Traitement informatique des données

Présentation du module:

Cette matière contribue au développement des acquis de l'étudiant en ce
Qui concerne les nouvelles bases de la bioinformatique

Connaissances préalables recommandées
Génétique
Biologie moléculaire
Microbiologie
Biochimie

Cours 02: les bases de données

Chargée de cours : ALIANE S

1. Introduction :

Les fichiers contenant l'information biologique sous la forme de séquences constituent l'élément central autour duquel les bases de données se sont constituées à l'origine.

On peut distinguer :

- Les bases de données généralistes : elles correspondent à une collecte des données la plus exhaustive possible et qui offrent un ensemble d'informations diverses.
- Les bases de données spécialisées : elles correspondent à des données plus homogènes établies autour d'une thématique. Cette spécialisation apporte une valeur ajoutée aux données concernées.

Il existe un très grand nombre de bases de données d'intérêt biologique. Le panorama de ces milliers de bases de données biologiques nécessitent cependant un préalable qui s'appuie sur une forme de "sagesse" :

- La maintenance et la mise à jour actives de bases de données biologiques publiques sur le Web demandent beaucoup temps et sont coûteuses.
- Sans soutien institutionnel ou plan de viabilité financière, la plupart des bases de données créées en tant que résultats de projets de recherche meurent ou sont archivées dans un délai de 10 à 15 ans.

1. Les bases de données généralistes

Les bases de données généralistes sont indispensables à la communauté scientifique car elles regroupent des données et des résultats essentiels dont certains ne sont plus reproduits dans la littérature scientifique.

Exemples de bases de données considérées comme des recueils de référence mondiaux :

NCBI ("National Center for Biotechnology Information")

EBI ("European Bioinformatics Institute")

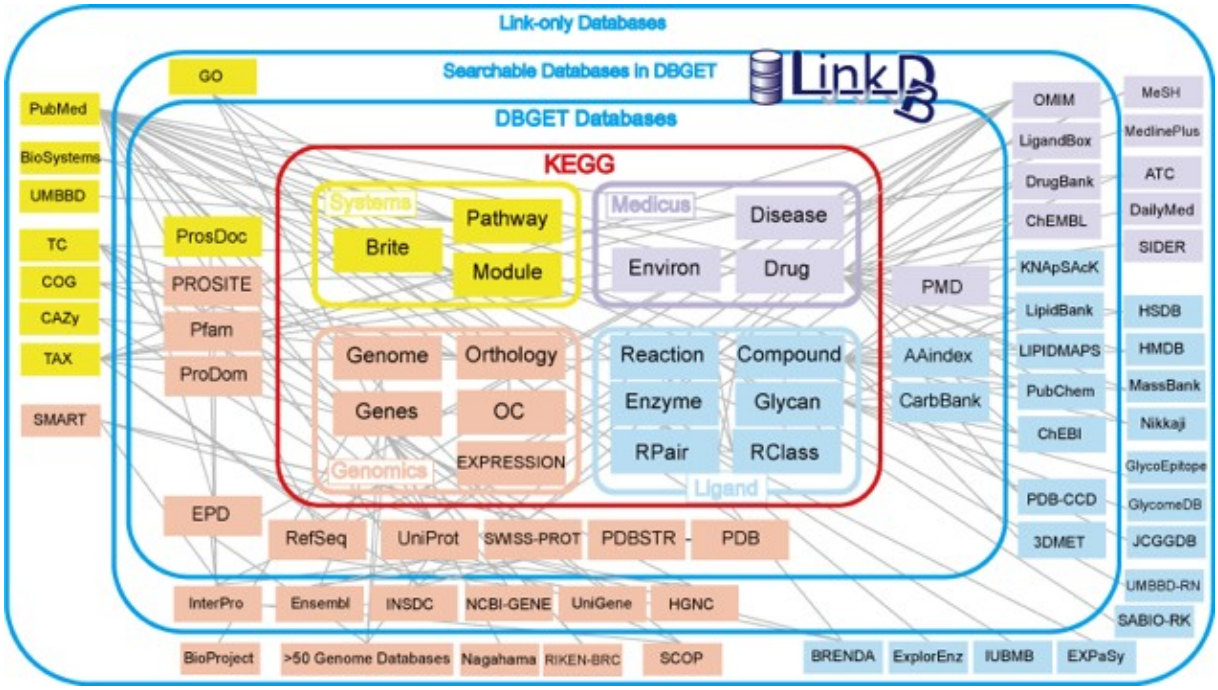
Uniprot

PDB ("Protein Data Bank")

KEGG ("Kyoto Encyclopedia of Genes and Genomes")

Dans le cadre de l'analyse des séquences, par exemple, le fait que la majorité des séquences connues soit réunie en un seul ensemble est un élément fondamental pour la recherche de similitudes avec une nouvelle séquence. D'autre part, la grande diversité d'organismes qui y est représentée permet d'aborder des analyses de type évolutif.

La figure ci-dessous montre le réseau de liens établis entre la base de données généraliste KEGG et les grandes autres bases de données généralistes.

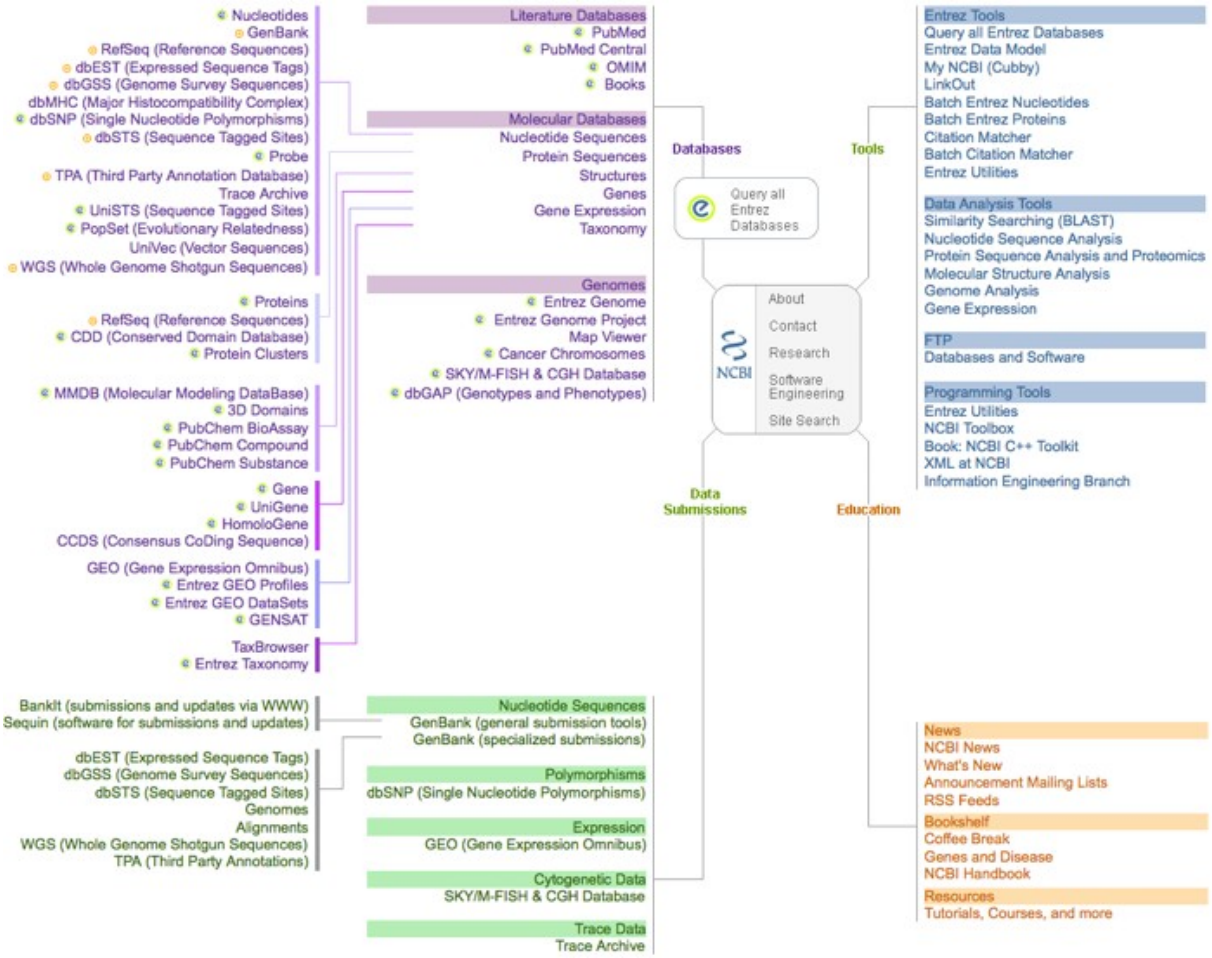


En réalité, figure originale permet la visualisation de ces liens et surtout de récupérer, pour un fichier donné, l'ensemble des fichiers équivalents dans les autres bases de données.

2. Exemple d'une base de données généraliste

Genbank - NCBI : Créée en 1982 par la société *IntelliGenetics* et diffusée maintenant par le NCBI ("*National Center for Biotechnology Information*", Bethesda - Maryland).

Figure ci-dessous : "*site map*" de l'ensemble de la base de données du NCBI.



Autres exemples de bases de données généralistes

DDBJ (*"DNA Data Bank of Japan"*) : Créée en 1986 et diffusée par le NIG ("*National Institute of Genetics*", Japon).

- UniProt ("*Universal Protein Resource*") : base de données mondiale des protéines créé par le consortium [EBI - SIB - PIR]. Voir par exemple *ExPASy Proteomics Server*.
- Swissprot & TrEMBL : Elle a été constituée à l'Université de Genève à partir de 1986. Elle est maintenant développée par le SIB (*Swiss Institute of Bioinformatics*) et l'EBI. Elle regroupe (entre autres) des séquences annotées de la PIR-NBRF ainsi que les séquences codantes traduites de l'EMBL (*TrEMBL*).

Voir un développement ci-dessous.

Ces grandes bases de données généralistes s'échangent systématiquement leur contenu depuis 1987 et adoptent un système de conventions communes (*The DDBJ/EMBL/GenBank Feature Table Definition*).

PIR-NBRF ("*Protein Information Ressource*") : banque de protéines créée sous l'influence du NBRF ("*National Biomedical Research Foundation*") à Washington. Elle diffuse maintenant des données issues du MIPS ("*Martinsried Institute for Protein Sequences*"), de la base

Japonnaise JIPID ("*Japan International Protein Information Database*") et des données propres de la NBRF.

GOLD ("*Genomes OnLine Database*") : base de données qui recensse les milliers de génomes séquencés ou en voie de séquençage.

"*Nucleic Acids Research*" (NAR) est un exemple de journal scientifique dédié plus particulièrement à la diffusion des bases de données biologiques.

4. Les bases de données spécialisées

Pour des besoins spécifiques liés à l'activité d'un groupe de personnes, ou encore par compilations bibliographiques, de nombreuses bases de données spécifiques ont été créées au sein des laboratoires. Certaines sont inconnues ou mal connues et attendent qu'on les exploite davantage.

Les bases de données spécialisées sont d'intérêt divers et la masse des données qu'elles contiennent peut varier d'une base à une autre. Ces bases correspondent à des améliorations ou à des regroupements par rapport aux données issues des bases généralistes.

Exemples de bases de données spécialisées

Late Embryogenesis Abundant Proteins database (LEAPdb - Hunault & Jaspard, 2010) : cette base de données contient des informations sur les protéines LEA impliqués dans la tolérance à de nombreux stress, notament la déshydratation et le froid.

small Heat Shock Proteins database (sHSPdb - Jaspard & Hunault, 2016) : cette base de données contient des informations sur les protéines de choc thermique de faible masse molaire.

RESID Database of Protein Modifications : base de données sur les acides aminés peu fréquents (sous-partie de la base de données PIR).

5. Le consortium de bases de données InterPro

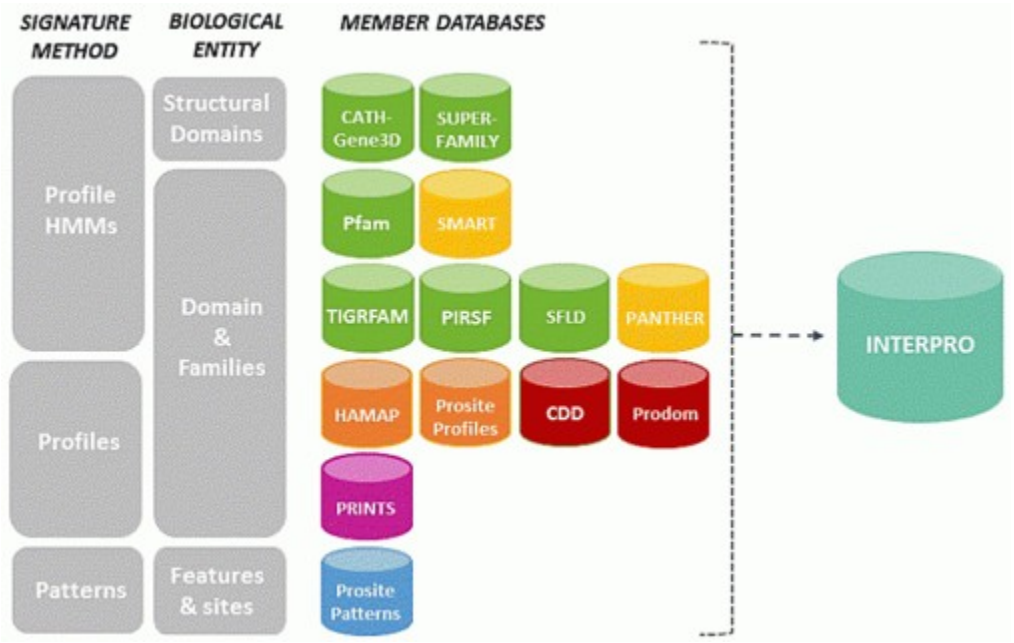
InterPro permet l'analyse de séquences de protéines en les classant dans des familles et en prédisant la présence de domaines et de sites fonctionnels.

InterPro est un consortium : pour mieux classer les protéines, InterPro utilise en effet les modèles ("*patterns*"), les profils ("*profiles*") et les signatures ("*fingerprints*") fournis par 14 bases de données membres (regroupées en une seule ressource) : CATH-Gene3D, SUPERFAMILY, Pfam, SMART, TIGRFAM, PIRSF, SFLD, PANTHER, HAMAP, Prosite, CDD, MobiDB, ProDom, PRINTS.

- Cela permet d'accéder au potentiel de prédiction de ces bases de données sans les consulter individuellement.

- En combinant ces différentes bases de données et les types de signature, InterPro capitalise leurs forces individuelles et fournit un outil puissant pour la prédiction de la fonction des protéines.

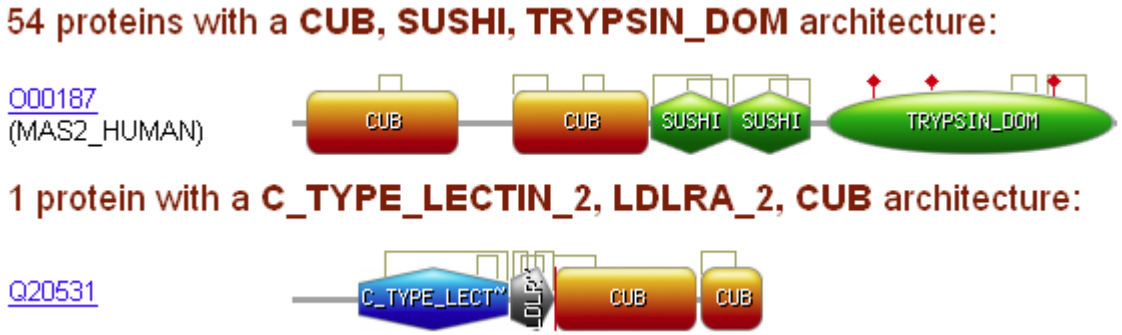
- InterPro simplifie et rationalise l'analyse des séquences des protéines en organisant la somme de toutes les informations de manière cohérente, en supprimant la redondance, en augmentant l'annotation des entrées et en ajoutant des liens vers les signatures et les protéines correspondantes.



Source : InterPro

6. La base de données Pfam (&asymp 18.300 familles - 2020)

La possibilité de diverses combinaisons de multiples domaines explique la très grande multiplicité des protéines. La caractérisation du ou des domaines d'une protéine permet d'en décrypter la/les fonction(s).



Source : Prosite

La base de données Pfam est une collection de familles de domaines des protéines : chaque famille est représentée par des alignements multiples des séquences et un modèle de Markov caché ("*hidden Markov model*" - HMM).

Chaque famille ou entrée Pfam (souvent désignée sous le nom "*Pfam-A entry*") est constituée d'un alignement de séquences généré de la manière suivante :

- On sélectionne un petit nombre de séquences de protéines que l'on considère comme représentatives de la famille Pfam.
- Ces séquences "souches" permettent d'obtenir un alignement de haute qualité ("*curated seed alignment*").
- Un profil HMM est construit avec HMMER à partir de cet alignement de haute qualité.
- Ce profil HMM est utilisé comme modèle pour rechercher les séquences homologues dans les bases de données (par exemple Uniprot).
- Un alignement est généré automatiquement avec toutes les séquences des protéines appartenant à la famille.

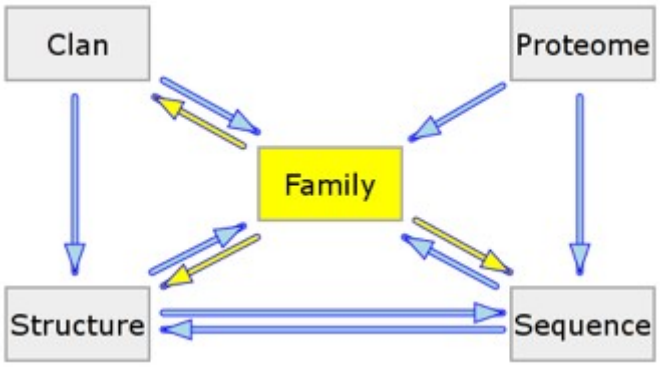
Script python de recherche du profil HMM d'une famille PFAM.

Les entrées Pfam sont classées en **6 catégories**, en fonction de la longueur et de la nature des parties de la séquence inclues dans l'entrée :

- Famille : ensemble de parties de séquences apparentées qui peuvent contenir un ou plusieurs domaines, sans preuve pour affirmer qu'il existe une subdivision. "famille" est la catégorie par défaut.
- Domaine : ensemble de parties de séquences apparentées qui forment une unité structurale.
- Répétition ("*repeat*") : unité courte "instable" tant qu'elle est isolée. Elle forme une structure "stable" quand plusieurs copies sont regroupées.
- Motif : unité courte trouvée dans les domaines non globulaires. Cette unité assure un rôle qui lui est propre (exemple : liaison à un métal).
- Superhélice ("*coiled-coil*") : régions d'une protéine qui contiennent de façon prédominante des motifs en double spirales (hélices alpha enroulées en faisceaux 2-7 - "*helix bundle*").
- Régions désordonnées : régions conservées de protéines avec un biais dans la composition en acides aminés et/ou régions dites intrinsèquement désordonnées ou non structurées.

Plusieurs entrées Pfam liées sont regroupées dans un clan. Leur inter-relation est définie par :

- la similarité de séquence
- la similitude de leurs structures 3D (si elles sont connues)
- la similitude entre leur profil HMM (telle que peut l'évaluer un algorithme comme HHsearch, par exemple)



Source : Pfam

7. Les bases de données de motifs

L'utilisation de bases spécialisées comme les bases de motifs est devenue un outil essentiel dans l'analyse des séquences pour tenter de déterminer la fonction de protéines inconnues ou savoir à quelle famille appartient une séquence non encore caractérisée.

a. Les bases de motifs nucléiques

La plupart de ces bases consiste à recenser dans des catalogues les séquences des différents motifs pour lesquels une activité biologique a été identifiée. Certains motifs sont simples et non ambigus, d'autres correspondent à des activités biologiques plus complexes et engendrent donc des séquences moins précises. Pour ces derniers types de motifs, des compilations ont été établies pour donner des listes annotées de motifs qui peuvent être communs à plusieurs séquences.

Il existe différentes bases de motifs nucléiques, notamment celles concernant les motifs de fixation des facteurs de transcription.

b. Les bases spécialisées de motifs protéiques

La base PROSITE peut être considérée comme un dictionnaire qui recense des motifs protéiques ayant une signification biologique.

Elle est établie en regroupant, quand cela est possible, les protéines contenues dans Swissprot par famille (Exemples : les kinases ou les protéases). On recherche ensuite, au sein de ces groupes, des motifs consensus susceptibles de les caractériser spécifiquement.

La conception de la base PROSITE repose sur quatre critères essentiels :

- collecter le plus possible de motifs significatifs
- avoir des motifs hautement spécifiques pour caractériser au mieux une famille de protéines
- donner une documentation complète sur chacun des motifs répertoriés
- faire une révision périodique des motifs pour s'assurer de leur validité par rapport aux dernières expérimentations

Voir un exemple : motif "*EF-hand*" des protéines fixant le calcium comme la calmoduline par exemple.

c. Exemples de logiciels et bases de données de profils PSSM

Pftools : ensemble d'outils logiciels (« *package* ») pour construire des profils dans le but de rechercher des séquences et les aligner. Parmi ces programmes :

- pfmake construit un profil à partir d'alignements multiples
- pfsearch pour fouiller une base de données de séquences de protéines sur la base d'un profil
- pfscan pour fouiller une base de données de profils sur la base d'une séquence de protéine

PRINTS : base de données de profils PSSM.

- PRINTS fournit des annotations détaillées des familles de protéines et un outil de diagnostic pour les nouvelles séquences.
- PRINTS est une base de données d'empreintes protéiques ("*fingerprints*") : groupe de motifs conservés issus d'alignements multiples de séquences. Ensemble, ces motifs constituent une signature caractéristique de la famille de protéines.

PRINTS est l'un des partenaires fondateurs du consortium de ressources bioinformatiques InterPro (base de données de familles de protéines, de domaines et de sites fonctionnels).

- Quelques bases de données du consortium InterPro : CATH-Gene3D, CDD, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY, TIGRFAMs
- ProDom : collection de motifs protéiques obtenues automatiquement avec PSI-BLAST.