Master Bioinformatique – Université de Rouen
M1 Semestre 2 – mis à jour janvier 2018
UE UE3 Bioinformatique en sciences omiques
H. Dauchel

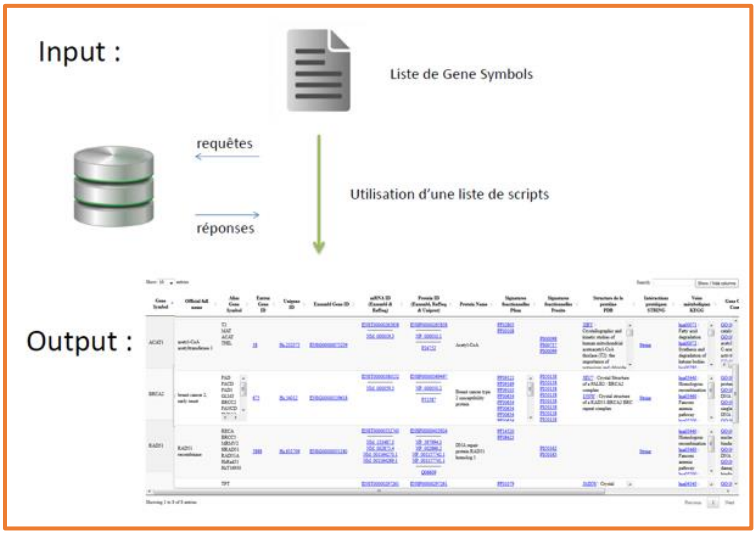
Projet « scripting pour l'agrégation automatique d'annotations »

Modalité

Le travail est individuel et personnel mais les échanges ne sont pas interdits.
Le travail final fera l'objet d'une présentation- démonstration individuelle.
Le travail est réalisé en autonomie avec des points bilans d'avancement.

Cahier des charges

Vous mettrez en place une série de scripts permettant d'agréger à partir d'une liste de gènes d'une espèce donnée, leurs annotations respectives dans un fichier tabulé interactif.



En bleu : données de départ (input)
En rouge : sources primaires
En noir, liste des informations requises dans le tableau final (output)

Informations générales : tout obligatoire

- ✓ **Gene Symbol** (ex : RAD51 sur **Gene, NCBI**) pour l'organisme ***Genre species*** (ex : *Homo sapiens*)
- ✓ Offical full name (Ex : RAD51 recombinase, Gene, NCBI)
- ✓ N° accès du gène : (ex : 5888 sur Gene, NCBI ; sur Ensembl: (ex : ENSG00000051180) + le lien de visualisation sur le genome browser Ensembl
- ✓ Protein names (ex : DNA repair-protein RAD51 homolog 1 sur UniprotKB)
- ✓ N° d'accès protéine(s) (ex : Q06609 sur UniprotKB, NP_001157741.1. sur RefSeq, ENSP0000026786 sur Ensembl)
- ✓ N° d'accès ARN messager (ex : NM_001164269.1. sur RefSeq, ENST00000267868 sur Ensembl)

Annotation fonctionnelle et structurale de la protéine

- ✓ Signatures fonctionnelles (ex : PF08423 domaines protéiques sur Pfam + graphical view, ex : PS50162 motifs et domaines sur Prosite + graphical view)
- ✓ Structure de la protéine (ex : 1b22, N-terminal Domain sur PDB ou autre banque de structures)

Annotation relationnelle :

- ✓ Gene ontologies (ex : Function : transcription; Biological process : DNA damage ; Cellular component : nucleus sur GO)
- ✓ Voies métaboliques (ex : hsa:5888 + les voies hsa03440 Homologous recombination sur KEGG)
- ✓ Interactions protéiques (ex : lien vers String ou autre banque)
- ✓ Orthologues (Ensembl)

Contraintes de la solution

- ✓ **Le modèle de la collecte des annotations est fourni (figure 1). Vous aurez à le précisez avec le nom de vos méthodes et scripts.** L'origine de l'annotation est dès que possible la source primaire de l'information.
- ✓ Pour agréger les annotations, vous utiliserez les connaissances acquises au cours de vos enseignements en réinvestissant obligatoirement une **diversité d'outils de scripting** : programmation PERL et bioPERL ; API (API REST ; API PERL Ensembl, API PERL e-utilities du NCBI) requêtes SQL ; outils de *ID mapping ;* HTML et construction d'URL ; JavaScript …
- ✓ Votre solution devra fonctionner **quelle que soit l'espèce**
- ✓ Le tableau interactif sera construit avec le **plug-in DataTables de la librairie jQuery en Javascript**: il comportera donc les liens .html fonctionnels vers les banques ressources et les autres facilités d'utilisation interactive (fonction de tri, recherche, ascenceur…).

Environnement et phase de travail

- ✓ Phase d'analyse :
  - repérage des liens croisés entre les portails et banques
  - recherche de documentations sur l'accès programmatique aux banques de données
  - repérage du fonctionnement du plug-in DataTables
- ✓ Phase de développement : Mise en œuvre de la programmation des scripts
  - Conseils : (figure 2)
    - Organisation modulaire selon les bases de données
    - Lancement par un script principal
    - Structure de données pour la collecte des annotations

Livraisons

- ➢ **Livrable : mercredi 28 février 2018 20h –** espace dépôt Moodle

Un fichier de la forme Prenom_Nom_Annotation.tar.gz comprenant :

- Un fichier final de votre modélisation de la forme Prenom_Nom_schema_conceptuel.pdf
- Le diaporama de votre présentation (Cf ci-dessous) Prenom_Nom_Presentation.pdf
- Les sources de vos scripts de la forme script.pl
- Un fichier input comportant un exemple d'un petit jeu de données (10 Gene symbols) de la forme Genesymbols.txt
- Un fichier output correspondant de la forme results.html

- ➢ **Présentation individuelle : jeudi 1er mars 2018 -14h-16h et vendredi 2 mars - 10h30-12h30 ; 10 min + 5 min questions**

Pas d'introduction !

1. Votre solution (livrable) :
   - **D1 : Modélisation** > schéma général de votre collecte d'informations d'annotations avec les méthodes
   - **D2 : Organisation du livrable**, ie du fichier source (fichier input, organisation des modules, script principal, fichier sortie) ;
   - **D3 : lancement de la démonstration**
2. **[D4-D9 max] Présentation des portions de votre code**: un exemple pour chaque type de solution de script (API, bioperl, construction d'URL, web services…) avec arguments et paramètres ; structure de données collectées ; création du tableau.
3. **D10 (max) Conclusions/ Auto réflexions**: difficultés et contournements, points positifs/négatifs (complet ou pas, temps d'exécution…).