Chapitre 2 : Les Banques et Bases De Données Biologiques

Introduction

Internet offre au biologiste une quantité écrasante d'information et d'outils pour analyser les données du vivant et on trouve assez facilement des listes de sites intéressant.

Certains serveurs proposent d'analyser les données en direct (réponse sur une page Web) ou en différé (réponse par e-mail). D'autres permettent de télécharger leurs programmes pour les installer localement. Théoriquement, la recherche des séquences semblables à une séquence donnée nécessite la comparaison de toutes les séquences de la banque avec la séquence requête.

Il est impossible de citer toutes les bases de données biologiques ici, cependant il est intéressant de connaitre et suivre les bases de données les plus importantes dans votre domaine (Tableau 1):

Tableau 1. Principaux serveurs généralistes de bioinformatique.

National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/genbank).
European Bioinformatics Institute of the European Molecular Biology Laboratory (EBI-EMBL, http://www.ebi.ac.uk/services).
DNA Data Bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp).
UniProt KnowledgeBase (http://www.uniprot.org) contient les séquences protéiques avec leurs annotations fonctionnelles.
Protein Data Bank (PDB, http://www.rcsb.org/pdb) contient les informations sur la structure tridimensionnelle des protéines.
ExPASy Molecular Biology Server: http://www.expasy.ch
Informatique appliquée à l'étude des Biomolécules des Génomes: http://www.infobiogen.fr
Institute for Genomic Research : http://www.tigr.org

La littérature scientifique peut être recherchée sur PubMed (http://www.ncbi.nlm.nih.gov/pubmed) ou Google Scholar (http://scholar.google.com). Vous pouvez chercher les articles par auteur ou mots clés.

1.1. Les banques nucléiques

Les données stockées dans ce type de banques sont des données issues de séquençage d'ADN et ARN. Trois banques nucléiques connues, elles partagent des informations et donc contiennent des ensembles presque identiques de séquences**.** Ces trois banques s'échangent systématiquement leur contenu depuis 1987 et ont adopté un système de conventions communes : « DDBJ/EMBL/GenBank »:

• La banque EMBL: créée en 1980 et financée par l'EMBO (European Moleculary Biology Organization), développée au sein du Laboratoire Européen de Biologie Moléculaire situé à Heidelberg (Allemagne), elle est maintenant diffusée par l'EBI: http://www.ebi.ac.uk/embl/. En 24 février 2014, la banque contient 369.5 millions séquences.

• La banque GenBank (Genetic Sequence Databank): créée en 1982 par la société IntelliGenetics et diffusée maintenant par le NCBI (National Center for Biotechnology Information) : http://www.ncbi.nlm.nih.gov/. En février 2014 la banque contient 171.123.749 séquences. GenBank contient une sous-banque de protéines, traduction des séquences nucléiques, appelée GenPept.

• La banque DDBJ (DNA Databank of Japan): créée en 1986 et diffusée par le NIG (National Institute of Genetics, Japon), a enregistré un total de 81.994.905 de séquences ADN le moi de décembre 2019 (DDBJ 2019).

1.2. Les banques protéiques

Les données stockées dans ces bases sont issus d'une traduction de séquences d'ADN ou par le séquençage de protéines (rare car long et coûteux):

• *La banque SwissProt* **:** est une banque protéique crée en 1986 à l'Université de Genève et maintenue depuis 1987 dans le cadre d'une collaboration, entre cette université (via ExPASy, Expert Protein Analysis System) et l'EBI. Celle-ci regroupe aussi des séquences annotées de la banque PIR-NBRF ainsi que des séquences codantes traduites de l'EMBL. En février 2014 la banque contient 542503 séquences compressant 192888369 acides aminés.

1.3. Les banques structurelles

Elles sont des banques spécialisées pour les structures 2D et 3D des protéines. Plusieurs banques connues dans ce contexte nous citons ici à titre d'exemple la banque PDB:

• *La banque PDB* (Protein Data Bank) créée en 1971, c'est la banque de référence des structures protéiques obtenues expérimentalement par cristallographie rayon X, spectroscopie RMN et cryo-microscopie électronique (technique la plus récemment utilisée). Les

coordonnées des atomes formant la structure d'une protéine, le détail de la séquence, les conditions de cristallisation sont les principales informations disponibles pour chaque structure de la banque. C'est à partir de cette banque que sont détectés les homologues structuraux. La Figure 1 représente l'évolution du nombre de structures protéiques enregistrées par année sur PDB, le moi de janvier 2020 a remarqué un total de 147.827 structures.

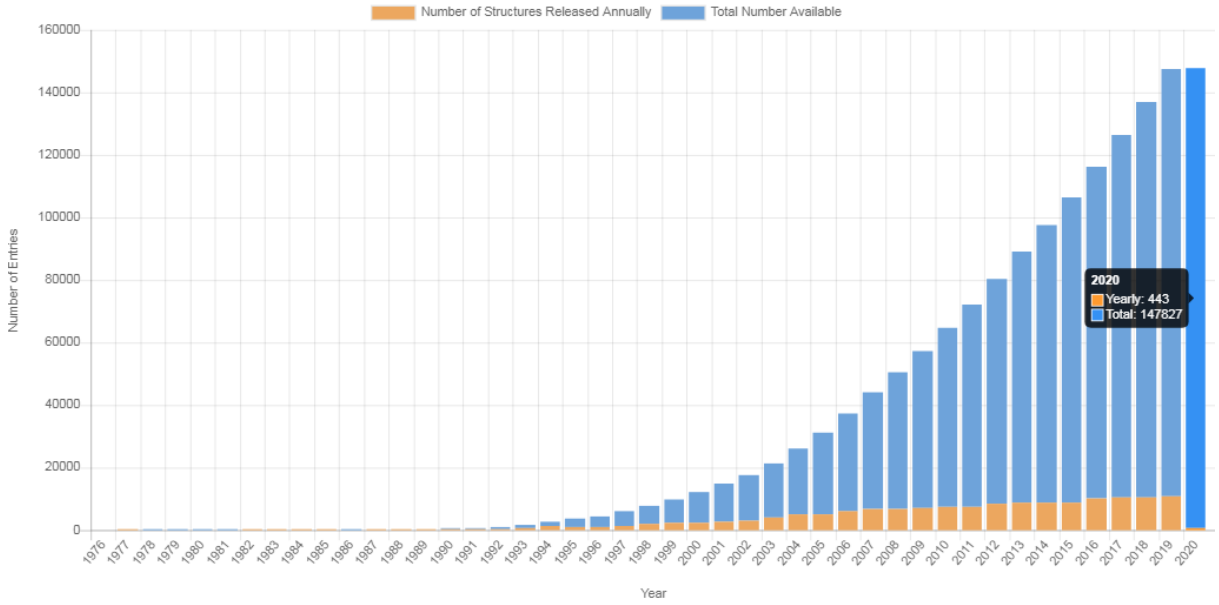

Figure 1. Statistiques des structures protéiques PDB réalisées par année.