**CHAPITRE II : LES BANQUES DE DONNEES**

1. **Définition**

Les banques de données biologiques sont des bases de données contenant des informations biologiques et des données largement diffusées par le réseau internet et sont généralement reliées entre elles par des liens.

1. **Caractéristiques d’une base de donnée est donc:** c’est un ensemble de données ;
* Structuré,
* Indexé,
* Périodiquement mise à jour,
* Accessibles au moyen d’un logiciel,
* Elles comportent souvent des outils associés (logiciels) nécessaires pour :
- l’accession à la Base ;
- la mise à jour de la Base
1. **Rôle des bases de données**

• Collecter les informations ; Stocker et organiser les informations ; Distribuer les informations ; Faciliter l’exploitation des informations.

1. **Classification de bases de données biologiques**•  Bases de données bibliographiques
•  Bases de données de séquences nucléiques ou protéiques : généralistes et spécialisées.





**V.1. Bases de données bibliographiques**

* **Exemple: PubMed ;** est une base de données bibliographiques, développé par le National Center for Biotechnology Information (NCBI) de la National Library of Medicine, centrée sur la documentation en sciences biologiques.

**IV.2. Bases de données de séquences nucléiques ou protéiques**

Les génomes sont des textes de taille gigantesque, déchiffrés par les méthodes de séquençage et d’assemblage qu’il faut ensuite stocker pour les analyser, les classer, les comparer, les réutiliser et finalement les comprendre.

Des informations pertinentes ont été extraites des travaux déjà effectués : gènes, gènes d’une maladie, protéines, motifs, métabolismes, etc.

Ces informations sont regroupées dans diverses bases de données

 **Il existe de nombreuses bases de données biologiques :** **laquelle choisir ?**

* Principaux centres de bioinformatique : NCBI et EMBL
* Il existe essentiellement deux catégories de bases de données:
**IV.2.1. Bases de données généralistes**

Correspondent à une collecte de données la plus exhaustive possible et offrent un ensemble hétérogène d’informations.

• Bases de données généralistes nucléiques
• Bases de données généralistes protéiques

* **Caractéristiques:**
* Très riches
– Grand nombre de séquences accessibles
– Grande diversité des organismes représentés
* Peu/pas de contrôles sur la qualité des entrées
– Les auteurs sont responsables des entrées!
=> Nombreux Problèmes/Erreurs
* Qualité des informations non homogènes
* Redondance (la même sequence peut être representée plusieurs fois)

**IV.2.1.1. Bases de données généralistes nucléiques**

Il existe trois importantes bases généralistes nucléiques. Ces trois bases de données échangent systématiquement leur contenu. Elles contiennent les séquences d’ADN et de protéines publiées dans les journaux/periodiques scientifiques ou soumises par les etablissements/centres de recherche publiques.



**IV.2.1.2. bases de données généralistes protéiques**





**IV.2.2. Base de données spécialisées**

Correspondent à des données plus homogènes établies autour d’une thématique et qui offrent une valeur ajoutée à partir d’une technique particulière ou d’un intérêt suscité par un groupe d’individus.

* **Caractéristiques:**
* Spécifique à un organism,
* Spécifique sur des domaines de protéines,
* Voies de régulations biochimiques,
* Expression de gènes,
* Variation génétique,
* Intéractions Protéine-Protéine.



**Les plus importantes bases de données biologiques**

