CHAPITRE 6 : BASES DE DONNEES

1- Introduction

C'est au début des années 80 que les premières banques de séquences sont apparues sous l'initiative de quelques équipes comme celle du Professeur Grantham à Lyon. Très rapidement avec les évolutions techniques du séquençage, la collecte et la gestion des données ont nécessité une organisation plus conséquente. Ainsi, plusieurs organismes ont pris en charge la production de telles bases de données.

2- Définitions

a- Base de données

Une base de données est un ensemble structuré et organisé permettant le stockage de grandes quantités d'informations afin d'en faciliter leur utilisation (ajout, mise à jour, recherche et éventuellement analyse dans les systèmes les plus évolues). Elles sont toutes organisées en fonction d'un modèle de données (*data model*) qui peut être de différents types :

- modèle hiérarchique (*hierarchical model*),
- modèle en réseau (*network model*),
- modèle relationnel (*relational model*),
- modèle oriente objet (*objectoriented model*),
- modèle semi structure *(semistructured model*),
- modèle associatif (*associative model*),
- modèle EAV (*Entity-Attribute-Value data model*) ou encore
- modèle contextuel (*context model*).

L'un des modèles les plus utilisés aujourd'hui est le modèle de bases de données relationnelles qui a été inventé en 1970 par **Edgar Frank Codd.**

b- Bases de données biologiques

Sont des bibliothèques répertoriant des informations sur les sciences de la vie ; collectées grâce à des expériences scientifiques, à la littérature publiée, aux technologies expérimentales à haut débit, et aux analyses informatiques. 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.

Ces bases de données sont des outils importants pour les scientifiques car elles leur permettent de comprendre et expliquer de nombreux phénomènes biologiques allant de la structure des biomolécules et leurs interactions à l'ensemble du métabolisme des organismes, et même l'évolution des espèces.

c- Système de gestion de base de données

Ensemble de programmesqui permettent l'accès à une base de données : transactions, intégrité, sécurité, administration.

d- Index

Structure de donnéespermettant au SGBD d'accéder de manière efficace au contenu d'une base de données. Attention, un index améliore les performances d'accès en interrogation mais pénalise les mises à jour et prend de la place sur le disque.

e- Banque de données

Les banques de données sont souvent de gros amas d'information en ligne (pas forcément structuré), produits par des institutions. Parfois seulement le stockage de références sur des documents.

3- **Accès**

La plupart des bases de données biologiques sont accessibles sur des sites web sur lesquels les utilisateurs peuvent parcourir les informations. En général, il est également possible de télécharger les données sous divers formats : texte, données de séquençage, structures protéiques ou liens. Par exemple :

- Des informations sous formes de textes peuvent être fournies par **PubMed** ou **OMIM,**
- Des données de séquençage sont disponibles sur **GenBank** (ADN) et **UniProt** (protéines),
- Des structures spatiales protéiques sont disponibles sur **PDB, SCOP et CATH**.

4- Les types de bases de données

Nous distinguerons deux types de banques :

a- Bases de données généralistes

Elles correspondent à une collecte de données la plus exhaustive possible et qui offrent finalement un ensemble plutôt hétérogène d'informations. En Europe, financée par l'**EMBO** (European Moleculary Biology Organisation), une équipe s'est constituée pour développer une banque de séquences nucléiques (**EMBL** data library) en 1986. Cette équipe travaille au

sein du Laboratoire Européen de Biologie Moléculaire qui se trouve près de Cambridge au sein de **l'EBI** (European Bioinformatics Institute). Du côté américain, une banque nucléique nommée **GenBank** a été créée en 1986 ; cette base de données est diffusée maintenant par le **NCBI** (National Center for Biotechnology Information). La collaboration entre ces deux banques a commencé relativement tôt ; elle s'est étendue en 1987 avec la participation de la **DDBJ (**DNA Data Bank of Japan) pour donner naissance finalement en 1990 à un format unique dans la description des caractéristiques biologiques qui accompagnent les séquences dans les banques de données nucléiques **(The DDBJ/EMBL/GenBankfeature).**

Parallèlement, pour les protéines, deux banques principales ont été créées :

- La première, sous l'influence du **NBRF** (National Biomedical Research Foundation) à Washington, produit maintenant une association de données issues du **MIPS** (Martinsried Institute for ProteinSequences), de la base Japonnaise **JIPID** (Japan International Protein Information Database) et des données propres de la NBRF. Elle se nomme la **PIR-NBRF** (Protein Identification Ressource, 1986).
- La deuxième, **Swissprot** a été constituée à l'Université de Genève à partir de 1986 et regroupe entre autres des séquences annotées de la **PIR-NBRF** ainsi que des séquences codantes traduites de l'**EMBL**.

Devant la croissance quasi exponentielle des données et l'hétérogénéité des séquences contenues dans les principales bases de séquences généralistes, d'autres bases spécialisées sont apparues.

b- Bases de données spécialisées

Elles sont constituées autour de thématiques biologiques ou tout simplement en vue de réunir les séquences d'une même espèce et d'en enrichir les annotations pour diminuer ou lever les ambiguïtés laissées par les grandes banques publiques. Elles ont pour but de recenser des familles de séquences autour de caractéristiques biologiques précises comme les gènes identiques issus d'espèces différentes. Elles peuvent aussi regrouper des classes spécifiques de séquences comme les enzymes de restriction, ou toutes les séquences d'un même génome. En fait très souvent ces bases correspondent à des améliorations ou à des regroupements par rapport aux données issues des bases généralistes. A titre d'exemple on peut citer la base sur les séquences nucléiques d'*Eschérichia coli* **ECD**, la base **NRL3D** des séquences protéiques dont la structure a été déterminée ou bien encore des bases de motifs nucléiques ou protéiques telle que **PROSITE** .