Lucie TOURNAYRE et Somia SAIDI

Enseignants: Fabien DUCHATEAU et Guillaume LAUNAY

Rapport TER

Etude de faisabilité pour l'automatisation de recherche de jeux de données en bioinformatique

Le but du TER est, à partir d'un fichier d'interactions de paires de protéines, de récupérer d'un côté sur Uniprot et d'un autre côté sur Genbank et Ensembl les informations relatives aux deux protéines en interaction et de les stocker dans une base de données.
Ces jeux de données s'appellent des workflows. Ils consistent, à partir du fichier d'interactions, à retrouver les protéomes puis les génomes des deux organismes interagissant.

Fichier d'interactions

tsv

↓ ↓

Fichier protéique Uniprot (Interactant A) **et** fichier protéique Uniprot (Interactant B)

txt,fasta,xml,rdf *txt,fasta,xml,rdf*

↓

Fichier génomique Genbank (Interactant A **ou** B)

gb, fasta

↓

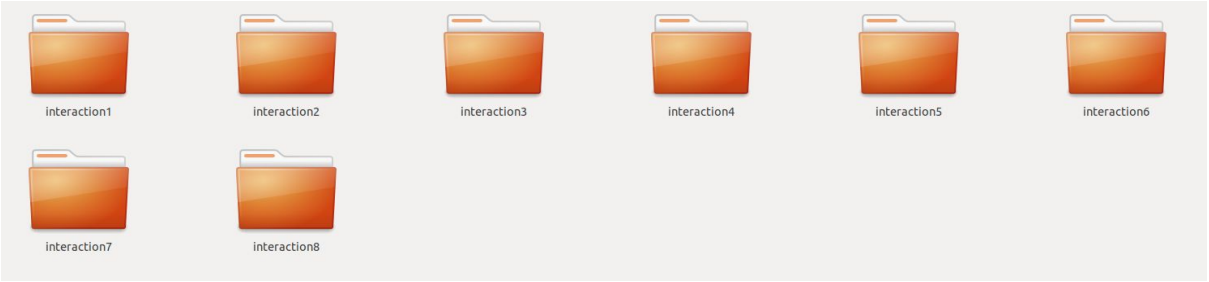
Fichier génomique Ensembl (Interactant A **ou** B)

json

Un workflow complet

Dans notre TER, nous avions pour consigne de nous intéresser uniquement au cas où un des interactant est humain et l'autre non (par exemple humain/virus). Une première sélection est faite.
Nous avons commencé par récupérer les identifiants Uniprot sur le jeu de données de base. Ensuite, à partir de ces identifiants Uniprot nous avons pu retrouver les séquences protéiques et les identifiants Genbank . Une fois sur le site NCBI, nous avons pu télécharger la séquence génomique des interactants en local. Nous avons également récupéré des informations sur les organismes en interaction sur Ensembl.

Toutes les données récupérées sont stockées en local sur l'ordinateur. Cette base de données se présente sous la forme d'un répertoire "dossier_TER". Dans ce répertoire, si on a bien réussi à obtenir les informations issues d'Uniprot, Genbank et Ensembl, des dossiers nommés « interaction 1 », « interaction 2 », etc sont créés pour chaque interaction humain/non humain trouvée.

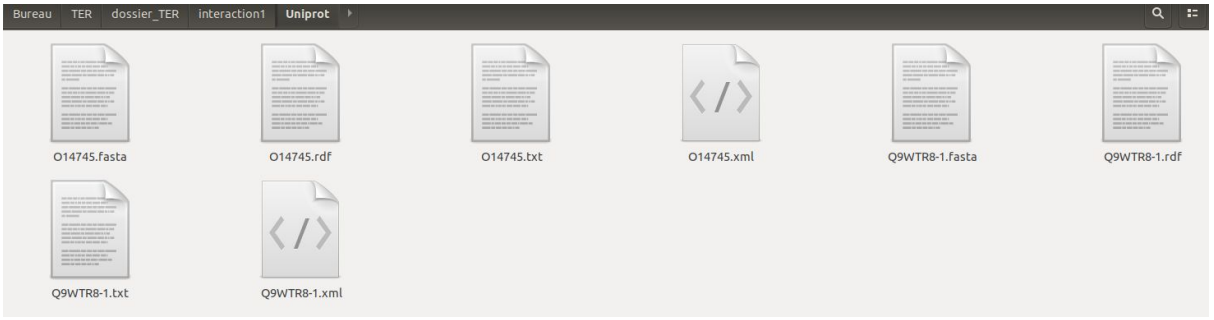


Chaque dossier « interaction » contient un sous dossier « Uniprot », un « Genbank » et un « Ensembl » dans lesquels sont stockées les données recueillis sur les différents sites. Ces sous-dossiers se créent uniquement si les URL construitent, pour accéder à la base de donnée en question, existent.

Nous avons travaillé avec différents formats dans ce TER :

- Fichier d'interaction : formats tsv
- Uniprot : formats rdf, xml, txt et FASTA
- Genbank NCBI : formats FASTA et gb
- Ensembl : format json

Chaque sous dossier contient donc les informations dans tous ces formats, si elles existent. Si un seul de ces formats n'a pas été récupéré, alors le sous-dossier n'est pas créé.



Exemple de contenu d'un sous-dossier Uniprot

Une fois ces données stockées sur l'ordinateur, nous avons mis en place un script supplémentaire afin de donner des statistiques relatives aux interactions intégrées. Ces statistiques sont stockées dans une base de données au format Json stockée en local sur l'ordinateur et nommée « BD_TER ».
Ainsi nous pouvons connaître le nombre d'interactions totales sur le fichier, le nombre d'interactions humain/non humain trouvées, etc. La nature du workflow est aussi indiquée : il peut être complet ou incomplet.
Un workflow complet permet de remonter au protéome et au génome de chaque interactant. Tandis que dans un workflow incomplet, des informations n'ont pas été trouvées que ce soit sur Uniprot, Genbank ou Ensembl.

Un dernier fichier a été créé pour le TER : il s'agit d'un fichier nommé "log_TER.log" dans lequel sont reportées toutes les informations et messages d'erreurs qui ont pu arriver au moment de l'exécution du script python. Nous nous somme particulièrement intéressées à l'erreur de type "Erreur HTTP 404", car nous construisons systématiquement des URL pour récupérer les protéomes et génomes.
Ce fichier log s'incrémente à chaque utilisation.

Connaissances/compétences acquises :

- Parser un fichier au format tsv, parser des pages web
- Découvrir et gérer plusieurs formats de fichiers
- Automatiser une recherche de données

- Passer d'une base de données à une autre (de Uniprot à Genbank par exemple) via des identifiants
- Sérialiser et désérialiser notre base de donnée statistique Json

Problèmes rencontrés :

- Au début nous avions parsé une colonne entière du fichier, alors qu'il fallait parser les protéines par paire seulement. En effet, nous nous intéressons à des interactions binaires.
- On ne faisait pas toujours attention à l'importance des exceptions, les pages web qu'on parse contenant des éléments mis en ligne, des données peuvent être inexistantes
- La récupération du format gb à partir de *Beautiful Soup* a été difficile car il fallait sélectionner du texte entre des balises mais celles-ci n'étaient pas affichées dans le code de la page. De plus, le *National Center for Biotechnology Information (NCBI)* autorise la récupération de leurs données qu'à partir du système global de recherches inter-bases de données *Entrez.*
- La récupération du troisième ID sur Uniprot (celui du gène) a posé problème. On ne parvenait à récupérer que le premier (qui était un ID de transcrit). Pour résoudre ce problème, on a remarqué que les identifiants avaient une petite différence propre à chacun :
    - ENS**T** pour transcrit
    - ENS**P** pour protéine
    - ENS**G** pour gène
    - ENS**E** pour exon

  On a donc récupéré seulement l'ID précédé de ENSG pour avoir l'ID du gène.