CHAPITRE II.

RECHERCHE DANS LES BASES ET LES BANQUES DE DONNÉES

Définition

Les bases de données biologiques sont des bibliothèques électronique et informatisé qui contiennent 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.

1. Rôle des banques et bases de données biologiques

Leur principale mission est de rendre publiques les séquences qui ont été déterminées ainsi des premiers intérêts de ces banques est la masse de séquences qu'elles contiennent. Entre autres ils ont pour mission l'archivage, le stockage, la diffusion et l'exploitation des données biologiques.

2. Contenues des bases de données biologiques

Ces bases de données peuvent contenir des informations : (ADN, protéines, gènes et génomes, taxonomie, autres, … etc.). On y trouve également une bibliographie et une expertise biologique directement liées aux séquences traitées.

3. Types de banques de données

Il existe un grand nombre de bases de données d'intérêt biologique. Nous nous limiterons à une présentation des principales banques de données publiques, basées sur la structure primaire des s séquences, qui sont largement utilisées dans l'analyse informatique des séquences.

Nous distinguerons deux types de banques, celles qui correspondent à une collecte des données la plus exhaustive possible et qui offrent finalement un ensemble plutôt hétérogène d'informations (banques de données généralistes) et celles qui correspondent à des données plus homogènes établies autour d'une thématique (banques de données spécialisées) et qui offrent une valeur ajoutée à partir d'une technique particulière ou d'un intérêt par un groupe de scientifiques.

3.1. Les banques de séquences généralistes

C'est au début des années 80 que les premières banques de séquences sont apparues sous l'initiative de quelques équipes dont la première à l'initiative de Grantham et C. Gautier à Lyon. Elles couvrent tous les secteurs de la biologie, toutes les espèces. Ainsi, plusieurs organismes ont pris en charge la production de telles bases de données.

3.1.1.Les banques de séquences nucléiques

❖ **EMBL (European Molecular Biology Laboratory ou Laboratoire Européen de Biologie Moléculaire)** : banque européenne créée en 1980 et financée par l'EMBO (European Moleculary Biology Organization), elle est aujourd'hui diffusée par l'EBI (European Bioinformatics Institute, Cambridge, UK) ;

❖ **GenBank** : créée en 1982 par la société IntelliGenetics et diffusée maintenant par le NCBI (National Center for Biotechnology Information, Los Alamos, US) ;

❖ **DDBJ (DNA Data Bank of Japan)** : créée en 1986 et diffusée par le NIG (National Institute of Genetics, Japon) ;

3.1.2. Les banques protéiques

❖ **PIR-NBRF (Protein Information Resource-National Biomedical Research Foundation) :** créée en 1984 par la NBRF (National Biomedical Research Foundation). Elle est maintenant un ensemble de données issues du MIPS (Martinsried Institute for Protein Sequences, Munich, Allemagne) et de la banque japonaise JIPID (Japan International Protein Information Database) ;

❖ **SwissProt** : créée 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 PIRNBRF ainsi que des séquences codantes, traduites de l'EMBL.

Elles contiennent la protéine obtenue de plusieurs manières différentes :

➢ *in silico* : déduite à partir de la séquence nucléique, par simple traduction du ou des exons la codant

➢ isolée à partir de la cellule

- ou encore par génie génétique

3.1.3 Avantages et inconvénients

❖ Avantages

- Ces banques sont d'une importance majeure car elles offrent des informations qui ne sont plus reproduites dans la littérature scientifique (livres ou articles)
- Ces informations sont gratuites
- On y trouve une bibliographie et une expertise directement liées aux séquences traitées

❖ Inconvénients

- Manques de vérification de séquences soumises
- Le temps d'insertion des séquences (retard … !)

3.2. Les bases de données de séquences spécialisées

Elles couvrent un secteur défini de la biologie. Pour des besoins spécifiques, de nombreuses
bases de données spécialisées ont été créées, Certaines sont pérennes et continuent d'être développées et mises à jour, d'autres sont laissées à l'abandon et enfin d'autres ont disparu. On en dénombre à cette date un peu plus d'un millier, accessibles directement par le Web. La nature ainsi que la quantité d'informations sont très variables.

3.2.1. Organisme

Ces banques regroupent les données pour **un organisme particulier**, ou un groupe, contenant tout ou partie des informations suivantes :

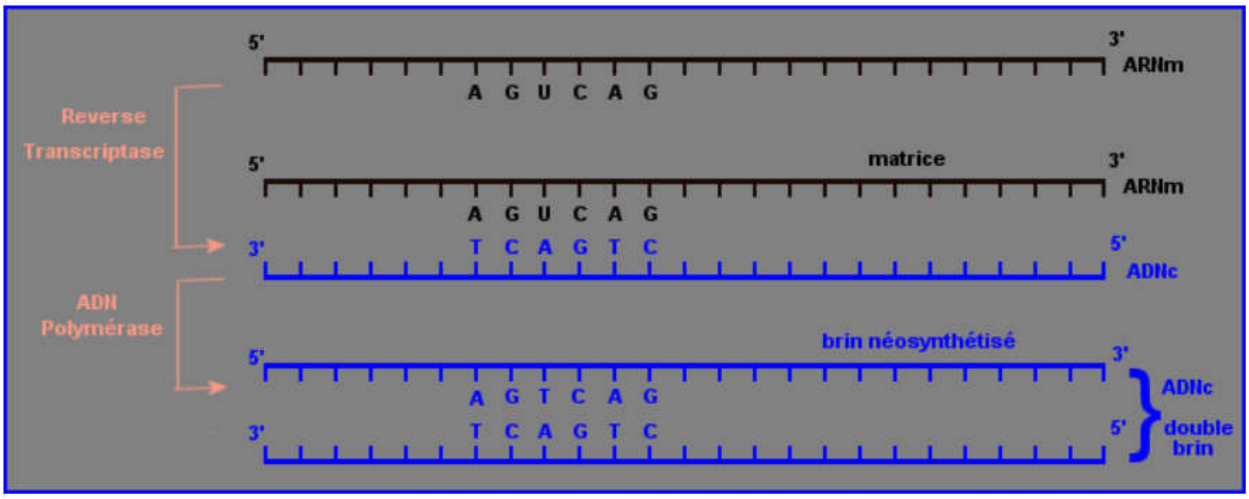
- **carte physique chromosomique** : la cartographie *physique* est de localiser les gènes sur les *chromosomes*.
- **carte génétique et liaison** :

-clonage positionnel pour les gènes :

- ✓ **EST (marqueurs de séquences exprimées) :** Un **marqueur de séquence exprimée**, ou ***expressed sequence tag*** **(EST)**, est une courte portion séquencée d'un ADN complémentaire (ADNc), utilisée comme marqueur pour différencier les gènes entre eux dans une séquence ADN et identifier les gènes homologues dans d'autres espèces.

  Parce qu'il est généralement assez facile de récupérer des brins d'ARNm des cellules, les biologistes récupèrent ces séquences et les convertissent en ADNc, qui est bien plus stable. Un ARNm étant forcément l'expression d'un gène du génome, cet ADNc n'est pas une copie exacte de la séquence ADN qui a généré l'ARN car, à la suite de l'épissage, l'ARN ne garde pas les régions non codantes de l'ADN (introns).

- ✓ **Banque d'ADNc :** L'**ADN complémentaire** (ou **ADNc**, **Acide désoxyribonucléique complémentaire**) est un simple brin artificiellement synthétisé à partir d'un ARNm, représentant ainsi la partie codante de la région du génome ayant été transcrit en cet ARNm. Il est obtenu après une réaction de transcription inverse d'un ARNm mature et équivaut donc à la copie ADN de l'ARNm qui a été extrait dans une cellule donnée à un moment donné.



- ✓ **Banque de vecteurs de clonage** : On appelle vecteur l'ADN dans lequel on insère le fragment d'ADN à étudier. L'ADN inséré est appelé insert ou ADN étranger ou ADN exogène. Cette séquence nucléotidique est capable de s'auto-répliquer.

  Les vecteurs sont donc des petits ADN dans lesquels on insère un fragment d'ADN que l'on veut étudier. Ces petits ADN sont généralement des plasmides ou des bactériophages. Il est nécessaire d'introduire ces bactériophages ou plasmides dans les bactéries pour réaliser une multiplication de ceux-ci.

Les plasmides : Les plasmides sont des petits fragments d'ADN circulaire présents dans la cellule bactérienne et indépendants du génome bactérien.

Les phages : Les phages sont des virus qui infectent les bactéries.

- Gène et expression

- Cytogénétique et anomalies chromosomiques

- Gène et maladie

- Oncogènes

3.2. Banques nucléiques spécialisée

Elles sont spécialisées dans les informations suivantes :

-EST, ADNc

- ARN

- Structure secondaire d'ARN

- Signaux et éléments de régulation

- Sondes, amorces

- Alignements

- Famille de gènes

3.3.Banques protéiques spécialisées

Elles sont spécialisées dans les informations suivantes :

- Alignement

- Classification structurale

- Familles de protéines

- Interactions

- Enzymes

- Modifications protéiques post-traductionnelles

- Pathologies

- Gels bidimensionnels

- Bases protéiques sur l'interaction et la thermodynamique des protéines

3.4. Banques immunologiques

Elles sont spécialisées dans les informations suivantes :

- Séquences

- Récepteur (cellule T, par exemple)

- Complex MHC (Major Histocompatibility Complex) : un système de reconnaissance du soi présent chez la plupart des vertébrés. Les molécules du CMH sont à la surface de toutes les cellules nucluéées pour le CMH de classe I et les cellules présentatrices de l'antigène

pour le CMH de classe II qui assurent la présentation de l'antigène aux lymphocytes T afin de les activer.

On distingue les complexes majeurs d'histocompatibilité de classe I et de classe II. Chez l'être humain, on parle d'antigène HLA.

- Système HLA

3.5. Banques Structure 2D ou 3D

Elles sont spécialisées dans les informations suivantes :

- Coordonnées 3D de protéines
- Structure secondaire des protéines
- Domaines structuraux : est une partie d'une protéine capable d'adopter une structure de manière autonome ou partiellement autonome du reste de la molécule
- Centre actif des enzymes
- Complexes récepteurs-ligands
- Atlas de topologie structurale des protéines

❖ Avantages

- Elles fournissent des informations détaillées, spécifiques du domaine biologique qui n'existent pas dans les systèmes généralistes
- Les données sont en général contrôlées, donc plus fiables et de meilleure qualité que dans les bases généralistes
- Elles évoluent en fonction des progrès scientifiques dans le domaine plus facilement

❖ Inconvénients

Ne cible pas toujours ce que l'on veut ; toutes les banques possibles n'existent pas