse quant à leur parenté évolutive (homologie). Il faut distinguer ces termes:

- **Identité:** estimation de la fraction de résidus identiques entre deux séquences (fondée sur un alignement).

- **Similarité:** estimation de la fraction de résidus similaires entre deux séquences, lorsque leur score de substitution est supérieur à 0.

- **Homologie:** désigne une parenté évolutive entre deux séquences, c'est-à-dire dérivant d'un ancêtre commun.

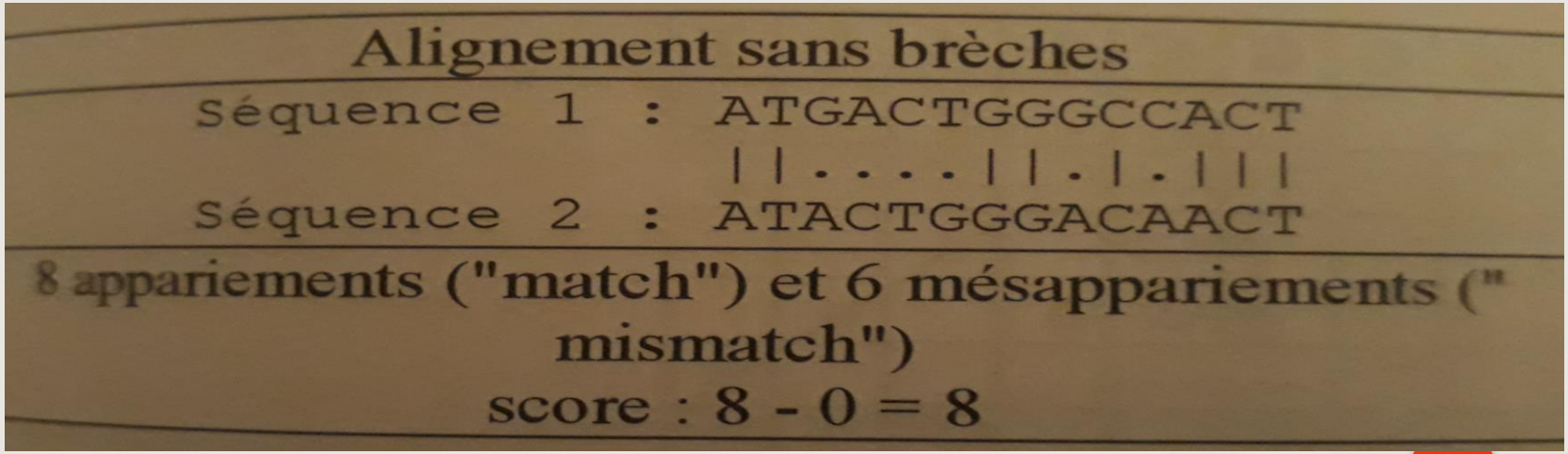
Alignement et détermination de score:

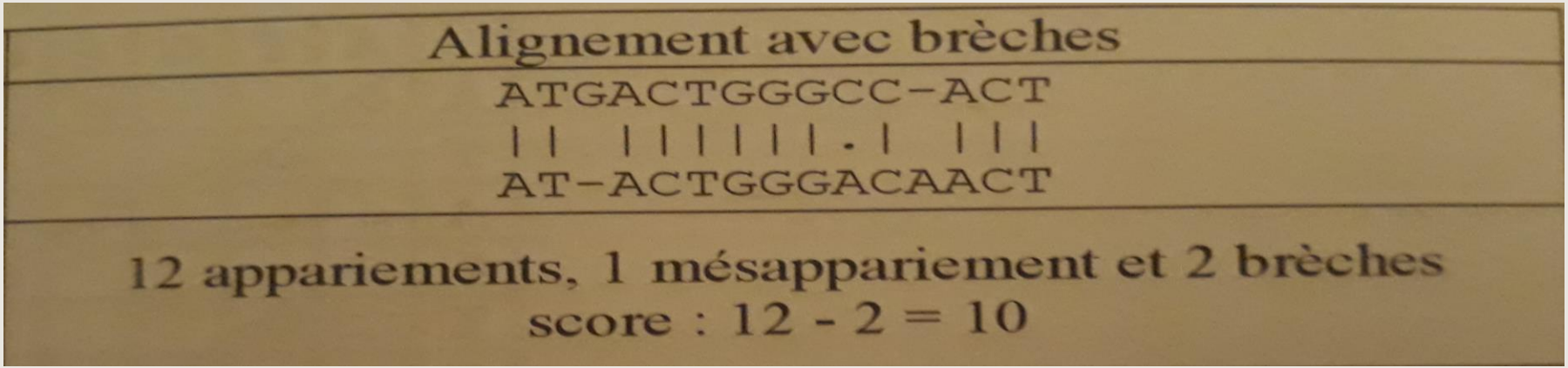
Aligner 2 séquences c'est chercher le maximum d'appariements entre les lettres qui les composent (Nucléotides ou résidus d'acides aminés) avec le minimum de misappariements et de brèches.

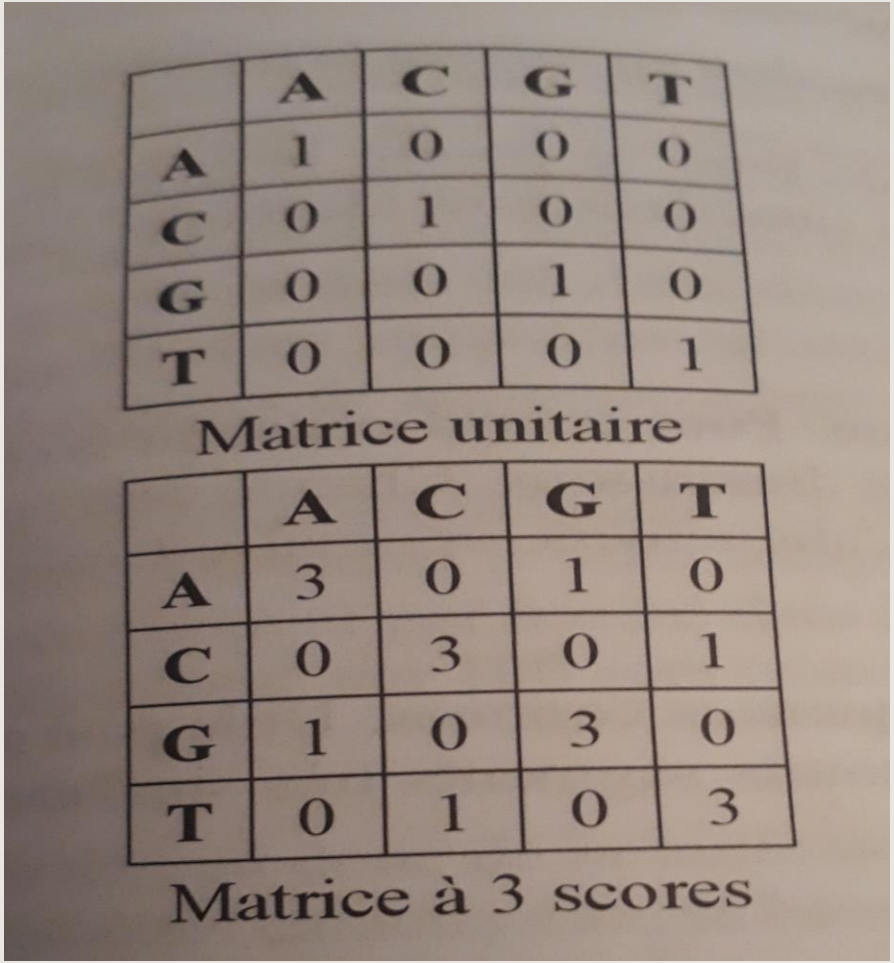
Remarque: les pénalités des brèches doivent être suffisamment couteuses pour éviter les alignement sans signification biologique.

Score= $\sum$ scores élémentaires - $\sum$ scores pénalités

Exemple: determination de score avec une matrice unitaire :
l'appariemement vaut +1,
le mésappariement vaut 0
et la brèche vaut -1




Alignement avec brèches
ATGACTGGGCC-ACT
|| |||||||.| |||
AT-ACTGGGACAACT
12 appariements, 1 mésappariement et 2 brèches
score : 12 - 2 = 10



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

Matrice unitaire

	A	C	G	T
A	3	0	1	0
C	0	3	0	1
G	1	0	3	0
T	0	1	0	3

Matrice à 3 scores

Matrice à 3 score distingue:

✓ les transitions (A↔ G et C ↔ T)

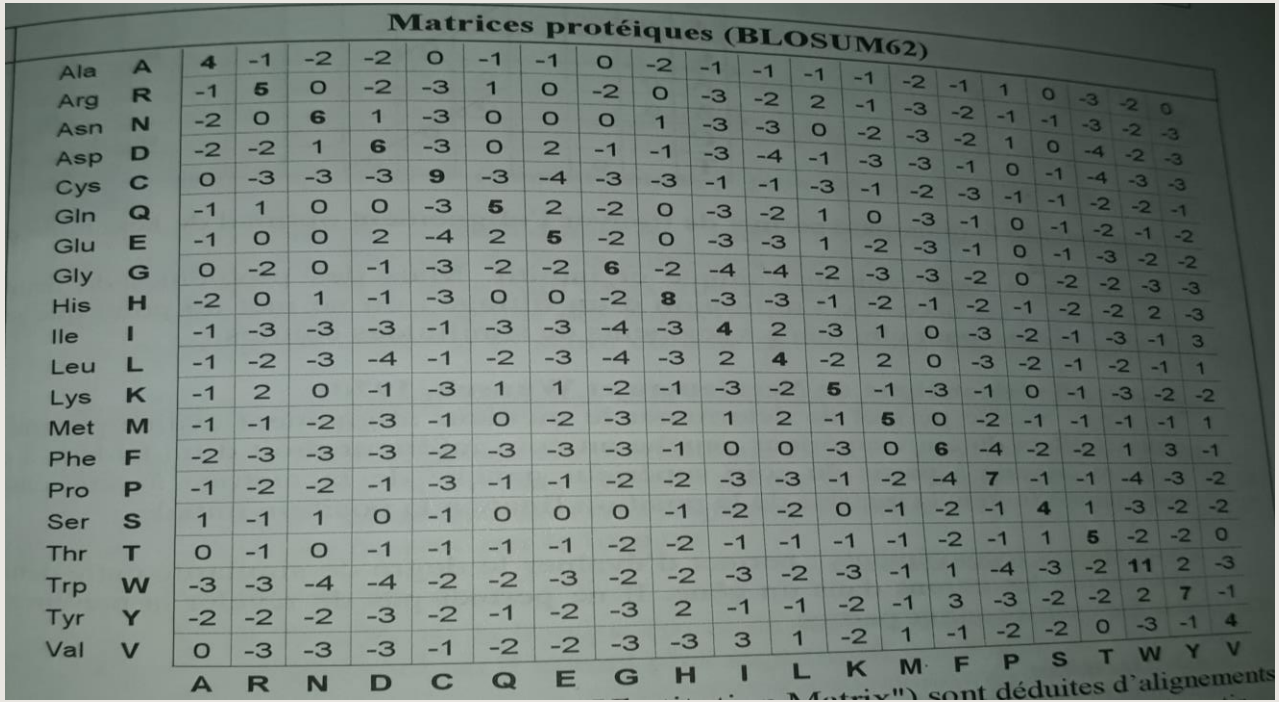
✓Les transversions (modification d'une base purique par une base pyrimidique et inversement)

Les matrices protéiques (BLOSUM 62)

Les matrices **BLOcks SUbstitution Matrix** sont déduites d'alignements de fraguements (Blocks) de protéines très éloignées. Déduites à partir de séquences ayant **62 %** de **similitude**. Elles sont bien adaptées aux recherches de séquences dans les banques de données (BLAST, FASTA).

Chaque score donne le cout de remplacement d'un résidu par un autre:

- ❑ les acides aminés rares ont des scores élevés (Trp, Cys, His..)
- ❑ les acides aminés communs ont des scores faibles (Ala, Leu, Ile,..)
- ❑ les substitutions entre acides aminés similaires sont peu pénalisantes (elles peuvent se produire sans affecter l'activité de la protéines (exp: Lys↔ Arg)



Matrices protéiques (BLOSUM62)

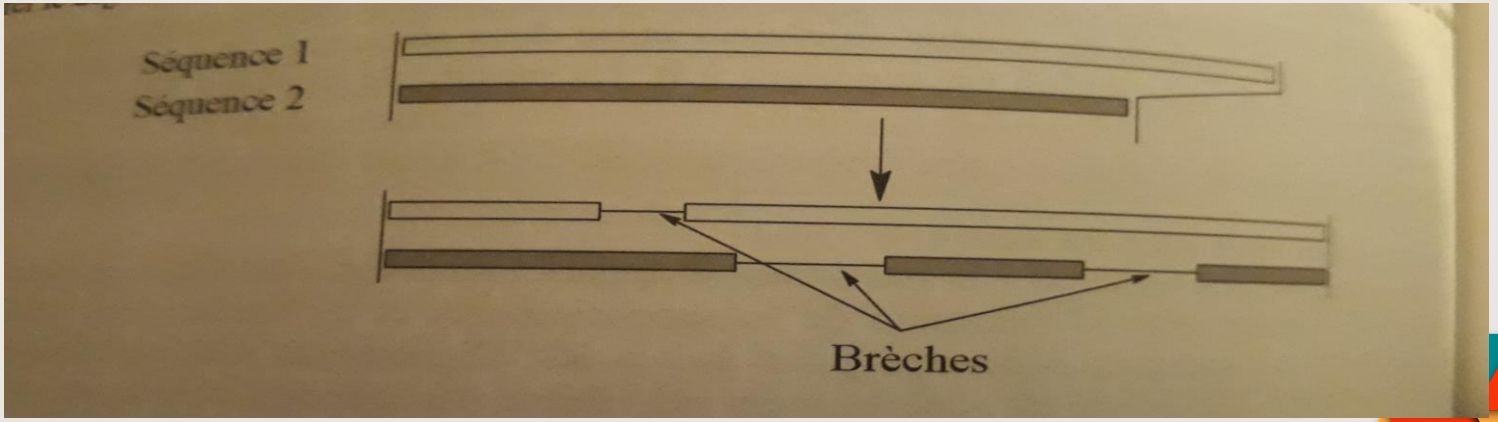
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	**4**	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
Arg	R	-1	**5**	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
Asn	N	-2	0	**6**	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
Asp	D	-2	-2	1	**6**	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
Cys	C	0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

Autre matrice PAM: Point Accepted Mutation

Déduites d'alignement globaux de famille de proteines très proches (Cytochromes, hémoglobines...)

Alignement global

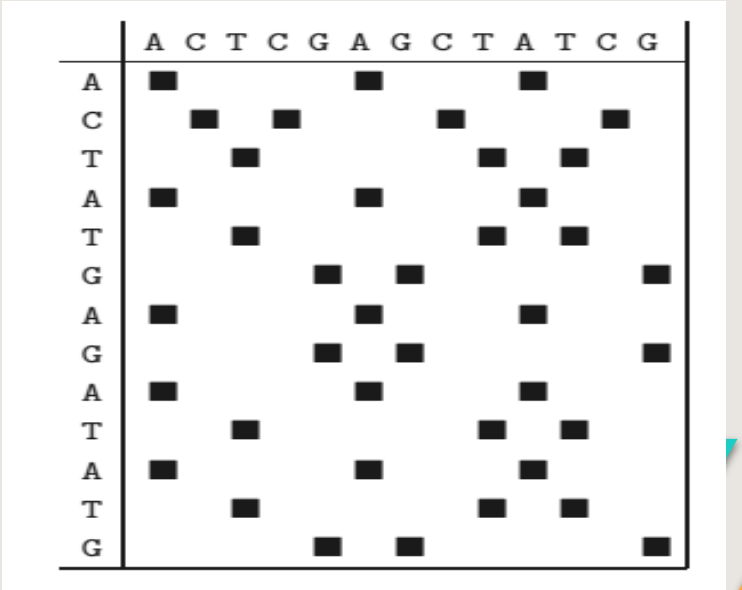
Alignement de 2 séquences sur la totalité de leur longueur en tenant compte de tous les résidus. Si les longueurs de séquences sont différentes, des insertions/ délétions sont introduites pour aligner les deux extrémités des 2 séquences.
Permet de mesurer le degré de similitude de deux séquences connues.

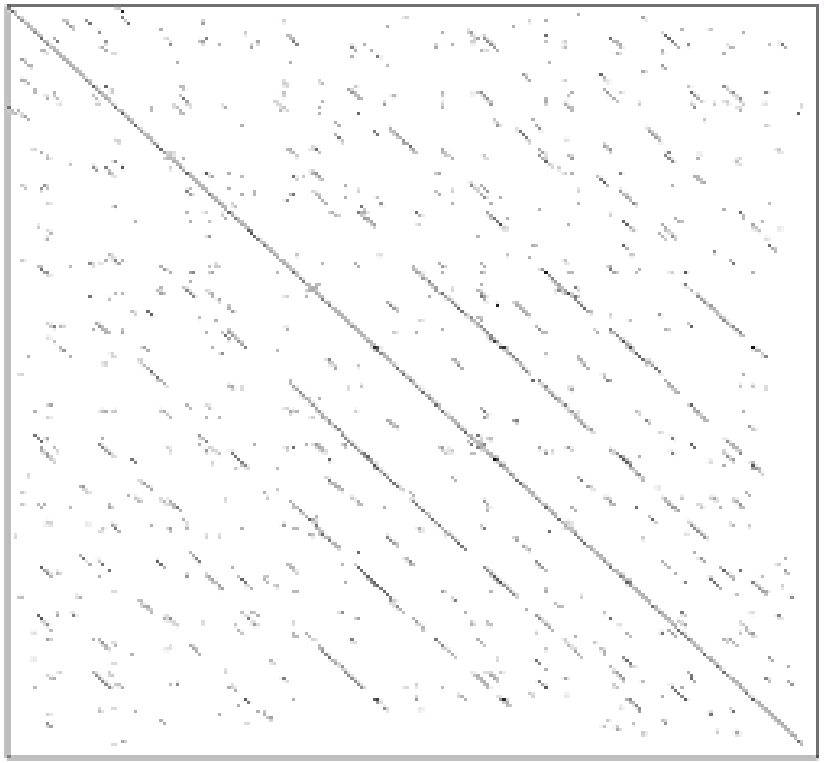


Exp: Le Dot Plot

un outil graphique pour la comparaison, dont le principe est de :
mettre les séquences le long des axes d'une matrice
mettre un point là où il y a un match
une diagonale (une suite de points en diagonale) ⇒ une région similaire.

match (identité) → ■
mismatch → □





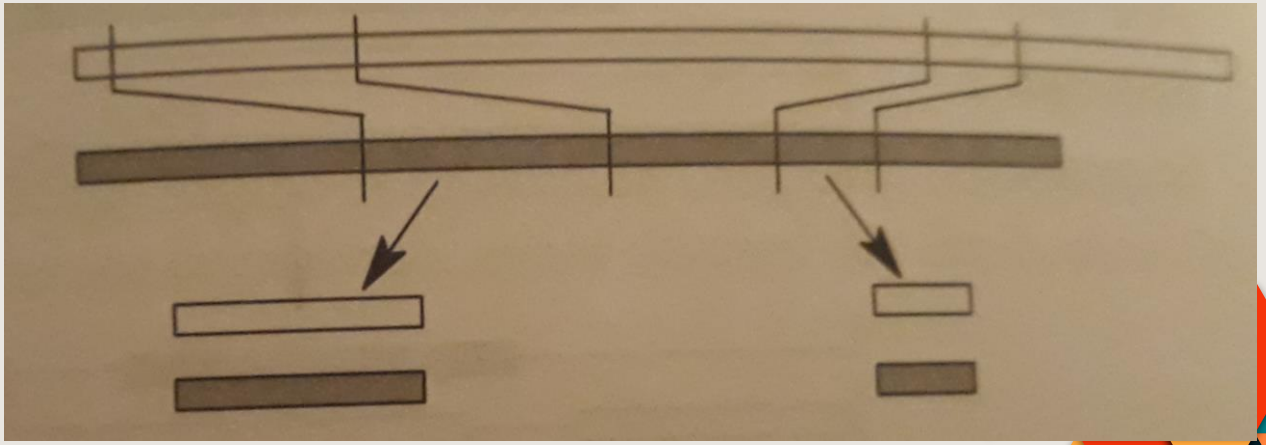
protéine ribosomale S1 de *Escherichia Coli* sur elle-même

Alignement local

Alignement de 2 séquences portant sur des régions isolées et permettant de trouver des segments qui ont un haut degré de similitude. Cette propriété en fait un outils idéal, rapide et efficace, de recherche dans les bases de données en comparant une séquence inconnue avec les séquences de la banque.

Séquence 1:

Séquence 2:

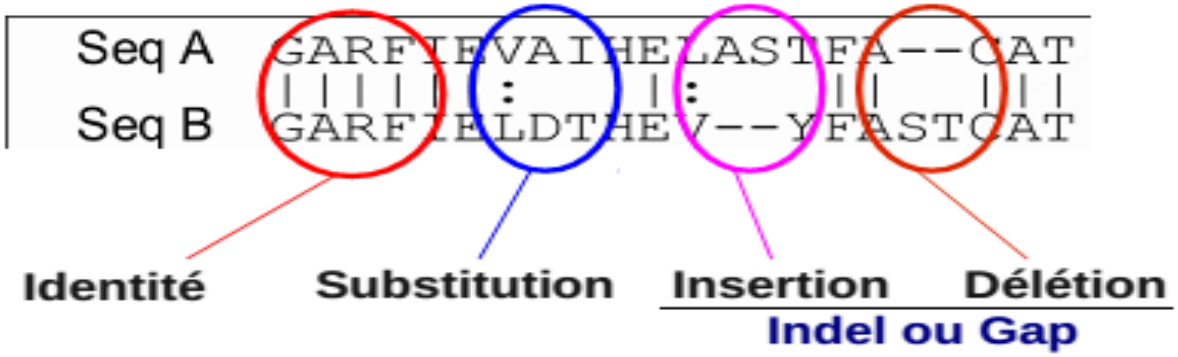


Exp: BLAST (Basic local alignement search tool, altschul 1990)

Spécialement développé pour confronter une séquence inconnu à une banque de séquences et trouver les alignements locaux statistiquement significatifs. L'idée de base exploitée par l'algorithme est que les bons alignements doivent contenir quelque part des petites régions très riches en identité. ce repérage initial permet de sélectionner rapidement les séquences de la banque potentiellement similaires à la séquence requête.

Qu'est ce qu'un alignement

Trois Situations sont possibles pour une position donnée de l'alignement:
1.Les caractères sont les mêmes: **Identité**
2.Les caractères ne sont pas les mêmes: **Substitution**
3.L'une des position est un espace: **Insertion/ délétion**



Identité ou Match (| ou * ou C)
Substitution non conservative ou Mismatch (néant)
Substitution conservative (+ ou : ou .)
Indel ou Gap (néant, - ou .)