Université Batna 2
Faculté des sciences de la nature et de la vie
Département de biologie des organismes

TD de
Génomique et protéomique fonctionnelle
M2BMC

Dr. LAANANI.I

2022-2023

Table des matières

TD 01 : Portails, bases de données et bioinformatique

L'objectif de ce TD est de faire le point sur les apports de la bioinformatique notamment par les différentes **bases de données** et outils **bioinformatiques** qu'elle a permis de créer ces dernières années et qui sont aujourd'hui autant d'outils incontournables pour les généticiens.

1. La bioinformatique:

La bioinformatique correspond à l'utilisation des outils **informatiques** pour **stocker** et **analyser** les données de la biologie afin de résoudre les problèmes scientifiques posés par la biologie dans son ensemble. Il s'agit dans tous les cas d'un champ de recherche multidisciplinaire qui associe des informaticiens, mathématiciens, physiciens et biologistes.

Il existe trois principaux outils de bioinformatique. On distingue les **portails**, les **logiciels** (applications) et les **bases de données** biologiques.

1.1. Les portails en bioinformatique:

Ce sont des serveurs ou centres de ressources qui proposent des applications (Tools & Software) et l'accès à des **banques de données**. Ils permettent ainsi d'accéder à des informations dans **les bases de données** et d'effectuer en ligne des calculs, des analyses et des comparaisons.

Exemples de portails en bioinformatique

-NCBI : National Center for Biotechnology Information (NIH, USA),

(https://www.ncbi.nlm.nih.gov/)

-EBI: European Bioinformatics Institute (EMBL, GB), (https://www.ebi.ac.uk/**)**

-Expasy: Expert Protein Analysis System (Swiss Institute of Bioinformatics, Suisse), (https://www.expasy.org/)

-HGMP: Human Genome Mapping Project resource centre (Cambridge, GB),

(https://gtr.ukri.org/organisation/9D890F57-EC84-48A7-80E5-C59D42839D6B)

1.2. Les bases de données :

Se trouvent dans les portails et permettent le partage des connaissances à travers la communauté scientifique. Elles contiennent des informations venant de divers champs de recherche tels que la génomique, la protéomique, la métabolomique, la phylogénétique et les puces à ADN. Parmi le contenu des bases de données, on trouve des informations à propos de la fonction, de la structure, de la localisation (cellulaire et chromosomique) des gènes et les effets cliniques de leurs mutations, ainsi que leurs similarités de séquence et de structure.

1. 2.1. Banques nucléiques : Il existe trois banques nucléiques internationales :

- **GenBank** (https://www.ncbi.nlm.nih.gov/genbank/**)** : la banque américaine gérée par le National Centre for Biotechnology Information(NCBI)
- **EMBL** (European Molecular Biology Laboratory databank, https://www.embl.org/), la banque européenne maintenue à l'European Bioinformatics Institute (EBI) ;
- **DDBJ** (DNA Database of Japan, https://www.ddbj.nig.ac.jp/index-e.html), la banque japonaise.

Ces trois banques gèrent l'ensemble des séquences nucléiques et leurs **annotations**, elles coopérant et échange quotidiennement leurs données afin de garantir une cohérence maximale dans la mise à disposition des séquences de la communauté scientifique, même si chacune de ces banques présente quelques petites spécificités mais la structuration des données y est semblable et leur contenu en séquences nucléiques est strictement identique.

1.2. 2. Banques protéiques :

- **La banque de données européenne Swiss-Prot** (https://www.uniprot.org/) : qui se caractérise par une excellente qualité **d'annotation** des données grâce à la contribution d'experts au détriment de l'exhaustivité (séquences protéiques **annotées manuellement** par des biologistes).
- **La banque TrEMBL**: qui contient l'ensemble des séquences protéiques conceptuelles obtenues par traduction automatique des séquences codantes contenues dans **EMBL**.

1. 2.3. Les bases de données spécialisées de génomes complets :

- Reference Sequence **(RefSeq)** du **NCBI**, est l'une des ressources les plus anciennes dédiées aux génomes complets procaryotes et eucaryotes.

Ressources pour les procaryotes :

-Pour les procaryotes les deux bases de données de génomes complets les plus couramment utilisée sont la section « **Microbial Genomes** » de la base **RefSeq** du **NCBI (**https://www.ncbi.nlm.nih.gov/genome**),** et la partie « procaryotes » de la base **Ensemble Genomes** (https://www.ensembl.org/).

Ressources pour les animaux :

-Une des principales ressources de données pour les génomes des eucaryotes supérieures est le « **projet Ensemble** », issu d'une collaboration entre l'EBI et le Sanger Institute et dédié à l'annotation automatique des génomes de métazoaires, dont l'objectif d'annoter et comparer les grandes séquences chromosomiques à partir l'ensemble des données disponibles.

-La base **FlyBase** (*https://flybase.org/)***:** pour l'annotation et l'analyse fonctionnelle des génomes de Drosophila;

-La base **WormBase** (https://wormbase.org//#012-34-5)**:** pour intégrer les informations disponibles sur le nématode ;

- La base ***Caernohabditis elegans*** **DataBase (AceDB,** https://www.sanger.ac.uk/tool/acedb/) : pour la gestion et l'annotation du génome du nématode modèle *C. elegans* et maintenant applicable à tout autre organisme procaryote ou eucaryote.

Ressources pour les plantes :

-La base The *Arabidopsis* Infomation Resource (TAIR, https://www.arabidopsis.org/) qui centralise la plupart des informations disponibles sur *Arabidopsis.*

Ressources pour les champignons:

-*Saccharomyces* Genome Database (SGD, https://www.yeastgenome.org/): est une base centrée sur la biologie moléculaire et la génétique de la levure de boulanger *Saccharomyces cerevisiae*.

1.2.4. Les bases de données centrales dédiées aux SNPs

- **dbSNP** *(https://www.ncbi.nlm.nih.gov/snp/* qui se trouve dans le portail *NCBI)*
- **Allele FREquency Database** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4264510/ qui se trouve dans le portail *NCBI)*
- **HapMap** *(Projet international https://www.ncbi.nlm.nih.gov/probe/docs/projhapmap/)*

1.3. Les formats de séquence :

1.3.1. Le format **BRUT** qui est la séquence seule.

Exemple :

cggcgccgcgagcttctcctctcctcacgaccgaggcagagcagt
attatggcgaaccttggctgctggatgctggttctctttgtggccaca
ggagtgacctgggcctctgcaagaagcgcccgaagcctggagga
ggaacactgggggcagccgatacccggggcagggcagcccgttc
tgttttgttatataaaaaaattgtaaatgtttaatatctgactgaaatt
aacgagcgaagatgagcacc

1.3.2. Le format FASTA :

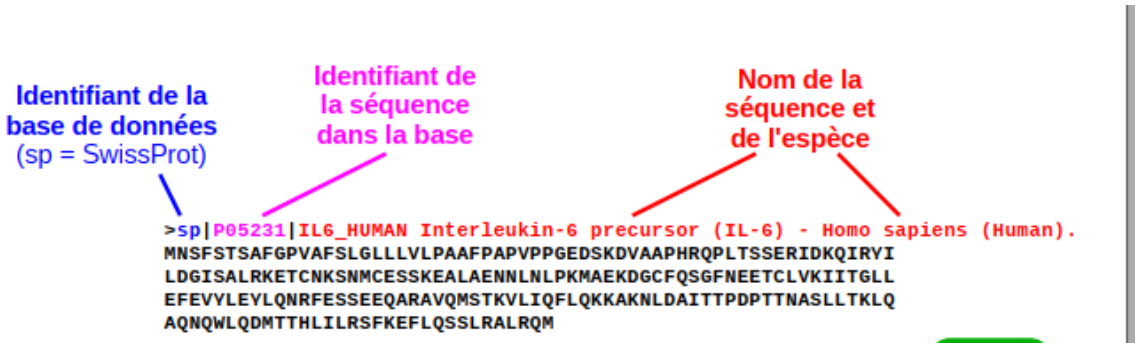
C'est le **format standard** le plus couramment utilisé. Un fichier FASTA contient une ou plusieurs séquences, soit de **nucléotides** ou **d'acides aminés**. Chaque séquence est précédée d'une ligne débutant par le symbole > suivi d'un entête contenant normalement le nom de la séquence et les informations complémentaires qu'on veut y ajouter.

Example:

1. **Gène:**

```
>HSGLTH1 Human theta 1-globin gene
CCACTGCACTCACCGCACCCGGCCAATTTTTGTGTTTTTAGTAGAGACTAAATACCATATAGTGAACACCTAAGA
CGGGGGGCCTTGGATCCAGGGCGATTCAGAGGGCCCCGGTCGGAGCTGTCGGAGATTGAGCGCGCGCGGTCCCGG
GATCTCCGACGAGGCCCTGGACCCCCGGGCGGCGAAGCTGCGGCGCGGCGCCCCCTGGAGGCCGCGGGACCCCTG
GCCGGTCCGCGCAGGCGCAGCGGGGTCGCAGGGCGCGGCGGGTTCCAGCGCGGGGATGGCGCTGTCCGCGGAGGA
CCGGGCGCTGGTGCGCGCCCTGTGGAAGAAGCTGGGCAGCAACGTCGGCGTCTACACGACAGAGGCCCTGGAAAG
GTGCGGCAGGCTGGGCGCCCCCGCCCCCAGGGGCCCTCCCTCCCCAAGCCCCCCGGACGCGCCTCACCCACGTTC
CTCTCGCAGGACCTTCCTGGCTTTCCCCGCCACGAAGACCTACTTCTCCCACCTGGACCTGAGCCCCGGCTCCTC
ACAAGTCAGAGCCCACGGCCAGAAGGTGGCGGACGCGCTGAGCCTCGCCGTGGAGCGCCTGGACGACCTACCCCA
CGCGCTGTCCGCGCTGAGCCACCTGCACGCGTGCCAGCTGCGAGTGGACCCGGCCAGCTTCCAGGTGAGCGGCTG
CCGTGCTGGGCCCCTGTCCCCGGGAGGGCCCCGGCGGGGTGGGTGCGGGGGGCGTGCGGGGCGGGTGCAGGCGAG
TGAGCCTTGAGCGCTCGCCGCAGCTCCTGGGCCACTGCCTGCTGGTAACCCTCGCCCGGCACTACCCCGGAGACT
TCAGCCCCGCGCTGCAGGCGTCGCTGGACAAGTTCCTGAGCCACGTTATCTCGGCGCTGGTTTCCGAGTACCGCT
GAACTGTGGGTGGGTGGCCGCGGGATCCCCAGGCGACCTTCCCCGTGTTTGAGTAAAGCCTCTCCCAGGAGCAGC
CTTCTTGCCGTGCTCTCTCGAGGTCAGGACGCGAGAGGAAGGCGC
```

2. **Protéine:**



1.4. BLAST:

BLAST « Basic Local Alignment Search Tool », est un programme couramment utilisé pour trouver des régions **d'homologie** entre différentes séquences. Il permet de trouver les régions similaires entre deux ou plusieurs séquences de nucléotides ou d'acides aminés, et de réaliser un alignement de ces régions homologues. On doit normalement donner une séquence d'entrée qui sera comparée à une banque de séquences nucléotidiques ou protéiques. Cette méthode est utilisée pour trouver des **relations fonctionnelles** ou **évolutives** entre les séquences et peut aider à identifier les membres d'une même famille de gènes. La comparaison peut être effectuée sur de grandes banques de séquences disponibles sur internet comme celles retrouvées entre autres sur le site de NCBI. Plusieurs variantes :

- **Blast n :** recherche d'une séquence nucléotidique dans une banque d'ADN
- **Blast p :** recherche d'une séquence protéique dans une banque de protéine