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.
- Ces bases de données, typiquement construites avec un usage particulier en tête, sont 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) pour interagir automatiquement avec la base de données
- Le **téléchargement** via FTP (File Transfer Protocol) pour l'analyse en local.

Cas d'étude

Nous allons essayer de collecter des informations au sujet du gène appelé
récepteur 2 du facteur de croissance épidermique humain
soit *"Human Epidermal Growth Factor Receptor 2"* en anglais.

Bases de données bibliographiques

- **Pubmed** `http://www.ncbi.nlm.nih.gov/pubmed/`
  Plus de 26 millions de références de la littérature biomédicale, de la base de données MEDLINE, de revues de sciences de la vie, et d'ouvrages en ligne.
- 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/` Base de données du National Center for Biotechnology Information, une division du National Institutes of Health (NIH, institutions nationales pour la recherche médicale et biomédicale des États-Unis). NCBI Gene intègre des informations diverses sur les gènes d'un grand nombre d'espèces :
  - 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 : le génome de référence déterminé en 2000 par le Human Genome Project est toujours sujet à de nouvelles améliorations. En effet, le séquençage de l'ADN humain est fait par morceaux, qui doivent ensuite être assemblés. Le séquençage de morceaux manquants ou la mise au point de nouveaux algorithmes d'assemblage donnent lieu à de nouvelles version du génome de référence (ex. hg18, hg19 ou GRCh37, GRCh38).
    - **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` Comme NCBI Gene, recense et résume des informations variées provenant de diverses autres bases de données.
  - **Q.7 Dans quels compartiments cellulaires notre gène est-il exprimé ?**
- Ensembl `http://www.ensembl.org/`

  une base de données d'annotation de gènes : transcrits, orthologues, ontologies, mutations, régulation, etc.
- HGNC (HUGO Gene Nomenclature Committee) `http://www.genenames.org/` est le consortium chargé de définir un unique nom et symbole pour chaque gène, afin de faciliter la communication entre scientifiques et les requêtes dans les bases de données.

  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 (Basic Local Alignment Search Tool) pour trouver des séquences similaires à celle du gène (ou de la séquence) auquel on s'intéresse.
- 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`
  une base de données d'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 (i.e. régions fonctionnelles)
  - Familles de protéines représentées par des alignements de séquences multiples (en particulier HMMs.)
  - 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 (Protein Data Bank) `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 : Les voies biologiques ("*biological pathways*") représentent une cascade d'interactions moléculaires conduisant à un résultat final. Par contraste, les réseaux d'interactions (PPIN, ou "Protein-Protein Interaction Networks") peuvent représenter toutes les interactions concernant un ou plusieurs gènes. De tels réseaux sont des graphes, dont les nœuds (sommets) représentent une protéine / un gène, et les arêtes représentent une interaction physique, une co-expression, une co-régulation, etc. entre les gènes / protéines qu'elles relient.

- **BioGRID** `http://thebiogrid.org/` The Biological General Repository for Interaction Datasets
  - Construite à partir de la litérature
  - Interactions **physiques**
  - Interactions **génétiques**, quand la mutation ou la sur- ou sous-expression de deux gènes résulte en un phénotype surprenant au vu des effets de chacune de ces modifications considérée individuellement.
  - **Q.18 Combien de molécules interagissent avec notre gène ?**
- **STRING** `http://string-db.org/`
  - Interactions **connues** (listées dans des bases de données validées manuellement ou déterminées expérimentalement)
  - Interactions **prédites :**
    - fusion (les gènes qui sont fusionnés dans d'autres espèces sont susceptibles d'avoir des fonctionalités identiques ou liées.)
    - voisinage conservé (les gènes sont proches sur la séquence génétique dans de multiples espèces, ce qui suggère que leurs fonctions sont liées)
    - co-occurence (protéines qui apparaissent dans la même voie métabolique)
    - co-expression
    - automatiquement extraites d'articles scientifiques
  - **Q.19 Combien de gènes interagissent avec le notre ?**
- **HPRD** `http://www.hprd.org/`

Bases de données de phénotypes médicaux

- **OMIM** Online Mendelian Inheritance in Man `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** : des maladies trop rares pour générer beaucoup d'intérêt de la part des chercheurs, des organismes de financement ou des industries pharmaceutiques.
  - **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, i.e. simples variations génomiques d'une seule paire de bases.
- **ClinVar** (Clinically Associated Human Variations) `https://www.ncbi.nlm.nih.gov/clinvar/`
  - Variants observés
  - Preuves cliniques et expérimentales de leur effet phénotypique
- **HGMD** (The Human Gene Mutation Database) `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 la base de données publique des médicaments `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/`