S1133: Bases de données en bioinformatique

Chloé-Agathe Azencott

2016

Objectifs

- Savoir **où et comment** chercher un type spécifique d'information sur **un gène**, **une protéine**, ou **une mutation.**
- Connaître les **différentes bases de données** disponibles.
- Appréhender la **complexité** de l'univers des bases de données bioinformatiques.

De nombreuses bases de données

- **Données publiques** issues de :
  - publications
  - expériences
  - analyses manuelles
  - extraction automatique
  - prédiction automatique.
- souvent **partiellement redondantes.**

Chaque base de données propose (généralement)

- Un accès simplifié via une **interface web :**
  - recherche
  - visualisation
  - lien vers d'autres bases
- Une **API** (Application Program Interface)
- Le **téléchargement** via FTP.

Cas d'étude

Récepteur 2 du facteur de croissance épidermique humain

"Human Epidermal Growth Factor Receptor 2"

Bases de données bibliographiques

- **Pubmed** `http://www.ncbi.nlm.nih.gov/pubmed/`
  Plus de 26 millions de références de la littérature biomédicale.
- Bases de données **non spécifiques :**
  - **Google Scholar** `https://scholar.google.com`
  - **Wikipedia** `http://wikipedia.org`
- **Q.1 Qu'avez-vous appris sur le récepteur 2 du facteur de croissance épidermique humain ?**
- **Q.2 Pouvez-vous trouver un symbole / une abbréviation pour ce gène ?**

Bases de données de gènes

- **NCBI Gene** `http://www.ncbi.nlm.nih.gov/gene/`
  - **nomenclature**
    - **Q.3 Quel est l'identifiant du récepteur 2 du facteur de croissance épidermique humain dans NCBI Gene ?**
    - **Q.4 Dans quelles autres bases de données trouve-t-on ce gène ?**
  - **séquence(s) de référence**
    - **Q.5 Sur quel chromosome est-il localisé ?**
  - **voies biologiques :** voies métaboliques, voies de signalisation
  - **mutations**
  - **phénotypes médicaux**
    - **Q.6 À quelles affections est-il associé ?**

    Peut servir de **point d'entrée** vers de nombreuses autres bases.
- **GeneCards** `http://www.genecards.org`
  - **Q.7 Dans quels compartiments cellulaires notre gène est-il exprimé ?**

Bases de données de gènes

- **Ensembl** `http://www.ensembl.org/`

  annotation de gènes : transcrits, orthologues, ontologies, mutations, régulation, etc.

- **HGNC** (HUGO Gene Nomenclature Committee)
  `http://www.genenames.org/`

  Q.8 Quel est le nom officiel de notre gène ?

Bases de données de séquences d'ADN

- **Visualiser** une séquence génétique
- **"Naviguer"** dans le génome
- **Ensembl** `http://www.ensembl.org`
- **Vega** `http://vega.sanger.ac.uk`
  - Télécharger la séquence
  - Faire une **recherche BLAST**
- **UCSC Genome Browser** `http://genome-euro.ucsc.edu/`
  - Annotations organisées en **"tracks"**
  - **Intégration verticale** d'information
  - Taille de **fenêtre variable**
  - Visualisation de données fournies par **l'utilisateur**.

Bases de données de protéines

- **HPRD** (Human Protein Reference Database) `http://www.hprd.org`
  information sur les **protéines**
  - **Ontologies** :
    - **Q.9 Dans quels processus biologiques ce gène est-il impliqué ?**
    - **Q.10 Quelle est sa fonction moléculaire ?**
  - **Expression protéique**
    - **Q.11 Dans quels tissus ce gène est-il exprimé ?**
  - **Modifications post-traductionnelles (PTMs)**
- **Uniprot** `http://www.uniprot.org`
  - **Q.12 Combien il y a-t-il d'isoformes connus de la protéine codée par notre gène ?**
- **Pfam** `http://pfam.xfam.org/`
  - **Domaines** des protéines
  - **Familles** de protéines
  - Visualisable depuis la fiche Uniprot
    - **Q.13 Combien de domaines ont-ils été identifiés dans notre protéine ?**

Bases de données de structure de protéines

- **PDB** `http://www.rcsb.org/pdb/home/home.do`
  - **Q.14 Combien de structures correspondant à notre protéine sont-elles enregistrées à la PDB ?**
  - **Q.15 Quelles sont les différences entre ces différentes structures ?**
- Protein structure classification
  - **SCOPe** `http://scop2.mrc-lmb.cam.ac.uk/` Structural Classification of Proteins
  - **CATH** `http://www.cathdb.info/` Class, Architecture, Topology, Homology
  - **Q.16 À quelles super-familles la structure PDB 1mfg appartient-elle ?**

Bases de données de voies biologiques

Répertorient les **mécanismes** biologiques (**voies métaboliques**, **voies de signalisation**, **voies de régulation**, etc.)

- **KEGG** `http://www.genome.jp/kegg/pathway.html`
  - **Q.17 Quelles sont les voies biologiques dans lesquelles notre gène est impliqué ?**
- **Reactome** `http://www.reactome.org/`
  - Visualisation dans le Reactome Pathway Browser
- **Wiki Pathways** `http://www.wikipathways.org`
- **Pathway Commons** `http://www.pathwaycommons.org/`
  - Visualisation dans PCViz

Bases de données d'interactions

- Réseaux d'interactions
- **BioGRID** `http://thebiogrid.org/`
  - Construite à partir de la litérature
  - Interactions **physiques**
  - Interactions **génétiques**
  - **Q.18 Combien de molécules interagissent avec notre gène ?**
- **STRING** `http://string-db.org/`
  - Interactions **connues**
  - Interactions **prédites :** fusion dans d'autres espèces ; voisinage conservé ; co-occurence ; co-expression ; extraction automatique d'articles.
  - **Q.19 Combien de gènes interagissent avec le notre ?**
- **HPRD** `http://www.hprd.org/`

Bases de données de phénotypes médicaux

- **OMIM** `http://omim.org/`
  - Informations sur toutes les **maladies génétiques** connues
  - Entrées = gènes ou phénotypes
  - **Q.20 À quelles maladies notre gène est-il lié ?**
- **Orphanet** `http://www.orpha.net/`
  - **Maladies orphelines**
  - **Q.21 Notre gène est-il lié à des maladies rares ?**
- **GeneCards**

Bases de données de mutations

- **dbVar** `http://www.ncbi.nlm.nih.gov/dbvar`
  Variations structurales : Copy Number Variations (CNVs), inversion, délétion, duplication, translocation, etc.
- **dbSNP** `http://www.ncbi.nlm.nih.gov/snp`
  SNPs : Single Nucleotide Polymorphims
- **ClinVar** `https://www.ncbi.nlm.nih.gov/clinvar/`
  - Variants observés
  - Preuves cliniques et expérimentales de leur effet phénotypique
- **HGMD** `http://www.hgmd.cf.ac.uk/ac/index.php`
- **GWAS catalog** `https://www.ebi.ac.uk/gwas/`
  Q.22 À quelles maladies des mutations de notre gène ont-elles été associées ?
- **SNPedia** `http://www.snpedia.com/index.php/SNPedia`
- **DECIPHER** `https://decipher.sanger.ac.uk/` : Données patients partagées.

Bases de données de ligands et médicaments

- **BindingDB** `http://www.bindingdb.org`
- **ChEMBL** `https://www.ebi.ac.uk/chembldb/`
- **DrugBank** `http://www.drugbank.ca/`
  - **Q.23 Quels sont les inibiteurs de notre protéine approuvés pour l'usage clinique ? Par quels laboratoires sont-ils commercialisés au Canada ?**
  - Pour répondre à cette question pour la France, croiser avec `http://base-donnees-publique.medicaments.gouv.fr`.

Quelques exemples d'autres bases de données

- Facteurs de transcription : **TRANSFAC** `http://www.gene-regulation.com/pub/databases.html`, **JASPAR** `http://jaspar.genereg.net/`
- Micro-ARNs : **miRBase** `http://www.mirbase.org/`
- Cancer genomics data : **The Cancer Genome Atlas (TCGA)** `http://cancergenome.nih.gov/`