

Chapitre 5 : Les bases de données

1. Les bases de données

Une base de données est un système électronique qui permet d'accéder facilement à un ensemble organisé de données, de les manipuler et de les mettre à jour. Elle est utilisée par les organisations comme méthode de stockage, de gestion et de récupération de l'information.

Une base de données est stockée sous la forme d'un fichier ou d'un ensemble de fichiers sur un disque ou un disque magnétique, un disque optique ou tout autre support de stockage. L'information contenue dans ces fichiers peut être divisée en enregistrements. Ces enregistrements sont constitués d'un ou de plusieurs champs. Un champ constitue une seule pièce d'information, et chaque champ contient généralement des informations se rapportant à un aspect ou attribut de l'entité décrite par la base de données.

Les enregistrements sont également organisés en tableaux qui contiennent des informations sur les relations entre les différents champs. À l'aide de mots-clés et de diverses commandes de tri, les utilisateurs peuvent rapidement rechercher, réorganiser, regrouper et sélectionner les champs dans de nombreux enregistrements pour récupérer ou créer des rapports sur des agrégats particuliers de données.

2. Bases de données biologiques

Les 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 des gènes et les effets cliniques de leurs mutations, ainsi que leur similarité 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.

On peut rechercher des similitudes entre l'objet génomique et différents bases de données.

Chapitre 5 : Les bases de données

A. Les bases de données génomiques :

L'INSDC (international nucleotide sequence Database collaboration) est constitué du regroupement des trois bases de données de séquence nucléotidiques DDBJ (Japon), Genbank (Etats-Unis) et EMBL (Europe) les trois bases de données échangent toute nouvelle donnée ou mise à jour dans le but d'une synchronisation optimale.

B. Les bases de données protéiques

Les principales banques de données de séquence protéiques sont Swiss-prot et PIR. On y trouve des informations sur la structure et fonction de la protéine ainsi que des liens vers diverses ressources sur des protéines apparentées. Ces bases sont complétées par TrEMBL. Une base de séquence de protéines issues de la traduction automatique des séquences codantes des bases de données nucléotidiques.

C. Les bases de données de motifs

Le motif ou sens strict est un segment court, continu et non ambigu d'une séquence nucléique ou protéique. Les motifs peuvent être :

- Une séquence exacte (ex : le site de restriction de l'enzyme EcoR/GAATTC).
- Une séquence consensus (ex : une séquence d'acides aminés répétée dans une protéine).

Des bases telles que PROSITE, PRINTS, MOTIFS peuvent être utilisées.

D. Les bases de données de domaines

Les protéines sont composées de domaines fonctionnels. Les bases de données de domaines telles que InterPro, ProDom, PFAM, eBlockS, SMART sont des collections de séquences protéiques présentant une parenté structurelle avec un domaine protéique fonctionnel connu.

E. Les bases de données d'expression

Pour l'annotation des génomes eucaryotes. La façon la plus directe d'identifier un gène est de montrer que le fragment génomique considéré est transcrit. Pour cela, des comparaisons de séquences nucléotidiques sont effectuées entre le génome à annoter et des bases de données d'expression, c'est-à-dire banques d'ADN complémentaire (ADNc) ou de marqueurs de séquences exprimées (EST pour expressed sequence tags en anglais). Les EST sont des petits morceaux de séquences d'ADNc la collection dbEST est un exemple de base de données d'EST.

Chapitre 5 : Les bases de données

Cette étape de l'annotation fonctionnelle permettra de valider les gènes prédits lors de l'annotation syntaxique et de trouver des gènes que les programmes informatiques de prédiction ont échoués à identifier.

3. Système de gestion de base de données

Un système de gestion de base de données (SGBD) est un terme bien connu dans l'analyse de données. Il fait référence à une collection de programmes qui permettent aux utilisateurs d'accéder à des bases de données et de manipuler, de maintenir, de signaler et de relier des données. Un SGBD est souvent utilisé pour réduire la redondance des données, partager les données de manière contrôlée et réduire les problèmes d'intégrité des données. SGBD n'est pas un système d'information, mais simplement un logiciel.

Le modèle relationnel, qui enregistre les données dans les formats de table, est le DBMS le plus utilisé. Le SGBD relationnel organise l'information en lignes, colonnes et tableaux, ce qui facilite la recherche d'informations pertinentes. Les bases de données relationnelles sont populaires parce qu'elles sont faciles à étendre, et de nouvelles catégories de données peuvent être ajoutées après la création de la base de données originale sans grande quantité de modification.

4. Index

Un index dans un Système de Gestion de Base de Données (SGBD) est une structure de données spéciale (généralement un Arbre B ou B-Tree) qui accélère considérablement la récupération des données en agissant comme une table des matières. L'objectif principal est de permettre au SGBD d'éviter un balayage complet de la table.

5. Accès

C'est référence aux méthodes et mécanismes utilisés pour interagir avec les données stockées. Ces méthodes décrivent comment le SGBD localise et récupère physiquement l'information sur le disque

6. Les types de bases de données

a. Les banques généralistes

On appelle banques généralistes, ou banques primaires, les ressources qui collectent, Gèrent, archivent et mettent à disposition de la communauté scientifique un ensemble De données primaires. Classiquement, on considère comme banques primaires les banques généralistes qui

Chapitre 5 : Les bases de données

contiennent des séquences nucléiques et protéiques obtenus par des méthodes expérimentales. Bien qu'actuellement la plupart des séquences protéiques ne soient pas obtenues expérimentalement, mais à partir des données de séquence nucléiques.

b. Les banques spécialisées

Les banques spécialisées se concentrent sur un **domaine précis, une seule espèce, ou un type d'information très spécifique**. Ces banques contiennent des données homogènes facilité pour mettre à jour les données, vérifier leur intégrité, offrir une interface adaptée,

- Collecte établie autour d'une thématique particulière
- Inconvénients : ne cible pas toujours ce que l'on veut ; toutes les banques possibles n'existent pas

7. Quelques bases spécifiques aux plantes

Plusieurs bases de données sont spécifiquement conçues pour les plantes, regroupant des informations allant de la taxonomie (classification) aux données génomiques et écologiques. Elles sont utilisées par les botanistes, les écologues, les agronomes et les bio-informaticiens.

a) Bases de Données Taxonomiques et Floristiques : Ces bases de données se concentrent sur l'identification, la localisation et la description des espèces végétales.

- **Flora Data (Tela Botanica) :** Une base de données collaborative française. Elle permet aux participants de recenser les espèces observées (flore) et d'améliorer la connaissance floristique du territoire. Elle offre des outils d'identification comme IdentiPlante pour l'aide à la détermination collective des espèces.
- **Plants of the World Online (Royal Botanic Gardens, Kew) :** Une référence mondiale pour la taxonomie et l'écologie des plantes. Elle fournit un grand ensemble de données sur la classification, la nomenclature et la distribution géographique des espèces.
- **GBIF (Global Biodiversity Information Facility) :** Bien que non exclusivement pour les plantes, c'est une immense plateforme qui centralise les données d'occurrence (où une espèce a été observée) de millions de spécimens végétaux provenant de collections et d'observations du monde entier.

Chapitre 5 : Les bases de données

- **Végébase** : Une base de données de connaissances et de choix des végétaux, souvent utilisée par les professionnels de l'aménagement paysager. Elle compile des informations sur les propriétés morphologiques, les besoins et le comportement des végétaux face aux contraintes (urbaines, sol, résistance).
- b) Bases de Données Génomiques et Bio-informatiques** : Ces ressources stockent les données génétiques, moléculaires et biochimiques pour comprendre la fonction des gènes et le développement des plantes.
- **GnpIS (Génomes et phénotypes d'espèces végétales)** : Un système d'information français (INRAE) qui intègre les données de génétique, génomique et phénotypage pour les plantes et les pathogènes végétaux. Il facilite l'accès aux séquences, aux cartes génétiques et aux annotations.
 - **URGI (Unité de recherche génomique-info)** : Une plateforme fournissant un support bio-informatique crucial pour la publication et l'intégration des données de génétique et de génomique végétale, notamment l'annotation des éléments transposables (séquences d'ADN mobiles).
 - **Banques de Plantes Modèles** : De nombreuses bases de données sont dédiées à des plantes modèles dont le génome est entièrement séquencé et bien étudié, comme *Arabidopsis thaliana* (l'arabette des dames) ou le riz. Les connaissances acquises sur ces modèles sont souvent extrapolées à d'autres espèces cultivées.
- c) Bases de Données sur les Utilisations (Phytothérapie)** : Certaines bases se focalisent sur les usages pratiques et médicinaux des plantes.
- **Bases de données de phytothérapie** : Elles compilent des informations sur l'identité, les propriétés médicinales et les précautions d'emploi des plantes utilisées en compléments alimentaires ou en médecine traditionnelle. Ces bases sont souvent gérées par des agences de santé publique ou des organisations professionnelles (comme l'Anses en France).
 - **PFAF (Plants For A Future)** : Axée sur les plantes comestibles, médicinales et utiles, elle répertorie les usages des espèces végétales.