bioinformatique

Table des matières

I Chapitre 1 : Les Banques et Bases De Données Biologiques

1. objectif spécifique

1- connaître la définition de bases de données et banques de données

2- faire la différence entre bases de données et banques de données

3- connaître la définition de banques nucléiques, protéiques, structurelles

4- connaître les étapes du travailles sur la base de données NCBI

2. 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 nous faut distinguer deux choses : qu'est ce qu'une base de données (BD) ?

différence entre banque de données et base de données ?

Une base de données, usuellement abrégée en BD ou BDD , Une base de données est un fichier ou un ensemble de fichiers permettant le stockage permanent ou temporaire des informations ainsi que l'accès à ces informations devenues structurées (http://www.webadev.com/lexique-b-base-dedonnees. php).

C'est un tableau dans lequel on intègre des informations de manière logique et structurée comme la liste d'un groupe d'étudiants :(1)*

2.1. Différence entre bases de données et banques de données

Il convient de dire qu'une banque de données est une base de données (car tableau structuré) mais qui contient des informations biologiques hétérogènes (virus, bactéries, champignons, végétaux, animaux) alors qu'une base de données est plus spécialisée (base spécifique à E. coli, à Bacillus, etc.).

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):

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

Tableau 1. Principaux serveurs généralistes de bioinformatique.

2.2. 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).(2)*

2.3. 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.

2.4. 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.(3)*

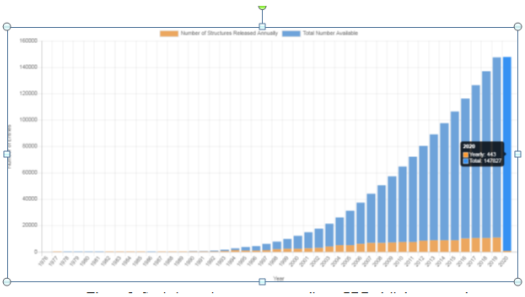

Figure 1. Statistiques des structures protéiques PDB réalisées par année.

2.5. Recherche d'Alignement des séquences du gène ARNr16S sur NCBI

1. Ouverture du lien NCBI sur internet par l'utilisation du moteur de recherche

Google

2. Choix du programme BLAST

3. Choix de l'outil nucleotide BLASTn

4. Insertion de la séquence ADN ou le Numéro d'Accès sur Gene Bank et activation

de l'outil BLAST

5. Lecture de la liste des résultats de l'Alignement

6. Lecture du détail des résultats de l'Alignement

7. Récolte des informations sur l'individu par le numéro d'accès sur Gene Bank :

Auteur, affiliation, publication, séquence, etc

pour voir la vidéo cliquiez *ici*

Glossaire

bioinformatique

est la science de l'utilisation de l'ordinateur dans l'acquisition, le traitement et l'analyse de l'information biologique

Abréviations

NCBI : National Center for Biotechnology Information

PDB : Protein Data Bank

Références

référence
Corpet F., Chevalet C. (2000) Génétique moléculaire : principes et application aux populations animales. INRA Prod. Anim. Numéro hors série : 191-195.

référence
Hochreiter S. (2013) Bioinformatics I, Sequence Analysis and Phylogenetics. Institute of Bioinformatics Johannes Kepler University Linz, Austria. 166p.

référence
Darlu P., Tassy P., (2004). La reconstruction phylogénétique. Concepts et méthodes. Masson. Paris. ISBN : 2-225-84229-9. 241 p.

référence
Imbs D., Hassan M.S. (2000) Bioinformatique Travail d'étude. Université de Nice Sophia Antipolis. 23p.

référence
- Saitou N., Nei M. (1987). The Neighbor-joining Method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4 (4): 406-425.

référence
Tamura K., Stecher G., Peterson D., Filipski A., Kumar S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Molecular Biology and Evolution: 30 2725-
2729.

Bibliographie

Boubendir A. (2019) Cours de Bioinformatique Centre Universitaire de Mila, Algérie.