Cheryn ALI - M2 Bioinformatique
Octobre 2019

Rapport Projet Intégration de données hétérogènes

Introduction

La génération de données toujours plus nombreuses et diversifiées, notamment en biologie, donne lieu à l'utilisation de méthodes d'intégrations de données afin de conduire à des analyses plus globales et à la compréhension de mécanismes biologiques complexes. Les analyses d'enrichissement sont ainsi couramment utilisées afin de recouper des ensembles d'éléments d'intérêt (Gènes, métabolites... etc) regroupés par différentes fonctions de regroupement (Co-expression, annotation... etc).

Ce projet consiste, dans un premier temps, à générer des données d'annotation de Gene Ontology (GO) des protéines de l'organisme *Solanum tuberosum*. Ce jeu de donnée constitue un premier ensemble de groupe de protéine. Dans un second temps, différente mesures d'enrichissement sont implémentées, comparées et testées sur la GO. Finalement, ces mesures sont utilisées dans le cadre d'une question biologique : les gènes co-exprimés dans les différents tissus de *Solanum tuberosum* sont-ils impliqués dans un processus biologique particulier et/ou sont-ils préférentiellement localisés sur un même chromosome?

I- Génération des données

L'annotation Gene Ontology permet d'associer à un gène ou son produit un ou plusieurs termes GO. Ces derniers indique les différentes informations disponible sur le produit de ce gène concernant son implication dans différents processus biologique, sa localisation cellulaire ainsi que sa fonction moléculaire (1). De multiple outils ont été développés afin de parcourir le graphe de la GO et d'en extraire des informations pertinentes. Dans le cadre de ce projet, STRINGdb et Neo4J sont utilisés. STRING (*search tool for recurring instances of neighbouring genes database*) est une base de données regroupant les différentes associations entre groupes de gènes tel que l'annotation GO, l'interaction protéine-protéine, l'homologie (2). Neo4J est un outil d'exploration et de visualisation de graphe (3). L'utilisation conjointe de ces deux outils nous permet donc d'extraire l'ensemble de l'annotation GO de l'organisme *Solanum tuberosum*. Cet organisme, plus communément connu sous le nom de "pomme de terre", représente un intérêt majeur dans le domaine agroalimentaire en tant que 4e plus importante culture au monde (4).

Méthodologie

La génération des données s'effectue avec le script 01_Generation_sets.r. Ce script parcours le graph de la GO à travers des requêtes Neo4j et génère les deux fichier .set. Ces derniers regroupent d'une part, les annotations et les identifiants externes des protéines qui leur sont directement associé et d'autre part, les annotations et les identifiants externes des protéines qui leur sont directement et implicitement associé. Ce script génère également un fichier .txt rassemblant des informations général sur ces annotations et 4 fichiers RDS. Un second script 02_Analyse_sets.r récupèrent ces fichiers afin de générer des fichiers .pdf contenant les plots de densité des nombres de protéines par annotation GO et de nombre d'annotations GO par protéine (directement et implicitement associé). Ces scripts se lancent en ligne de commande selon l'exemple suivant :

```
Rscript 01_Generation_sets.r -i 4113 -p
/home/guest/Desktop/Cheryn_ALI_Projet_IDH/Premiere_partie/
```

```
Rscript 02_Analyse_sets.r -i /home/guest/Desktop/Cheryn_ALI_Projet_IDH/Premiere_partie/DATA/ -o
/home/guest/Desktop/Cheryn_ALI_Projet_IDH/Premiere_partie/RESULTATS/
```

Ces scripts nécessitent une connexion à Neo4j active avec le graphe de la GO et les annotations de l'espèce d'intérêt intégrés. Ils nécessitent également l'installation préalable des librairies suivante : STRINGdb, tidyverse, neo4r, optparse, psych, ggbubr, ggplot2, ggthemes.

Il est important lors de l'exécution des scripts de respecter l'architecture de dossier suivant:

```
.
├─── Premiere_partie
│    ├─── DATA
│    └─── RESULTATS
```

Résultats

Le génome de S.tuberosum code 39 021 protéines dont seulement 224 sont annotés par au moins un terme GO. De plus, le terme Biological Process annote directement 205 protéines et implicitement 209 protéines. Les courbes de densité du nombre de protéines par terme (fig.1) montre une abondance de terme lié a très peu de protéine. Inversement, les courbes de densité du nombre de termes par protéine (fig.2) montrent une abondance de protéines annotées par plus de 50 GO termes.

De plus, les protéines semblent être directement liées aux nœuds de la GO à différent niveaux d'une même branche (comme schématisé dans la figure S1). Certaines protéines sont également annotées directement par un trop grand nombre de termes. Par exemple, la Cysteine protease inhibitor 1 (external_id = PGSC0003DMT400026285) est annoté par 468 GO terme dans nos résultats alors que UniProt ne référence que 8 annotations. Ceci indique une très mauvaise annotation du génome de la pomme de terre par la Gene Ontology avec la présence de nombreuse redondances.

Les mêmes scripts ont été exécutés sur *Anaplasma marginal*, pathogène bactérien des bovins. 191 protéines parmi 948 sont annotées par au moins un terme GO et les mêmes comportements des nombres de protéine/term et terme/protéines sont observés (les plots correspondant peuvent être consultés dans le dossier Premiere_partie/A_marginal/RESULTATS). De même que précédemment, on constate que le génome de A.marginal est mal annotés par la Gene Ontology.

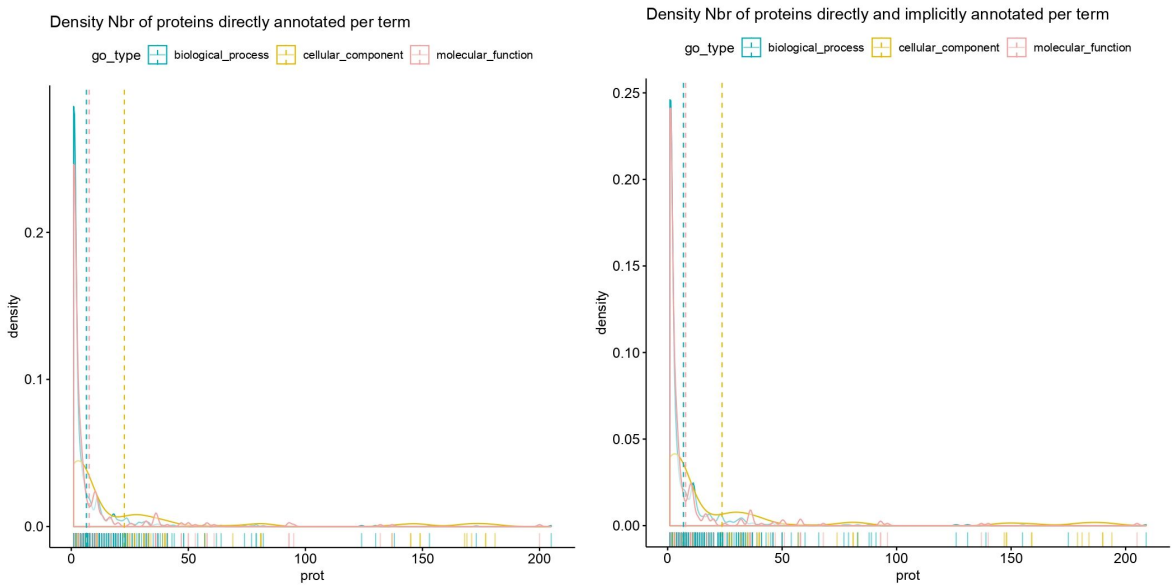

Figure.1 - Courbe de densité des nombres de protéines par GO terme annoté directement (à gauche), directement et implicitement (à droite) chez *Solanum tuberosum* .

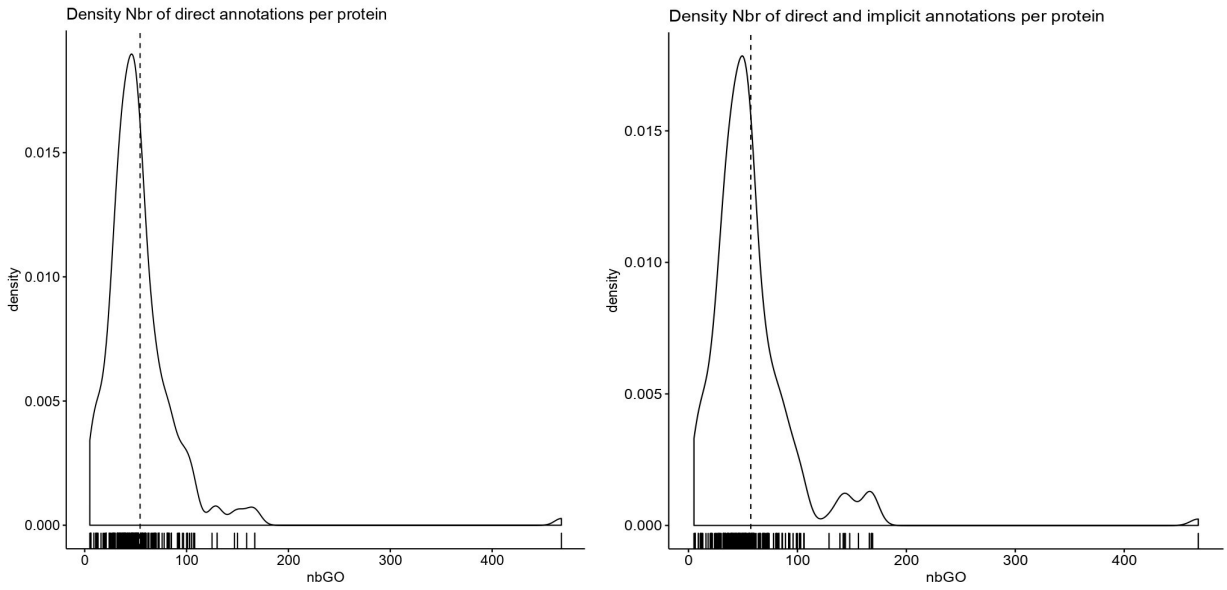

Figure.2 - Courbe de densité des nombres de termes GO par protéines annotés directement (à gauche), directement et implicitement (à droite) chez *Solanum tuberosum*.

II - Comparaison des mesures d'enrichissement

Les analyses de recoupement de voisinage permettent de déterminer si l'intersection entre deux groupe d'éléments est due au hasard ou représente un enrichissement significatif. Il existe plusieurs mesures permettant de tester cette significativité dont la loi hypergéométrique (également nommé test de Fisher), la chi2 et la loi binomial. Ces mesures calculent un p.valeur en se basant notamment sur 4 variables: la taille de la population (g), la taille de l'ensemble requête (q), la taille de l'ensemble cible (t) et le nombre d'éléments en communs. Dans le cadre de ce projet, le comportement du retour de p.valeur de ces mesures est observé en variant ces variables et à travers certain exemples particuliers. Le retour de la mesure de dissemblance coverage est observé de la même façon.

Méthodologie

Dans un premier temps, les mesures coverage et chi2 sont implémenté dans le script 03_blastset.py. Un second script, 04_comparaison_mesures.py, est consacré à la comparaison des mesures en variant q et t de 2 à 50 ainsi que c de 2 à min(q,t) pour un g égale à 250 . Il génère un graphique en 3 dimension dont la couleur est le reflet de la p.valeur (couleur claire pour p.valeur significative). Ces scripts se lancent en ligne de commande selon l'exemple suivant:

```
./03_blastset.py --query 'PGSC0003DMT400081790 PGSC0003DMT400067503 PGSC0003DMT400022652 PGSC0003DMT400001303 PGSC0003DMT400000066' --sets ~/Cheryn_ALI_Projet_IHD/Premiere_partie/RESULTATS/GOterm_prot_dir_Solanum_tuberosum.sets --adjust --alpha 0.05 --measure 'binomial'

./04_comparaison_methodes.py -m binomial -c viridis
```

Finalement un troisieme script 05_cas_particuliers.sh regroupe les commandes shell permetant de d'executer le script 03_blastset.py avec des requêtes choisie afin d'illustrer les cas particulier schématisé à la fig.4. Ces requêtes sont choisie manuellement à partir du fichier GOterm_prot_dir_Solanum_tuberosum.sets ce qui permet de définir la taille de q voulu et une target cible constituant ainsi un résultat "vrais positif".

Les librairies pythons suivantes doivent être préalablement installées: argparse, os, scipy, numpy, matplotlib et mpl_toolkits. Un exemple de sortie du script 03_blastset.py est donné à la fig.S2

Résultats

Les résultats des variations de taille de requêtes, cible et élément en commun pour chaque mesure sont représentés à la fig.3. La mesure coverage ne renvois de résultats satisfaisants qu'en cas de concordance parfaite entre q et t. Chi2 semble être plus permissif que le loi hypergéométrique ou la binomial et renvois davantage de résultats significatif. Ces deux dernière mesures semblent être similaires malgré le caractère légèrement plus restrictif de la binomial.

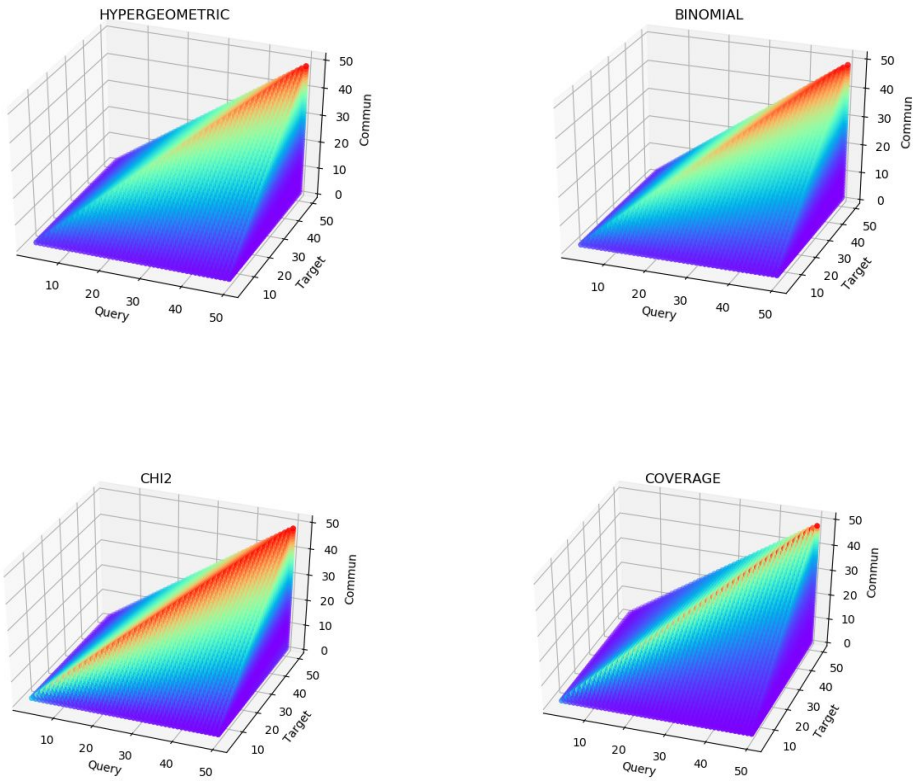

Figure.3 - Représentation graphique des p.values obtenues avec la loi hypergéométrique, la binomial, le chi2 et le coverage, en fonction des taille de groupe requête, cible et d'élément en commun. (couleur rouge = très significatif).

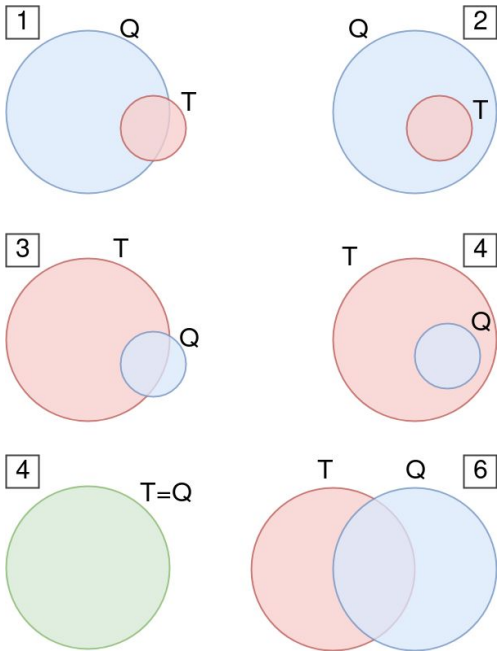

Figure.4 - Représentation schématique des tailles des groupes requête (Q) par rapport au taille de groupe cible (T) d'intérêt des cas particuliers. À noter que le cas 6 représente une cible aléatoire.

L'étude de cas particulier permet d'avoir une vision plus concrète des résultats de ces mesures (les cas étudiés sont schématisé à la fig.4).

Cas 1: Q>>T

Taille Q = 18; Taille T cible (GO:0006563) = 5; Taille éléments en commun = 3.
Aucune des mesures ne renvois de résultats significatif. Ceci s'explique par la taille trop petite du groupe cible et des éléments en commun.

Cas 2: Q⋑T

Taille Q = 15; Taille T cible (GO:0006563) = 5; Taille éléments en commun = 5.
L'ensemble des mesures renvois un résultat significatif pour la cible d'intérêt en premier rang et deux autres groupes. La chi2 est la plus permissive des mesure avec 12 autre resultats significatif renvoyé.

Cas 3 : Q << T

Taille Q = 15; Taille T cible (GO:0097159) = 93; Taille éléments en commun = 5.
Malgré le fait qu'un tiers des élément requête soit inclus dans un même groupe aucune mesure n'a donné de résultats significatif .

Cas 4 : Q ∈ T

Taille Q = 19; Taille T cible (GO:0046872) = 56; Taille éléments en commun = 19.
L'ensemble des mesures renvois la cible d'intérêt en premier rang. 3 autre resultats sont également renvoyé excepté par la coverage.

Cas 5 : Q = T

Taille Q = 5; Taille T cible (GO:0051287) = 5; Taille éléments en commun = 5.
L'ensemble des mesures renvois un résultat significatif pour la cible d'intérêt en premier rang. 3 autre résultats significatifs sont renvoyé par la binomial, 4 autre par l'hypergéométrique et la coverage, 11 autre par la chi2.

Cas 6 : Éléments choisie aléatoirement parmis l'ensemble de la population

Taille Q = 15.
Comme attendu, aucune des mesures ne renvois de résultats significatifs .

Ces constatations ne permettent pas de trancher quand à la mesure la plus adapté à l'utilisation. En effet, en l'absence jeux de données où l'on connaîtrait l'enrichissement "vrais" entre les groupes, il est difficile de comparer ces mesures de manière à désigner celle dont le résultats serait le plus juste.

III- Intégration de données

Les gènes co-exprimés dans les différents tissus de *Solanum tuberosum* sont-ils impliqués dans un processus biologique particulier et/ou sont-ils préférentiellement localisés sur un même chromosome ? Les différentes mesures d'enrichissement sont utilisées afin de fournir des éléments de réponse à cette question.

Méthodologie

Les données utilisées dans cette partie sont des donnée de transcriptions (en FPKM) des gènes à travers différents tissus de S.tuberosum récupéré de la base de donnée Spud (5). Étant donnée le nombre important de gène, de conditions et les capacités limité du matériel informatique à ma disposition, j'ai décidé de concentrer mon analyse sur les 10% de gènes les plus variables (6338 gènes) à travers 9 tissus (fleur, feuille, pétiole, apex de la pousse, tige, stolon, jeune tubercule, tubercule mature et racine).

La première étape consiste à clusteriser les gènes en se basant sur la corrélation de leur expression à travers les tissus. Un clustering hiérarchique est effectué en choisissant la corrélation de Pearson comme indice de similarité et Ward.D2 comme méthode d'agglomération. Le nombre de cluster (k) est fixé arbitrairement à 5 en se basant sur une observation de la heatmap d'expression (fig.5). Une méthode plus rigoureuse consisterait à évaluer le clustering pour différentes valeurs de k et choisir le nombre de cluster en fonction de cette évaluation. Cette méthode n'a pas été implémenté, car elle s'est révélée gourmande en temps de calcul.

Dans un second temps, chaque groupe de gène aurait idéalement été recoupé avec la GO grâce au script 03_blastset.py. Le script 07_test_cluster.sh permet d'automatiser l'ensemble des exécutions. Cette étape ne peut pas être effectué, car le nombre de protéine annotés présent dans les fichiers .sets est largement inférieur au nombre de protéines présent dans chaque cluster.

Finalement, un test d'enrichissement de Fisher est effectué entre chaque cluster et le groupe de gène présent sur chacun des 13 chromosomes de S.tuberosum. Une p.valeur est considérée significative si elle est inférieur à 0.0007692308 (correction par nombre de tests). Un test de chi2 est également effectué pour tester l'indépendance entre la localisation chromosomique et la co-expression.

Le clustering codé dans le script 06_Analyse_Biologique.r est une adaptation du code écrit par Nicolas Chanard, Alexandre Voisin et moi-même dans le cadre du projet tutoré de l'année dernière. Ce dernier génère un fichier texte contenant les noms des gènes de chaque cluster (un cluster par ligne) en plus des multiple plot.

Les analyses R nécessitent l'installation préalable des librairies suivante : tidyverse, gplots, ggplot2, limma, made4 et cluster. La modification chemins stockés dans les variables input_path et output_path est nécessaire en cas d'exécution du script 06_Analyse_Biologique.r .

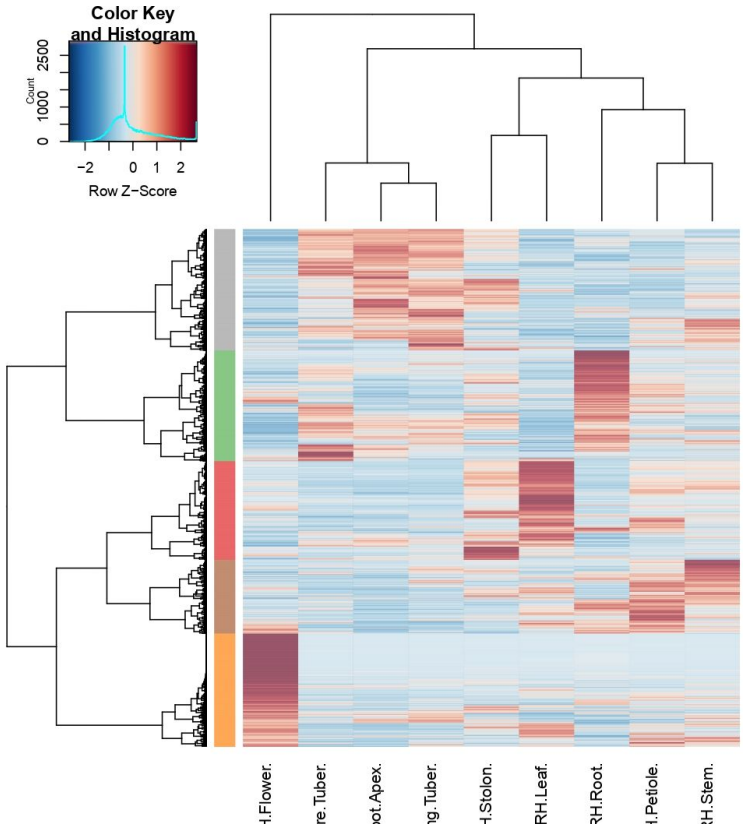

Figure.5 - Heatmap de l'expression de 6338 gènes de S.tuberculosum à travers 13 tissus. Clustering hiérarchique avec (1-corrélation de Pearson) en méthode de distance, Ward.D2 en méthode d'agglomération et nombre de cluster = 5).

Résultats

Les tests de Fisher sont retournés significatifs pour l'intégralité des croisements entre chromosomes et clusters malgré des tailles d'élément en commun très variables (fig.6 et fig.7). Le test de Chi2 est également significatif.

Malgré ces résultats positifs, je pense que nous ne pouvons pas conclure quant à un éventuel lien entre la localisation chromosomique et la co-expression des gènes à travers les tissus de la pomme de terre.

	cluster_1	cluster_2	cluster_3	cluster_4	cluster_5
chr01	1.31686191847422e−249	2.71918166677753e−209	6.80629178857789e−212	0	2.06520629158573e−172
chr02	1.76913190928021e−231	1.87619595952037e−183	2.64272296853514e−176	0	4.72453206545585e−166
chr03	7.85909433163811e−220	1.31923560530156e−184	1.49534674936319e−187	9.26832275765359e−307	5.65131845799111e−161
chr04	2.35881925640106e−211	8.71445264288766e−136	2.96252922237916e−180	2.09067122986488e−274	2.65659932071356e−121
chr05	7.41715497889305e−172	2.05671350937791e−146	9.84860853204406e−122	3.98440687598878e−218	1.35465021036764e−110
chr06	5.26773302299835e−194	2.86133799375714e−170	1.11064024297787e−135	1.22868366594801e−246	3.60038251420469e−148
chr07	1.57329993080267e−149	2.69361799125553e−146	5.25253451843754e−139	9.9651335319281e−242	5.49784113071884e−105
chr08	6.55440817066863e−135	2.76651762752714e−119	3.03790995092074e−118	3.751334595912e−207	6.85251550039649e−105
chr09	6.29321317501997e−153	3.65491024418856e−128	2.47197185839424e−142	1.03527200753122e−229	8.5170830446798e−125
chr10	7.70384893630627e−160	7.41717159009573e−144	1.67529512064708e−115	6.07793228602091e−236	1.08313297164172e−125
chr11	3.68723171591352e−135	5.57924256886988e−114	2.32460666306876e−108	1.16277400860782e−197	1.35355394476712e−89
chr12	5.98284602885193e−172	1.15202687684912e−154	1.88875806645209e−111	7.64423392914566e−217	9.92399099729061e−110
chr00	6.03477857016443e−37	4.09539290146117e−32	1.09916846669048e−41	1.26911397252899e−55	1.21556114858893e−51

Figure.6 - Tableau de p.valeurs des test de Fisher.

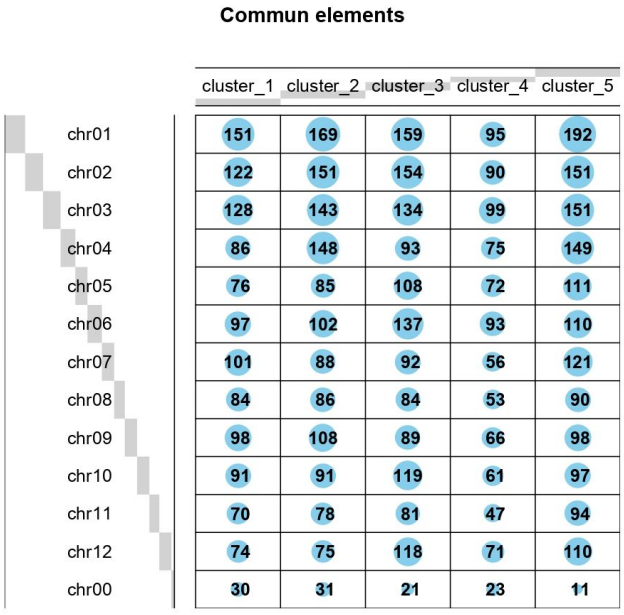

Figure.6 - Tableau de nombre d'éléments en communs.

Conclusion

Comme l'a si bien exprimé Mr Gilles Richard "Data is Power" mais le véritable défi réside dans la compréhension de ces données dans toutes leurs complexités et interconnexions. Les mesures de recoupement de voisinage sont des outils participant à établir des liens entre données de différents type. Cependant, comme nous l'avons vu au travers de ce projet, ces mesures sont sensible à de nombreux paramètre et leurs résultats doivent être interprété avec prudence.

Références

(1) http://geneontology.org/docs/go-annotations/

(2) STRING: a database of predicted functional associations between proteins. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. Nucleic Acids Res. 2003 Jan.

(3) https://neo4j.com/

(4) http://solanaceae.plantbiology.msu.edu/index.shtml

(5) http://solanaceae.plantbiology.msu.edu/pgsc_download.shtml

Figures supplémentaires

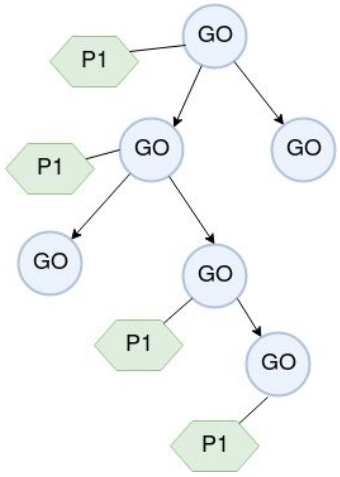

Figure.S1 - Schéma de redondance d'annotation GO (P1 = protéine 1).

```
(devenv) [guest@pc9 Seconde_partie]$
(devenv) [guest@pc9 Seconde_partie]$
(devenv) [guest@pc9 Seconde_partie]$ ./03_blastset.py --query 'PGSC0003DMT400081790 PGSC0003DMT400067503 PGSC0003DMT400022652 PG
SC0003DMT400001303 PGSC0003DMT400000066' --sets ~/Desktop/Cheryn_ALI_Projet_IHD/Premiere_partie/RESULTATS/GOterm_prot_dir_Solanu
m_tuberosum.sets --adjust --alpha 0.05 --measure 'hypergeometric'
GO:0051287      1.3057253890805872e-09  5/6     molecular_function: NAD binding PGSC0003DMT400022652, PGSC0003DMT400081790, PGSC
0003DMT400000066, PGSC0003DMT400067503, PGSC0003DMT400001303
GO:0050662      6.535155572348349e-07   5/15    molecular_function: coenzyme binding    PGSC0003DMT400022652, PGSC0003DMT4000817
90, PGSC0003DMT400000066, PGSC0003DMT400067503, PGSC0003DMT400001303
GO:0016614      1.5468928684437333e-05  4/11    molecular_function: oxidoreductase activity     PGSC0003DMT400081790, PGSC0003DM
T400067503, PGSC0003DMT400001303, PGSC0003DMT400000066
GO:0016616      1.5468928684437333e-05  4/11    molecular_function: oxidoreductase activity     PGSC0003DMT400081790, PGSC0003DM
T400067503, PGSC0003DMT400001303, PGSC0003DMT400000066
GO:0048037      1.7568535110079323e-05  5/27    molecular_function: cofactor binding    PGSC0003DMT400022652, PGSC0003DMT4000817
90, PGSC0003DMT400000066, PGSC0003DMT400067503, PGSC0003DMT400001303
(devenv) [guest@pc9 Seconde_partie]$
```

Figure.S2 - Sortie console de la requete ./03_blastset.py --query 'PGSC0003DMT400081790 PGSC0003DMT400067503 PGSC0003DMT400022652 PGSC0003DMT400001303 PGSC0003DMT400000066' --sets ~/Desktop/Cheryn_ALI_Projet_IHD/Premiere_partie/RESULTATS/GOterm_prot_dir_Solanum_tuberosum.sets --adjust --alpha 0.05 --measure 'hypergeometric'