**CHAPITRE 6 : BASES DE DONNEES**

**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.

**Définition**

Les bases de données biologiques sont des bibliothèques répertoriant des informations sur les [sciences de la vie](http://fr.wikipedia.org/wiki/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](http://fr.wikipedia.org/wiki/G%C3%A9nomique), la [protéomique](http://fr.wikipedia.org/wiki/Prot%C3%A9omique), la [métabolomique](http://fr.wikipedia.org/wiki/M%C3%A9tabolomique), la [phylogénétique](http://fr.wikipedia.org/wiki/Phylog%C3%A9n%C3%A9tique) et les [puces à ADN](http://fr.wikipedia.org/wiki/Puces_%C3%A0_ADN). Parmi le contenu des bases de données, on trouve des informations à propos de la fonction, de la structure, de la localisation ([cellulaire](http://fr.wikipedia.org/wiki/Cellulaire) et [chromosomique](http://fr.wikipedia.org/wiki/Chromosome)) des [gènes](http://fr.wikipedia.org/wiki/G%C3%A8ne) et les effets cliniques de leurs [mutations](http://fr.wikipedia.org/wiki/Mutation), 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](http://fr.wikipedia.org/wiki/Biomol%C3%A9cules) et leurs interactions à l'ensemble du métabolisme des organismes, et même l'[évolution](http://fr.wikipedia.org/wiki/%C3%89volution) des [espèces](http://fr.wikipedia.org/wiki/Esp%C3%A8ces).

– **Système de gestion de base de données** : ensemble de programmesqui permettent l'accès à une base de données : transactions, intégrité, sécurité, administration.

**– 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.

**Accès**

La plupart des bases de données biologiques sont accessibles sur des [sites web](http://fr.wikipedia.org/wiki/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](http://fr.wikipedia.org/wiki/PubMed) ou [OMIM](http://fr.wikipedia.org/wiki/OMIM),
* Des données de séquençage sont disponibles sur [GenBank](http://fr.wikipedia.org/wiki/GenBank) (ADN) et [UniProt](http://fr.wikipedia.org/wiki/UniProt) (protéines),
* Des structures spatiales protéiques sont disponibles sur [PDB](http://fr.wikipedia.org/wiki/PDB), [SCOP](http://fr.wikipedia.org/wiki/SCOP) et [CATH](http://fr.wikipedia.org/w/index.php?title=CATH&action=edit&redlink=1).

**Les types de bases de données**

Nous distinguerons deux types de banques :

**1- Les banques généralistes**: celles qui correspondent à une collecte des données la plus exhaustive possible et qui offrent finalement un ensemble plutôt hétérogène d'informations.

En Europe, financée par l'EMBO (EuropeanMolecularyBiology Organisation), une équipe s'est constituée pour développer une banque de séquences nucléiques ([**EMBL**](http://www.ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html) 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 (EuropeanBioinformatics Institute).**](http://www.ebi.ac.uk/ebi_home.html)

Du côté américain, soutenue par le NIH (National Institute of Health) une banque nucléique nommée [**GenBank**](http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html) a été créée à Los Alamosen 1986. Cette base de données est diffusée maintenant par le [**NCBI (National Center for Biotechnology Information).**](http://www.ncbi.nlm.nih.gov/) 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 BiomedicalResearchFoundation) 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**](http://www-nbrf.georgetown.edu/pir/) (ProteinIdentification Ressource, 1986).
* La deuxième, [**Swissprot**](http://expasy.hcuge.ch/sprot/sprot-top.html) 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. Elles se 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.

**2- Les banques spécialisées**

Pour des besoins spécifiques, de nombreuses bases de données spécialisées ont été créées au sein des laboratoires. Certaines ont continué d'être développées, d'autres n'ont pas été mises à jour et ont disparu car elles correspondaient à un besoin ponctuel. D'autres enfin sont inconnues ou mal connues et attendent qu'on les exploite davantage. Toutes ces bases de données spécialisées sont d'intérêt très divers et la masse des données qu'elles représentent peut varier considérablement d'une base à une autre. 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**](http://bmc1.bmc.uu.se/srs/srsc?-info+ECD), la base [**NRL3D**](http://www.bis.med.jhmi.edu/) des séquences protéiques dont la structure a été déterminée ou bien encore des bases de motifs nucléiques ou protéiques telle que [**PROSITE**](http://expasy.hcuge.ch/sprot/prosite.html) .

**3- Conclusion**

Le nombre de données dans le domaine de la biologie ne cesse d'augmenter en particulier avec le séquençage des génomes de différents organismes mais l'on assiste également à une grande diversification des informations produites (séquences primaires, structures moléculaires, cartographie, collection de souches ou de clones...). Toutes ces données sont regroupées dans des banques de données très variées dans leur volume et leur nature.

Ici, nous avons étudié essentiellement les grandes banques car malgré leurs imperfections, ce sont les plus utilisées du fait de leur compatibilité avec des logiciels comme BLAST et FASTA. Pour que les bases de données soient plus facilement exploitables, et que les utilisateurs puissent extraire les sous-ensembles de séquences qui les intéressent, des logiciels leur sont généralement offerts soit localement, soit sur les nombreux serveurs publics : les systèmes d’interrogation dédiés qui sont programmés exclusivement pour la manipulation de séquences biologiques. Mais ils sont généralement adaptés aux principales bases de données (EMBL, Genbank …). On peut citer parmi eux des logiciels comme[**ACNUC**](http://www.infobiogen.fr/srs5/man/srsman.html) ou [**SRS**](http://pbil.univ-lyon1.fr/databases/acnuc.html)**.**

Cependant le recensement et la description de toutes les bases publiques généralistes et spécialisées liées à la structure primaire des séquences mais aussi liées à d'autres caractéristiques biologiques peuvent être consultée dans des bases de données spécifiques qui donnent des listes actualisées des différentes banques d'intérêt biologique comme [**LMB**](http://www.dna.affrc.go.jp/htdocs/LIMB/)**D** (Listing of MolecularBiologyDatabases) ou [**DBCAT**](http://www.infobiogen.fr/services/dbcat/).