Université Batna 2 **Année universitaire : 2020-2021**

Faculté des Sciences de la nature et de la vie

Département d'Ecologie et environnement

Résumé

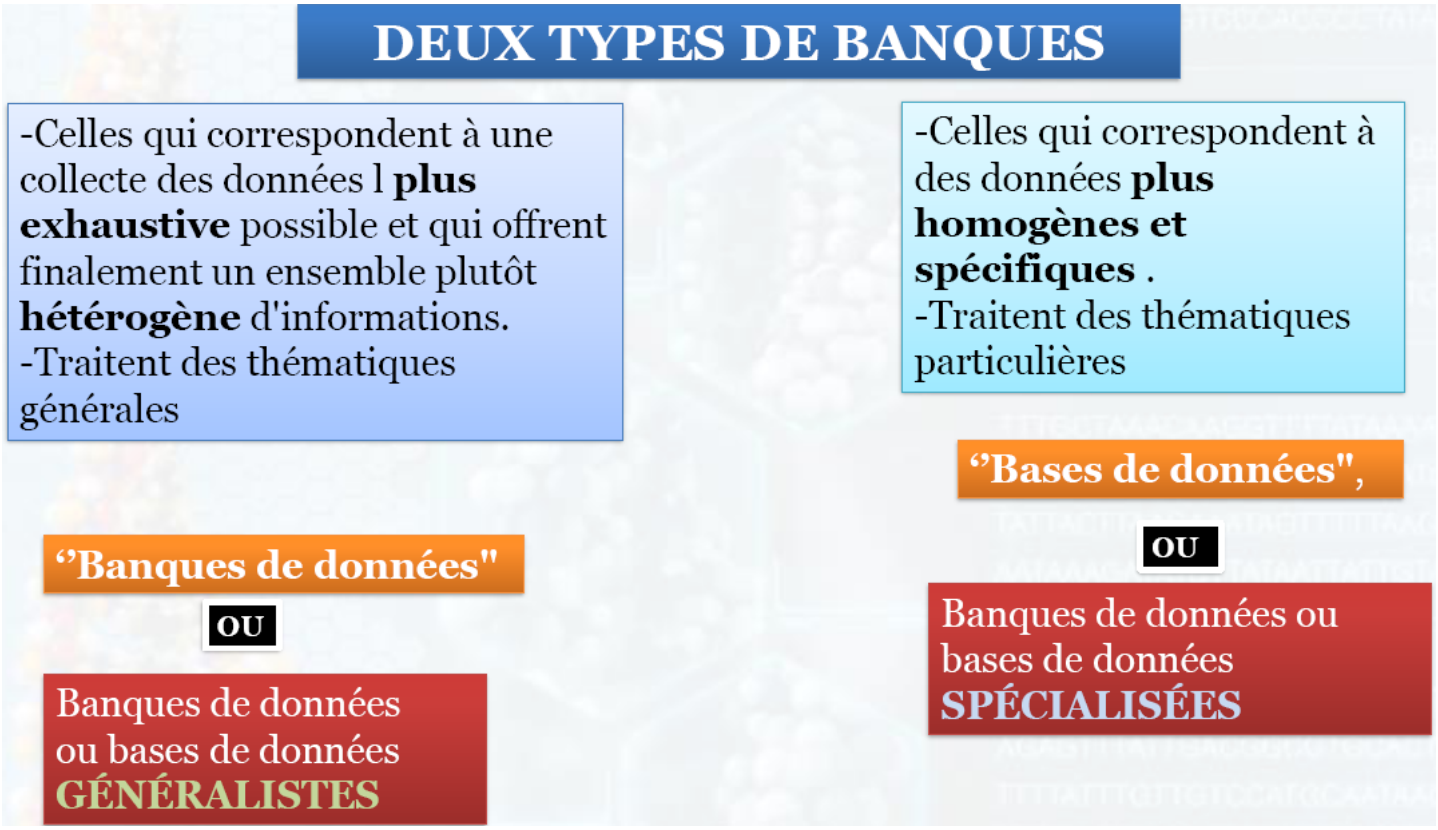
BIOINFORMATIQUE DÉFINITION :

La bioinformatique est la discipline de l'analyse « ***in silico***[1] » de l'information biologique renfermée dans les séquences nucléotidiques (séquences de nucléotides) et protéiques (séquence des acides aminés).

La bioinformatique propose des méthodes et des logiciels qui permettent :

- La collection, le stockage et la gestion des données biologiques et leur distribution à travers les réseaux.
-Le développement des (logiciels/algorithmes) pour analyser les problèmes de biologie moléculaire.
- L'analyse, la comparaison et la prédiction de la structure des gènes.
- La modélisation et la prédiction de la structure et de la fonction des protéines.
-Les études phylogénétiques et l'évolution moléculaire des êtres vivants.

Chapitre I. LES BANQUE DE DONNÉES BIOLOGIQUES



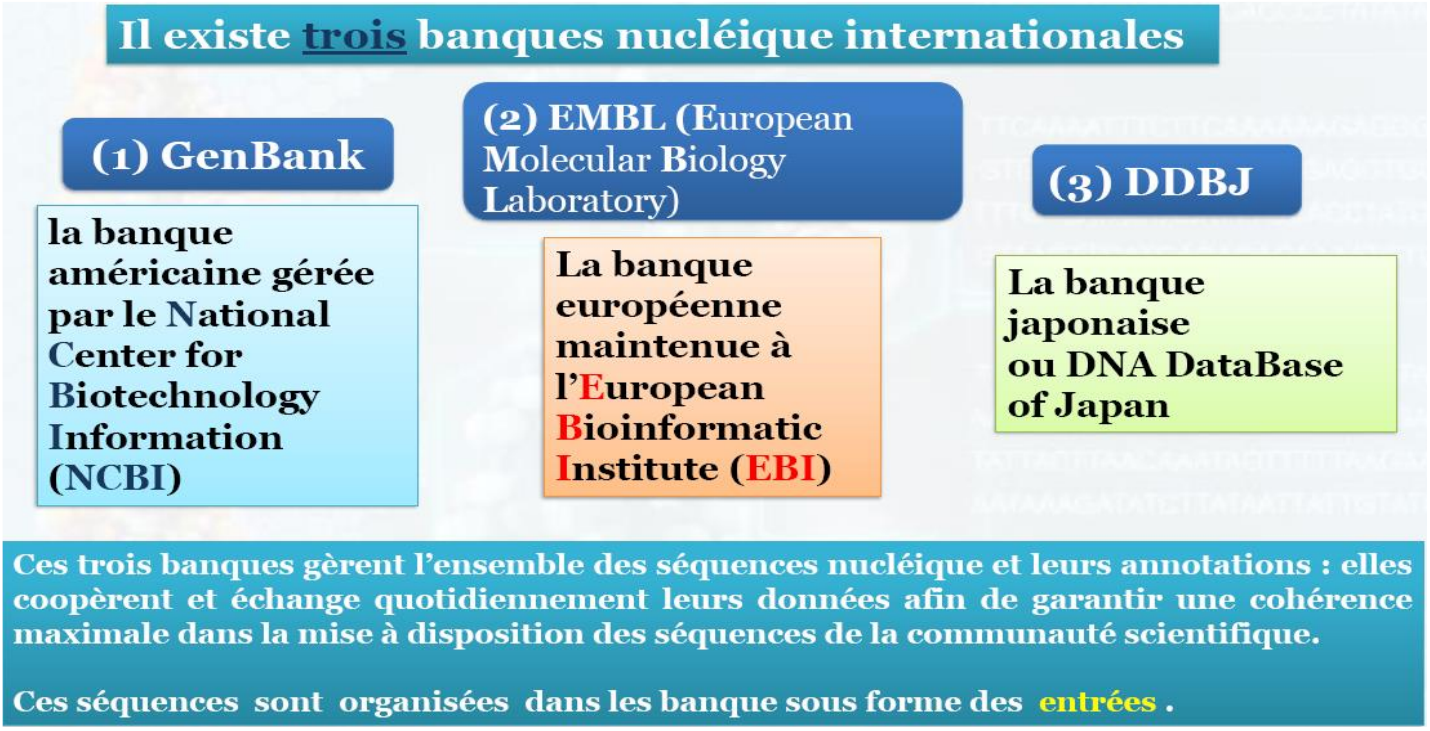
---

[1] ***in silico : se réfère à l'outil informatique. Lorsqu'on dit in silico cela veut dire l'utilisation des processeurs, logiciels informatiques pour gérer, traiter et analyser l'information biologique contenu essentiellement dans les séquences nucléiques et protéiques.***

I. Les Banques Généralistes :

Les banques généralistes sont indispensables à la communauté scientifique car elles regroupent des données et des résultats essentiels dont certains ne sont plus reproduits dans la littérature scientifique.

I. 1. Banque Nucléiques



Ces trois banques (GenBank, DDBJ, EMBL) sont interconnectées[2] (inter-reliées) du fait qu'elles échangent leurs informations. Il suffit de consulter le contenu d'une de ces 3 banques pour accéder au contenu de ces 3 banques en même temps.

I. 2. Banque Protéiques (exemples)

Swissprot & TrEMBL[3] : Elle a été constituée à l'Université de Genève à partir de 1986. Elle est maintenant développée par le SIB (Swiss Institute of Bioinformatics) et l'EBI. Elle regroupe (entre autres) des séquences annotées de la PIR-NBRF ainsi que les séquences codantes traduites de l'EMBL (TrEMBL[4]).

UniProt ("Universal Protein Resource") : c'est la base de données des protéines

[2] **Ces banques s'échangent systématiquement leur contenu depuis 1987 et adoptent un système de conventions communes (The DDBJ/EMBL/GenBank Feature Table Definition).**

[3] **SwissProt et TrEMBL sont toutes les deux des banques généralistes contenant des séquences protéiques. La différence réside dans le fait que les données introduites dans la banque de données SwissProt sont manuellement expertisées avec des ajouts de commentaire décrivant la fonction de la protéine, sa localisation cellulaire etc.., et des annotation dans la partie feature de certaines caractéristiques comme la présence de fragments transmembranaires, de motifs, de domaines fonctionnels. Ces annotations peuvent être extraites de publications ou obtenu à partir d'analyses réalisées par les annotateurs.**
TrEMBL contient les séquences protéiques obtenues par traduction automatique des CDS (régions codantes) des données présentes dans EMBL.

[4] **Attention** : il faut distinguer entre EMBL et TrEMBL : EMBL est la banque de données européenne généraliste de séquences d'acides nucléiques maintenue à l'EBI. Les banques généralistes d'acides nucléiques contiennent toutes les séquences d'acides nucléiques produites dans les laboratoires publiques. TrEMBL est elle aussi une banque de données généraliste mais elle contient des séquences protéiques. Elle est construite par traduction automatique de toutes les CDS de la banque EMBL. Les CDS (CoDing Sequence) correspondent aux régions codantes des gènes (du codon initiateur au codon stop).

II. Les banques spécialisées : elles regroupent des données plus homogènes établies autour d'une thématique ou d'une méthode spécifique de production des données.

Exemples de banques spécialisées : La base de données KEEG pathway (voies métaboliques), Flybase, Prosite (domaines des protéines), Pfam (proteins familly), TRANSFAC, SWISS 2D PAGE, ……

Exemple : bases spécialisée pour un génome spécifique, bases de séquences immunologiques, de voies métaboliques, de cartes génétiques, de motifs protéiques, d'expression de gènes, de structures, . . .

Exemples de formats liés aux logiciels de traitement des séquences

1. Format FASTA

Sans doute le plus répandu et l'un des plus pratiques car très simple. La séquence, sous forme de lignes de 80 caractères maximum, est précédée d'une ligne de titre (nom, définition ...) qui doit commencer par le caractère ">".
Plusieurs séquences peuvent être mises dans un même fichier.

Avec le format FASTA, un seul fichier peut contenir plusieurs enregistrements (séquences). Chaque enregistrement commence par ">".

```
>J00265.1 HUMINS01 Human insulin gene, complete Cds
CTCGAGGGGCCTAGACATTGCCCTCCAGAGAGAGCACCCAACACCCTCCAGGCTTGACCGGCCAGGGT
GTCCCCTTCCTACCTTGGAGAGAGCAGCCCCAGGGCATCCTGCAGGGGGTGCTGGGACACCAGCTGGC
CTTCAAGGTCTCTGCCTCCCTCCAGCCACCCCACTACACGCTGCTGGGATCCTGGATCTCAGCTCCCT
GGCCGACAACACTGGCAAACTCCTACTCATCCACGAAGGCCCTCCTGGGCATGGTGGTCCTTCCCAGC
CTGGCAGTCTGTTCCTCACACACCTTGTTAGTGCCCAGCCCCTGAGGTTGCAGCTGGGGGTGTCTCTG
AAGGGCTGTGAGCCCCCAGGAAGCCCTGGGGAAGTGCCTGCCTTGCCTCCCCCCGGCCCTGCCAGCGC
CTGGCTCTGCCCTCCTACCTGGGCTCCCCCCATCCAGCCTCCCTCCCTACACACTCCTCTCAAGGAGG
CACCCATGTCCTCTCCAGCTGCCGGGCCTCAGAGCACTGTGGCGTCCTGGGGCAGCCACCGCATGTCC
TGCTGTGGCATGGCTCAGGGTGGAAAGGGCGGAAGGGAGGGGTCCTGCAGATAGCTGGTGCCCACTAC
CAAACCCGCTCGGGGCAGGAGAGCCAAAGGCTGGGTGTGTGCAGAGCGGCCCCGAGAGGTTCCGAGGC
TGAGGCCAGGGTGGGACATAGGGATGCGAGGGGCCGGGGCACAGGATACTCCAACCTGCCTGCCCCCA
TGGTCTCATCCTCCTGCTTCTGGGACCTCCTGATCCTGCCCCTGGTGCTAAGAGGCAGGTAAGGGGCT
GCAGGCAGCAGGGCTCGGAGCCCATGCCCCCTCACCATGGGTCAGGCTGGACCTCCAGGTGCCTGTTC
TGGGGAGCTGGGAGGGCCGGAGGGGTGTACCCCAGGGGCTCAGCCCAGATGACACTATGGGGGTGATG
GTGTCATGGGACCTGGCCAGGAGAGGGG
```

Chapitre 2 : Alignement de s équences biologiques et matrices de comparaison

Alignement de sequences d'ADN (ou d'acides amines) :
operation de base en bio-informatique qui a pour but d'identifier des zones conservees entre sequences.

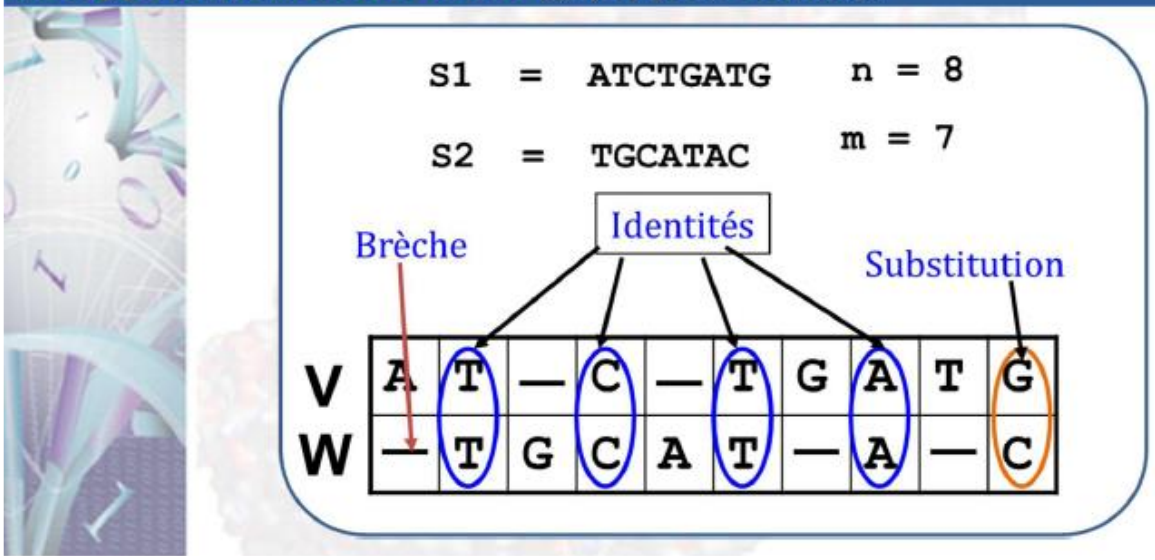
```
CAGCA-CTTGGATTCT-GG
CAGC---TTG--TACTCGG
```

- Utilité de l'alignement :
  - identifier des sites fonctionnels
  - prédire la ou les fonctions d'une protéine
  - prédire la structure secondaire (voire tertiaire ou quaternaire) d'une protéine
  - établir une phylogénie (évolution: parenté entre les organismes)

- 3 événements mutationnels élémentaires
  - substitution — AGACT → AGATT
  - insertion } indel — AGACT → AGACAT
  - délétion } indel — AGACT → AGAT
- Score d'une opération
  - substitution : score de similarité
  - indel : pénalité
- Le score de l'alignement est la somme des scores élémentaires

Principe de l'alignement et détermination du score

Aligner deux séquence, c'est rechercher le maximum d'appariement entre les lettre qui les composent (nucléotides ou résidus d'acides aminés) avec le minimum de mésappariement et des brèches (gaps) (voir schéma).



Brèche = Gap[5] = indel ; Identités = Match = même résidu ; Substitution = Mismatch

Mesure du degré de similitude: La plupart des méthodes d'alignement de séquences biologiques, et en particulier les méthodes d'alignement de séquence de protéines cherchent à optimiser un score d'alignement. Ce score est relié au taux de similarité entre les deux séquences comparées.

- Exemple (Mismatch: -1, Match: 3, Indel: -2) :

A	C	C	G	A	T	G	A	
A	C	–	G	C	T	-	A	
3	+3	-2	+3	-1	+3	-2	+3	= 10

[5] **brèches ou "gap" : espace artificiel introduit dans une séquence pour contre-balancer et matérialiser une insertion dans une autre séquence. Il permet d'optimiser l'alignement entre les séquences.**

- Exemple (Mismatch: -1, Match: 2, Indel: -4) :

A	G	T	T	G	T	T	C	
T	G	–	G	G	T	A	C	
-1	+2	-4	-1	+2	+2	-1	+2	= 1

Un alignement sera considéré comme bon s'il fait correspondre un nombre élevé d'identités, et un nombre minimal d'insertions, de délitions et de substitutions.

Ceci conduit naturellement a l'idée **d'évaluer la qualité d'un alignement** en lui attribuant une note :
une prime à l'alignement pour chaque identité
une pénalité pour chaque opération de modification (substitutions et breches).

La notation de l'alignement **(score total)** peut ainsi être calculée en sommant les primes d'identité et les pénalités des breches (d'insertions/délétions) et substitution effectuées.

Les pénalités des brèches doivent être suffisamment coûteuses pour éviter les alignements sans signification biologique.

Le coût d'extension d'une brèche déjà ouverte est généralement plus faible par rapport à celui de son ouverture.

la recherche de similitude entre séquences nécessite la détermination d'un score de similarité

$$Score_{Total} = \Sigma Score_{élémentaires} - \Sigma Score_{pénalités}$$

Exemple de détermination de score avec la matrice unitaire (l'appariement vaut +1, le mésappariement vaut 0 et une brèche vaut -1)

Alignement sans brèches	Alignement avec brèches
Séquence 1 ATGACTGGGCCACT Séquence 2 ATACTGGGACAACT	
Séquence 1 ATGACTGGGCCACT Séquence 2 ATACTGGGACAACT	Séquence 1 ATGACTGGGCC-ACT Séquence 2 AT-ACTGGGACAACT
8 appariements (match) et 6 mésappariement (mismatch). Score 8 - 0= 8	12 appariements, 1 mésappariement et 2 brèches. Score 12 - 2= 10

Matrices de substitution

- Matrices nucléiques
  - Il existe peu de matrices pour les acides nucléiques car il n'y a que 5 lettres pour leur alphabet
  - La plus fréquemment utilisée est la matrice dite unitaire (ou matrice identité) où toutes les bases sont considérées comme équivalentes

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

Match : 1
Mismatch : 0
Indel : 0

- Matrices des acides aminés : beaucoup plus complexe !
  - Pam [1978], Blosum [1992], Gonnet [1992]…
  - Basées sur: nombres de mutation nécessaires pour changer d'acide aminé, propriétés physico-chimiques, évolution…

Matrices de score

Matrice nucléique

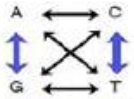
	A	C	G	T
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1

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 trois score

La matrice à trois scores distingue :

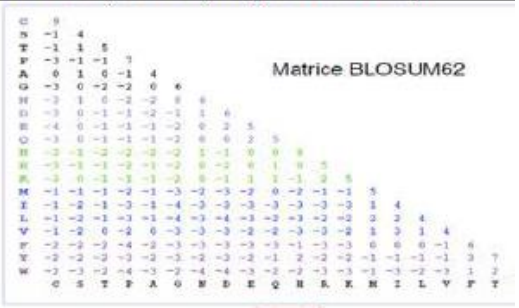


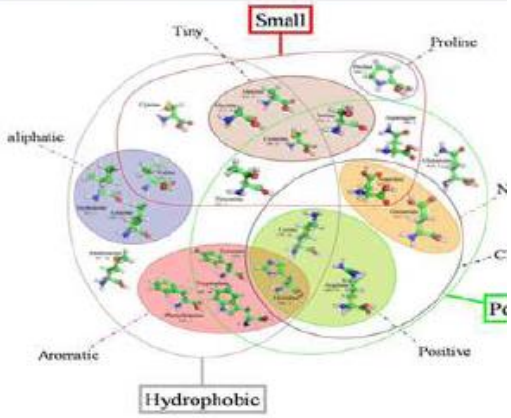
P(transition) > *P*(transversion)

la matrice à trois scores distingue :

-les transitions (A ↔G et C↔T)
-les transversions (A↔C, G↔T, T↔A et C↔G) : Modification d'une base purique par une base pyrimidique et inversement.

Matrice protéique (BLOSUM62)





Transition = purine ↔ purine ; pyrimidine ↔ pyrimidine

Transversion = purine ↔ pyrimidine et inversement

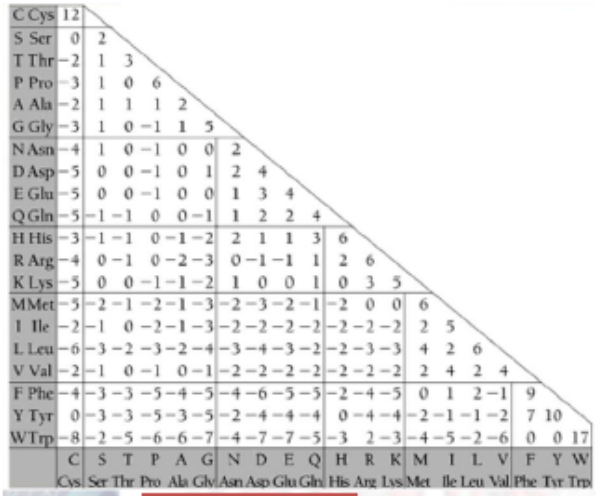
- En Bioinformatique la transition est beaucoup plus favorable que la transversion parce que on reste toujours dans la même famille et/ou les propriétés physicochimiques vont rester les mêmes

- En Bioinformatique : on favorise Match (le faite que le résidu soit reste le même), c'est pour cette raison les valeurs match vont avoir les scores les plus élevés, alors que les valeurs mismatch sont plus ou moins neutres et on fait le choix de pénaliser fortement les événements insertions/délétions.

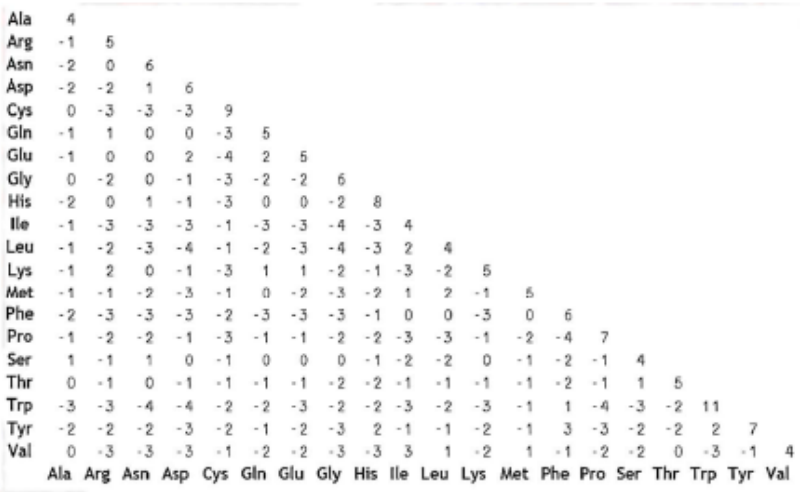
Les matrices **BLOSUM** (de Steve **Henikoff 1950**) (**BLO**cks **SU**bstitution **M**atrix) sont déduites d'alignements de fragments (Blocks) de **protéines très éloignées**.
Par exemple: BLOSUM62 est déduite à partir d'un alignement de séquences ayant 62% de similitude. Ces matrices sont bien adaptées aux recherches de séquences dans les banques de données (Blast, FASTA).
Chaque score donne le coût de remplacement d'un résidu par un autre. On note que :

-Les acides aminés rares ont un score élevés (Trp, Cys, His)
-les acides aminés communs ont des scores faibles (Ala, Leu, Ile,.....)
-les substitutions conservative entre acides aminés similaire sont peu pénalisantes. Ces substitutions peuvent se produire sans affecter l'activités de la protéine (ex : Lys↔Arg) .

Autre matrice : **PAM** (de Margaret Oakley **Dayhoff 1925-1983**) (**P**oint **A**ccepted **M**utation) déduites d'alignements globaux de famille de protéines très proches (exemple : cytochromes, hémoglobines .....)



PAM 250



Blusom 62

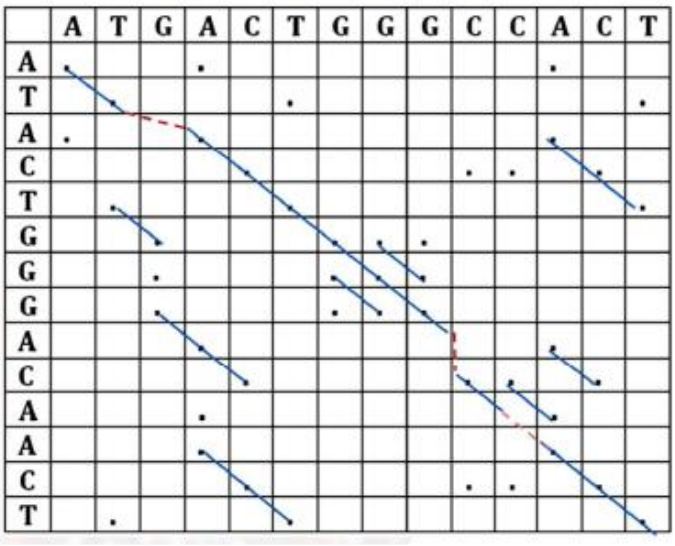
Remarque :

Pour les matrices PAMs : le numéro indique le nombre de mutations acceptés dans chaque 100 acide aminé par exemple : PAM 1 signifie qu'une seule mutation est acceptée dans chaque 100 acide aminé.

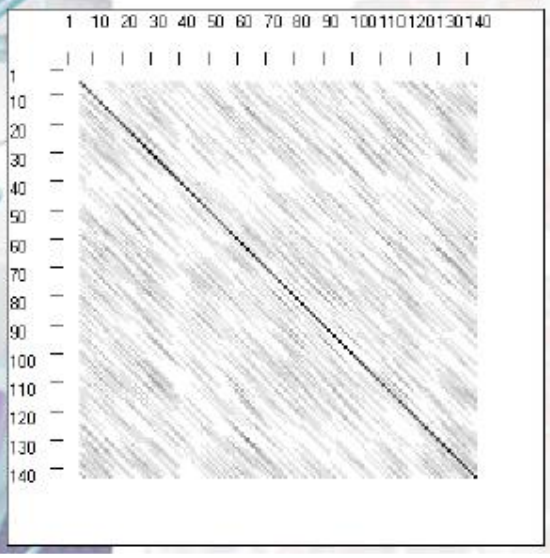
Pour les matrices BLOSUMs : le chiffre indique le pourcentage d'identité par exemple pour la matrice BLOSUM62 : 62 signifie qu'on garde uniquement les BLOKs ayant un pourcentage d'identité (%) supérieur à 62%.

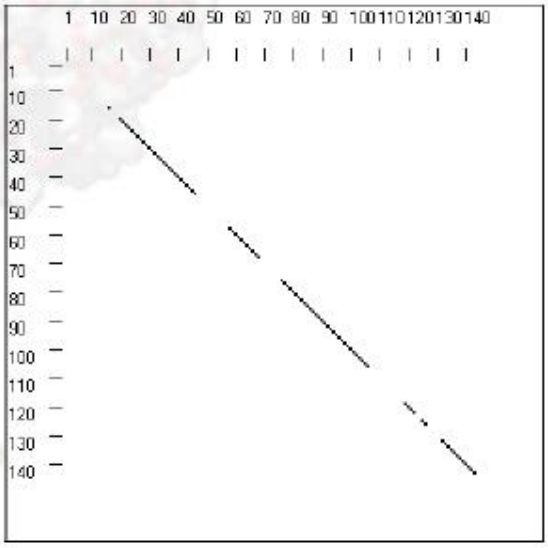
Dot-plot

Principe : outil visuel de comparaison de séquences. Les séquences sont positionnées perpendiculairement dans un tableau et on met un point à chaque appariement. La multiplicité des points forme des diagonales. Les décalages correspondent des insertions / délétions et les segments parallèles indiquent des répétitions. Si on reprend l'exemple précédent cela donne le tableau suivant :



Avantage et inconvénients : simple et intuitif. Mais des problèmes de bruit de fond se posent pour les longue séquences.
Cela nécessite l'utilisation d'un filtrage : on ne met un point que si n caractères sont identiques dans une fenêtre donnée, pour éliminer les segments de similitudes courtes.





- La différence entre la matrice Dot plot et les matrices de substitutions (matrices de scores) :

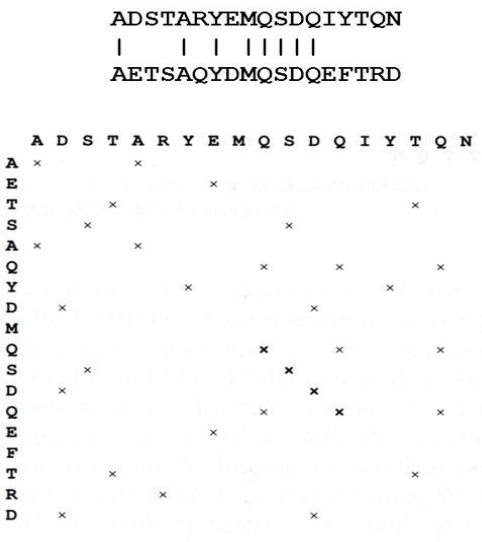
Dot plot = basée sur l'identité absolue (méthode trop sévère qui n'accepte aucun possibilité de mutation)

Matrices de substitutions = vont indiquer la possibilité qu'un résidu (qu'un acide aminé ; une base azotée) soit remplacé par un autre ; ce qui fait : certains changements/mutations vont être acceptable (similarité[6]).

Annexe :

Matrice de points (Matrice de pixels; dot-plot)

- Le dot plot est une représentation graphique simple des résidus identiques entre les deux séquences.
  - Les deux séquences sont représentées sur les deux axes
  - Un point (dot) est tracé pour chaque correspondance entre deux résidus de séquences.
  - Les lignes diagonales révèlent les régions alignables entre les deux séquences.



- Diapo: Emese Meglecz

[6] **Similarité : c'est le pourcentage d'identités et/ou de substitutions conservatives entre des séquences. Le degré de similarité est quantifié par un score. Le résultat de la recherche d'une similarité peut être utilisé pour inférer l'homologie de séquences. Homologie : 2 séquences sont homologues si elles ont un ancêtre commun. L'homologie se mesure par la similarité : une similarité significative est signe d'homologie**

Banques de séquences proteiques:

Swiss Prot:
- Niveau élevé d'Annotation (manuelle):
-Description de la fonction des proteines, structure des domaines et modification post-traductionnelle.... etc

TrEMBL :
Données générées par traduction automatique des informations génétiques de la banque de données EMBL (d'où TrEMBL =Traduction EMBL)
Annotation automatique

Prosite:
Base de données de familles et domaines de protéines

GenePept:
Traduction automatique des CDS de GenBank

PIR (Protein Information Ressource):
Groupe établit par le National Biomedical Research Foundation (NBRF)
Identification et interpretation de l'information des séquences proteiques

Expazy:
Base de données protéomique

En 2002: consortium UniProt (Universal Protein Resource) formé par le groupe SwissProt–TrEMBL et le groupe PIR

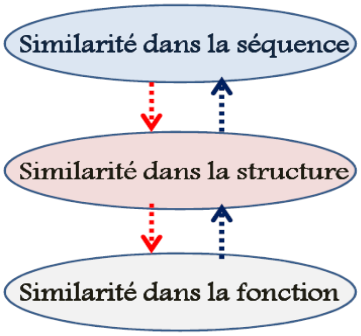
8

Autres types de banques de données:

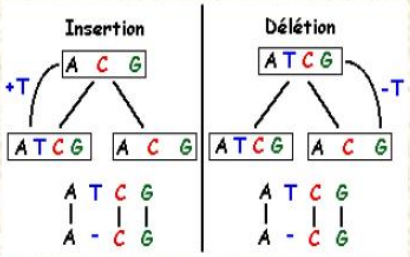
Banques de Structure:
Ex: la Protein Database PDB dédiée au structures proteiques déterminées experimentalement

Banques dédiées à un organisme particulier:
Ex: *Arabidopsis thaliana* (TAIR, ABRC….)
Colibri (*E. coli*)
Subtilis (*Bacillus subtilis*)
Flybase (Drosophile)

Analyser ma séquence pour ?



indel :
- "in" = insertion
- "del" = délétion



similarité : c'est le pourcentage d'identités et/ou de substitutions conservatives entre des séquences. Le degré de similarité est quantifié par un score. Le résultat de la recherche d'une similarité peut être utilisé pour inférer l'homologie de séquences.

homologie : 2 séquences sont homologues si elles ont un ancêtre commun. L'homologie se mesure par la similarité : une similarité significative est signe d'homologie sauf si les séquences présentent une faible complexité.