Stage M1 2018/2019

Cahier de Labo :

Développement d’un réseau biologique reliant les polluants organiques persistants aux AOPs

ACHEBOUCHE RAYANE

INSERM UMR 1124

Les différents fichiers générés et utilisés pendant mon stage M1 au sein de l'équipe MetaTox (INSERM UMR 1124), depuis le 1 avril 2019 jusqu'au 30 mai 2019, sont dans le répertoire **~Bureau/Rayane**.
Dans ce dossier se trouvent différents répertoires correspondant aux différentes étapes de mon stage. Il s'y trouve également 2 fichiers readme qui informent de la méthode générale adoptée lors du stage ainsi que les versions des logiciels utilisés. La plupart des scripts utilisés et présents dans les répertoires, sont commentés de manière à faciliter leur exécution. Ainsi les ouvrir dans un éditeur de texte type *Gedit*, *Geany*, ou *Notepad++* est indispensable. Nous allons parcourir chaque répertoire pour définir à quoi correspondent chacun des fichiers :

ACP_R :

Le fichier **Liste_CAS_CID_Smiles_Class.csv** renseigne quelles sont les molécules choisies ainsi que leurs identifiants dans les bases de données et le classement utilisé lors des ACP.
Le fichier **SMILES.txt** correspond à la liste des smiles utilisés afin de générer le fichier **SDF_FINAL.sdf** via le web server CACTUS
Listes_R.txt est un fichier facilitant la créaton de certaines listes utilisées dans le script **Script_ACP_FINAL.R.**
Le script **Script_ACP_FINAL.R** a été réalisé sur Rstudio et donc les fichiers **FINAL.Rproj, .RData** et **.Rhistory** lui sont associés. Le script s'ouvre sur RStudio et s'exécute étape par étape selon ce que l'on souhaite effectuer. Chaque partie du script est donc bien séparée.
Les fichiers **.png** générés sont identifiables directement par leur titre et correspondent à des figures générés à l'aide du script mentionné ci-dessus.
Les fichiers Edragon_XXX correspondent aux calculs des descripteurs par Edragon.
Le fichier **.log** associé renseigne sur l'exécution du logiciel, et le fichier **_NH.csv** correspond au fichier résultats de base mais sans en-tête pour une lecture plus facile par le script.

Comptox :

Le fichier **Comptox_ALL_compounds_matrix.ods** est le fichier dans lequel pour chacun des composés l'information Comptox/Toxcast a été compilée.
Le fichier **Comptox_ALL_compounds_matrix_2.csv** correspond à sa conversion en **.csv** avec comme séparateur "**;**" encodé en UTF-8
Le fichier **Script_clean.py** permet de passer de nettoyer le fichier précédent pour sa meilleur exploitation par les scripts de Mme Karine Audouze, et génère donc le fichier final contenant toutes les données => **Tableau_Comptox_final.csv**

CYTO :

Le script **Merge_table.py** permet de créer un fichier exploitable par cytoscape afin de relier des noms de gene à des Event ainsi que des Event à des AOP. Il génère le fichier **Gene_MIE_AOP_Name_2.csv**

Cytoscape :

Ce dossier correspond à la création d'un réseau biologique à partir de l'ensemble des données. Un cutoff sur le score OS a été mis en place afin de filtrer les protéines les plus associées (OS = 9 -> 14)

Le fichier **Cyto_Sess_2.cys** correspond au fichier session de cytoscape qui a permis de générer les différents fichier **.png**

One_Compound :

Ce dossier contient le script **one_comp.py** qui permet de relier un composé à différentes AOPs, sous un même fichier exploitable par Cytoscape à partir des fichiers Con___XXXXX générés par les scripts du dossier SCORING_PRED, et du fichier **Gene_MIE_AOP_Name_2.csv**. Le fichier de sortie généré contient les protéines et les Event, AOP associés pour un composé donné. Dans le dossier figure également un exemple pour l'anthracene, avec les fichiers correspondants et les illustrations.

SCORING_PRED :

Ce répertoire est divisé en trois parties (chacune correspondant à un répertoire) et quelques fichiers. Les fichiers présents sont les fichiers **Prot-POP-unix** qui ont été générés à partir des scripts de Madame Audouze, le fichier **W02_scoring_of_PPI .pdf** qui référence l'utilisation de fonctions de scoring pour la prédiction de score, et qui seront utilisées dans le script **script_scoring.py** afin de générer le fichier **outfile_Scoring.csv**. Ce fichier liste les associations protéine-protéines et les scores prédits associés.

Prot-POP-unix.con = fichier listant les associations prot-prot et les scores associés. Pour plus d'infos sur les scores, voir avec Mme Audouze.

Prot-POP-unix.che = fichier listant par ID pour chaque association prot-prot les composés impliqués dans l'association

À partir de ces scores reliés aux associations on prédit de différentes manières pour un composé X, dans quelles associations peuvent-ils être impliqués

EXTR correspond à la prédiction sur tous les composés de POPs de notre liste :

extr2.py = Script principal qui génère les résultats et les range dans un dossier Results (vérifier que le dossier n'existe pas déjà)

Prot-POP-unix_2.con = Fichier Prot-POP-unix.con auquel on a rajouté les scores prédits à l'aide du script script-scoring.py

PRED correspond à la prédiction sur un seul composé X qu'il fasse parti ou non de notre liste, à partir d'une liste de protéine sous format **.txt** :

pred2.py = Script principal qui génère les résultats dans le dossier courant

Con_XXXXX.con = Fichier résultat généré sur la base du fichier **XXXXXX.txt** qui lui même correspond à la liste des protéines dans lequel le composé **XXXXXX** est impliqué.

Not_Found_XXXXX.txt = Fichier listant les protéines qui n'ont pas été retrouvées comme impliquées dans une association protéine-protéine sur la base du fichier **outfile_Scoring.csv**

TEST_LISTE correspond à la prédiction sur un ensemble de composés contenu dans un même dossier :

List_prot_script2.py = Script principal permettant le calcul des résultats qui seront rangés dans des dossiers séparés portant le nom du composé en question.

Le script ayant été testé sur une liste de composés fournis par Mme Audouze (voir fichier **Liste_composes.csv**), des résultats sont déjà présents. Les composés n'ayant pas d'info sur Comptox/Toxcast sont dans ____No_info, ceux n'ayant pas d'essais actifs sont dans ____Inactive_Cytotox, et ceux ayant des essais actifs mais n'étant impliqués dans aucune protéine sont regroupés dans ____Active_No_prot