Université de Bordeaux

Rapport De Projet

Master Bio-informatique

Clusterisation de protéines semblables

Auteurs :
Lilia AHMED ZAID
Franz CERIL
Thomas RIQUELME
Amel YAHOU

Superviseur :
Pascal DESBARATS

14 décembre 2016

Table des matières

Introduction

Le nombre de protéines à étudier dans un organisme entier est trop important pour pouvoir examiner expérimentalement chaque protéine. Par ailleurs des outils bioinformatiques permettent de généraliser les résultats obtenus pour une protéine à d'autres protéines semblables.
Les étapes pour trouver des protéines homologues peut nécessiter l'utilisation de plusieurs outils bioinformatiques (ex : alignement de séquences, conservation de domaines au cours de l'évolution, comparaison de structure etc). A ce stade, le data mining peut faciliter l'étude des protéines semblables en allant fouiller dans des données récupérées sur une base de données. Il permet de structurer les données en fonction de l'analyse que l'on souhaite effectuer et de découvrir des motifs inattendus sur un grand jeux de données.

Dans le cadre de notre étude, nous allons essayer de répondre à la problématique suivante : Quelles sont les protéines semblables ?

Nous avons fait le choix d'affiner cette problématique pour chercher les protéines semblables entre deux espèces : l'homme (Homo Sapiens) et la souris (Mus Musculus).

En effet pour la recherche, il peut être intéressant d'identifier des protéines semblables entre l'homme et la souris afin de conduire des expériences sur la souris, une espèce modèle, et faire des découvertes potentiellement applicable à l'homme ensuite.

1 Analyse du sujet

1.1 Récupération des données

Pour identifier les protéines semblables entre l'homme et la souris, nous avons d'abord téléchargé toutes les protéines de ces deux espèces sur la base de données protéique UniProt, en sélectionnant un large panel de critères (Figure 1, page 3).

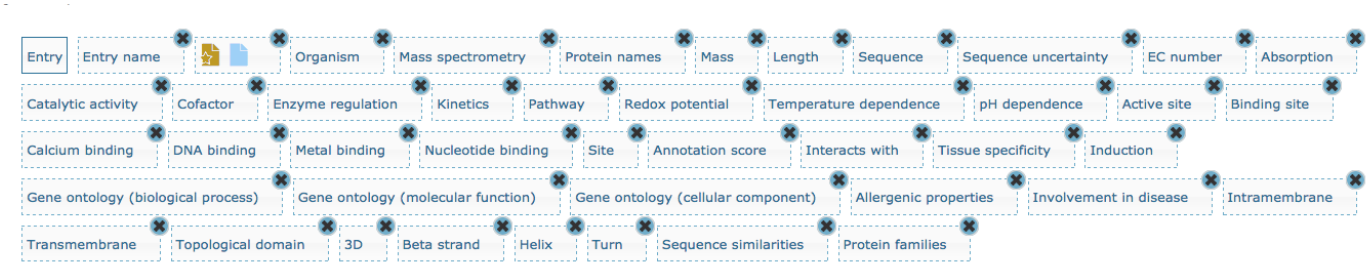

FIGURE 1 – Critéres pré-selectionné sur Uniprot

1.2 Test de représentativité des critères

Un test de représentativité a été réalisé sur notre jeu de données pour chaque critère préalablement sélectionné (sur Uniprot) afin de vérifier que ceux-ci sont bien renseignés dans nos données avant de faire notre choix. Nous n'avons gardé que les critères ayant un taux de représentativité supérieur à 50% (Figure 2, page 3).

```
Entry = 100.0 %
Entry_name = 100.0 %
Status = 100.0 %
Protein_names = 100.0 %
Gene_names = 98.253062549 %
Organism = 100.0 %
Length = 100.0 %
Mass = 100.0 %
Sequence = 100.0 %
Annotation = 100.0 %
Gene_ontology(biological_process) = 81.0216609427 %
Gene_ontology(molecular_function) = 76.1107655696 %
Gene_ontology(cellular_component) = 90.4702669082 %
Sequence_similarities = 91.4221585224 %
Protein_families = 71.0484328943 %
```

FIGURE 2 – Représentativité des critères sélectionnés

1.3 Sélection des critères

On peut classer les protéines selon différents critères : homologies de séquences, caractéristiques physico-chimiques, taille, masse ...etc.
Plutôt que de les classer en fonction de leur homologies de séquences, de caractéristiques physico-chimiques ou structurels, nous avons préféré nous basé sur un classement orienté "fonctions". Nous avons donc choisi de baser notre analyse sur deux critères principaux qui sont : la **Gene Ontology** qui représente le processus biologique dans lequel est impliqué la protéine et la **famille de protéines** à laquelle appartient la protéine. En dernier critère nous avons choisi l'organisme auquel appartient la protéine pour distinguer les protéines de l'Homme de celle de la Souris.

Nous avons ordonné nos critères selon la **priorité** (Figure 3, page 4) suivante :

1. Gene Ontology
2. la famille de protéines
3. l'organisme

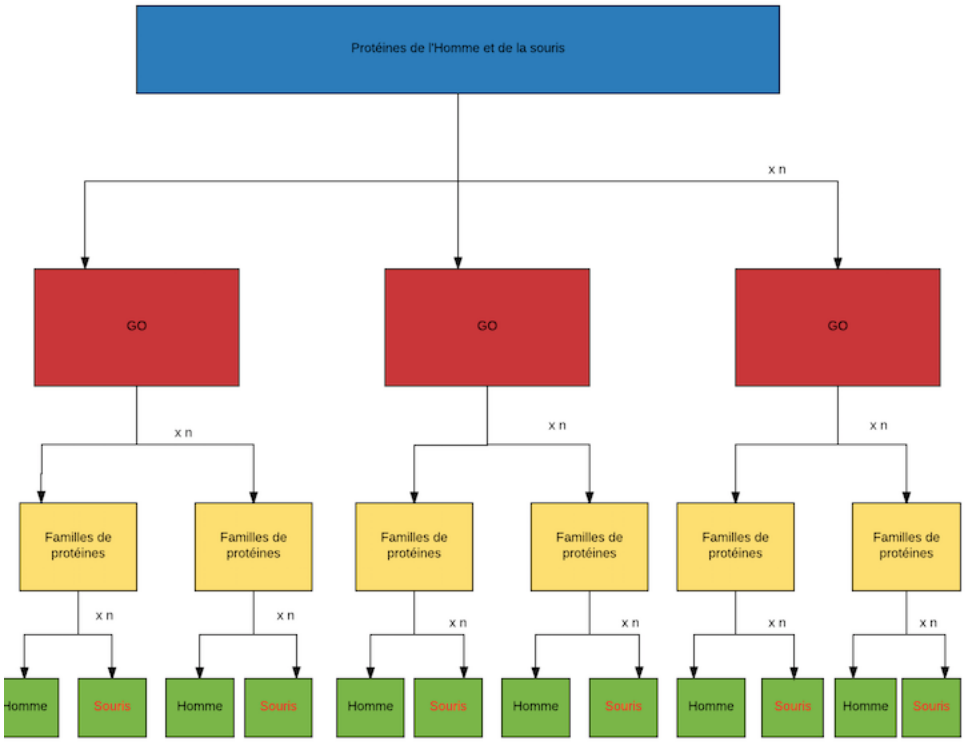

FIGURE 3 – Schéma représentant l'ordre de priorité des critères

2 Implémentation

Pour réaliser nos clusters selon nos critères "processus biologique", "famille de protéines" et "organisme", un clustering hiérachique est appliqué suivant la méthode DIvisive ANAlysis (DIANA). Cette méthode est une approche Top-Down de Clustering. C'est-à-dire qu'au départ nous possédons un ensemble de protéines de l'homme et de la souris, puis cet ensemble va être filtré par nos critères dans l'ordre de priorité choisi (1.processus biologique, 2.famille de protéines et 3.organisme) comme représenté sur la Figure 3.

En premier, le critère "processus biologique" est appliqué. Les protéines sont regroupées en clusters correspondant aux processus biologiques dans lesquels elles rentrent en jeu. Pour enregistrer ces clusters, la structure de données choisie est un dictionnaire qui aura comme clés le nom des processus biologiques et comme valeurs associées le nom des groupes de protéines qu'il contient (les familles).

Puis en deuxième, le critère "famille de protéines" est appliqué. Nos clusters correspondants aux processus biologiques des protéines vont être subdivisés en clusters correspondant aux familles auxquelles appartiennent nos protéines. De même, pour enregistrer ces sous clusters, des dictionnaires imbriqués dans les précédents sont utilisés. Ceux-ci comporteront comme clés le nom des familles de protéines et comme valeurs le nom des protéines appartenant à ces familles.

Enfin en troisième, le critère "organisme" est appliqué. Les protéines appartenant à une même famille au sein d'un même processus biologique sont distinguées en fonction du type d'organisme auquel elles appartiennent : homme ou souris. Pour enregistrer à quel organisme appartient une protéine, un dictionnaire séparé est utilisé et non pas imbriqué dans les dictionnaires précédents. Cela s'explique par le fait que l'on ne veut pas créer de sous groupes "homme" et "souris" dans le dictionnaire des familles de protéines. Celui-ci contiendra directement les protéines on se servira du dictionnaire des organismes pour rajouter ajouter un attribut "organisme" a nos protéines dans notre visualisation.

Pour coder cela, l'algorithme décrit ci-dessous a été appliqué.

2.1 Parcours des lignes du fichier de données

Au préalable un tableau de données a été récupéré sur la base de donnée Uniprot. Celui contient les caractéristiques de toutes nos protéines pour nos différents critères. A l'aide d'un script python, ce tableau est parcouru ligne par ligne pour chaque protéine (une ligne correspond aux caractéristiques d'une protéine).

2.2 Traitement ligne par ligne

2.2.1 Stockage du nom, de la famille, du processus biologique et de l'organisme dans des variables

La ligne de donnée est splitée pour séparer les données de chaque colonne.

Le contenu des colonnes correspondant au nom de la protéine, à la famille de la protéine, aux processus biologiques et à l'organisme auquels appartient la protéine est stocké respectivement dans des variables le temps de traiter une ligne. Il faut noter qu'une protéine appartient généralement

à plusieurs processus biologiques, donc le contenu de cette colonne est splité pour séparer tous les processuss auquels appartient la protéine, et ces processus sont stockés dans une liste de processus.

2.2.2 Ajout des protéines dans des dictionnaires imbriqués

Pour chaque processus dans lequel la protéine est impliquée, si ce processus n'existe pas dans le dictionnaire des processus il est ajouté comme clé. Comme valeur, il lui est attribuée un dictionnaire contenant comme clé la famille de la protéine impliqué dans ce processus. Enfin la protéine est ajoutée comme valeur au dictionnaire de la famille.

2.3 Répétition de ses étapes pour les lignes suivantes pour compléter les dictionnaires

A chaque nouvelle ligne correspondant à une nouvelle protéine, les nouveaux processus dans lesquels elle est impliqué sont ajoutés au dictionnaire des processus. La famille à laquelle appartient cette nouvelle protéine est ajouté comme clé aux dictionnaires des familles pour chacun de ces processus et la protéine est stockée comme valeur à chaun des dictionnaires de familles.

A la fin du traitement de toutes les lignes du fichier, le dictionnaire des processus contiendra tous les processus dans lesquels nos protéines rentrent en jeu. A l'intérieur de chacun des processus se trouvera tous les dictionnaires des familles impliquées dans ces processus et les protéines appartenant à chacune de ces familles.

2.4 Ecriture des résultats de clustering dans des fichiers textes

Le contenu de nos dictionnaires imbriqués a été écrit dans un fichier texte selon un format qui permettra une visualisation des clusters sous forme de graphe dans le logiciel Cytoscape. Ce logiciel reconnait en effet des tableaux à trois colonnes dont la première colonne décrit le noeud source, la seconde le noeud cible et la troisième le type d'intéraction qui les relie.

La première partie de nos résultats décrit la première partie de l'arbre. La première colonne (le noeud source) correspond au cluster général contenant tous les processus biologiques, toutes les familles et toutes les protéines ; la seconde colonne(les noeuds cibles) correspond aux différents processus biologiques reliés au cluster général ; et la troisième colonne correspond au lien général qui relie le cluster général à tous les processus biologiques.

Ensuite, la seconde partie de nos résultats décrit la seconde partie de l'arbre, c'est-à-dire les subdivisions des clusters "processus biologiques" en clusters de "familles de protéines". La première colonne contient donc les processus biologiques (noeuds sources), la seconde les familles de protéines (noeuds cibles), et la troisième le lien biologique qui relie les familles de protéines impliquées dans un même processus biologique.

Enfin, la troisième partie de nos résultats décrit la dernière partie de l'arbre, c'est-à-dire la subdivision des clusters familles de protéines en toutes les protéines qui les constituent. La première colonne contient donc les familles de protéines (noeuds sources), la seconde les protéines qui les consituent (noeuds cibles) et la troisième le lien familial qui les relie.

D'autre part, le contenu du dictionnaire associant à une protéine l'organisme à laquelle elle appartient a été écrit dans un second fichier. Celui-ci contiendra la liste de toutes les protéines et le type d'organisme auquel elles appartiennent (homme ou souris). De même le contenu de ce fichier prendra la forme d'un tableau à deux colonnes (une colonne protéine et une colonne organisme) reconnaissable par cytoscape pour spécifier un attribut organisme aux noeuds des protéines (cela permettra par la suite de leur donner des symboles différents pour plus de clarté au lieu de relier la protéine à l'espèce par une arête).

3 Résultats

3.1 Résultats du clustering hiérarchique

Notre analyse sur un jeu de données de 36979 protéines par clustering hiérarchique a aboutit à 14518 clusters de processus biologiques de la Gene Ontology, regroupant 130971 clusters de familles de protéines imliquées dans ces processus biologiques, eux mêmes regroupant 274285 protéines appartenant à ces familles impliquées dans ces processus.

Ce résultat comporte un trop grand nombre de données pour être visualisé sur un graphe. Un choix a été fait de représenter un échantillon de ces résultats avec le logiciel Cytoscape pour visualiser les clusters sous forme de graphe. Cet échantillon a été choisi arbitrairement a partir du clustering des 100 premières protéines de notre jeu de données pour voir dans quels processus elles étaient impliquées et à quelles familles de gènes elles appartenaient. Ce choix a été fait arbitrairement car on ne sait pas ce que l'on cherche dans le graphe, mais on peut imaginer que des chercheurs puissent utiliser nos données pour cibler un processus biologique et ainsi visualiser les familles de protéines et les protéines impliquées dans le processus d'intérêt pour leur recherche.

Après un test de construction du graphe le jeu de données résultant était toujours trop important pour être bien visualisé. Cela est dû au fait que les 100 premières protéines de notre jeu de données étaient déjà impliquées dans 1600 processus biologiques, ce qui aboutit à un graphe trop fourni. Pour obtenir un nombre de données résultants plus petit et bien visualisable un second filtre a été appliqué pour visualiser que les protéines et familles découlant des 100 premiers processus biologiques dans lesquelles nos 100 premières protéines étaient impliquées.

On obtient les graphes représentés ci-dessous. Premièrement un graphe représentant uniquement les clusters processus biologiques et les protéines impliquées dans ces processus. Deuxièmement un graphe représentant uniquement les clusters familles de protéines et les protéines appartenant à ces familles. Et enfin troisièmement un graphe représentant les clusters processus biologiques, les sous clusters familles de protéines et les protéines impliquées dans ces processus.

3.2 Clusters processus biologiques de la Gene Ontology

Le critère de la Gene Ontology (GO) regroupe les protéines qui sont impliquées dans les mêmes processus biologiques. Nous pouvons observer à travers nos résultats de visualisation du cluster (Figure 4, page 9) que la relation entre les protéines et les processus biologiques de la GO sont complexes.

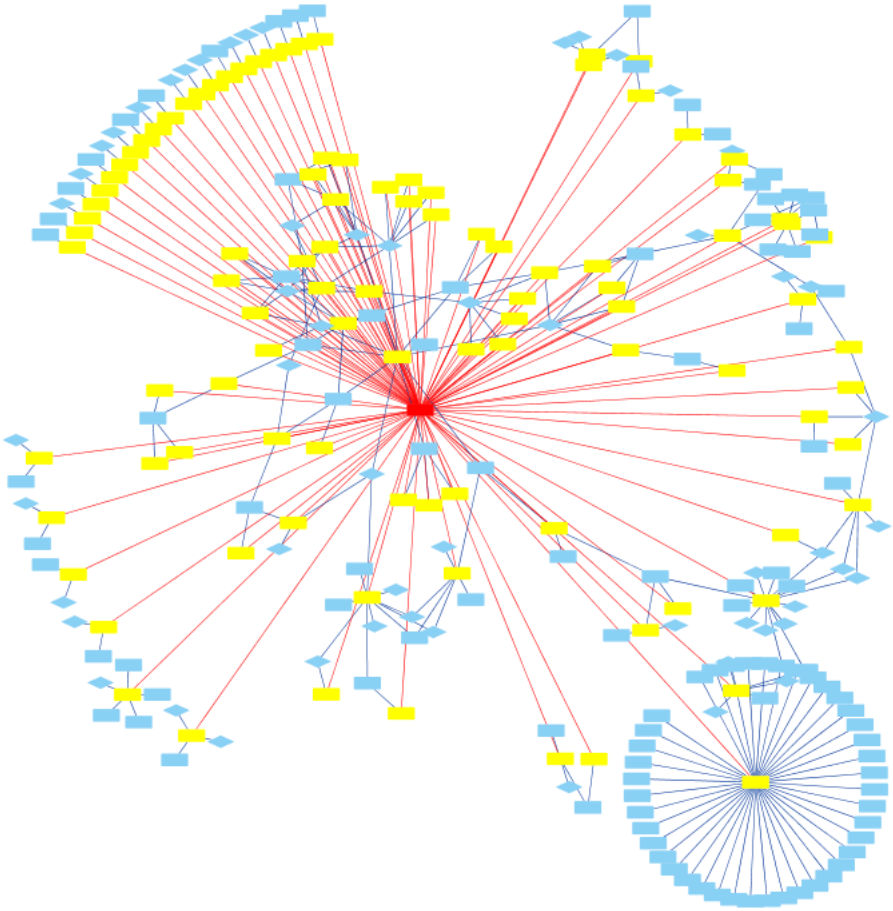

FIGURE 4 – Visualisation des clusters des processus biologiques de la Gene Ontology avec Cytoscape

Le graphe obtenu n'est pas un arbre comme on aurait pu l'imaginer. Si les processus biologiques (représentés en jaunes) sont issus d'un même noeud père (le cluster général représenté en rouge), toutes les protéines (représentées en bleu) ne découlent pas uniquement d'un même processus (père). En effet, il existe des protéines qui peuvent être impliquées dans plusieurs processus biologiques différents, par exemple la protéine **LRRK2** chez l'Homme est impliquée dans trois processus biologiques différents (Figure 5, page 10). Ce qui explique la complexité apparente du graphe.

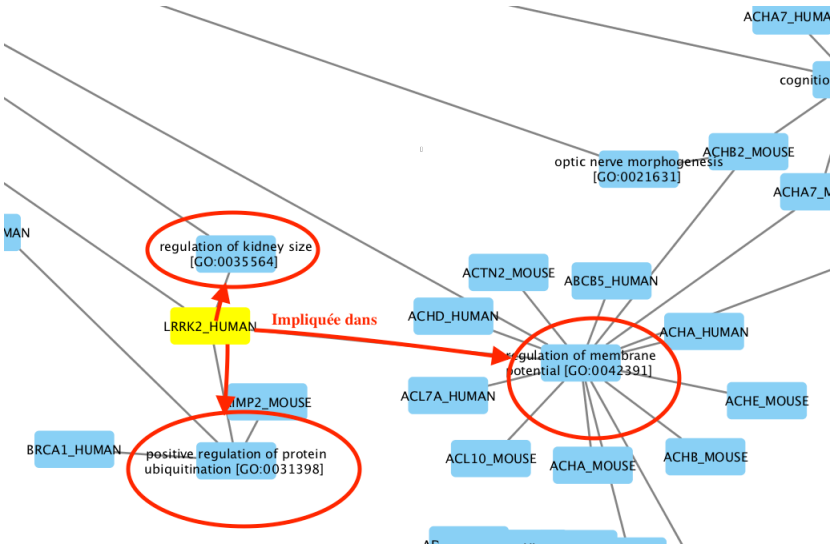

FIGURE 5 – Visualisation des processus biologiques de la protéine LRRK2 chez l'homme avec Cytoscape

3.3 Clusters familles de protéines

Le clustering pour les familles de protéines aboutit à la construction d'un arbre. En effet, contrairement au clustering pour les processus biologiques, ici les protéines ne peuvent appartenir qu'à une seule famille de protéine (Figure 6, page 10).

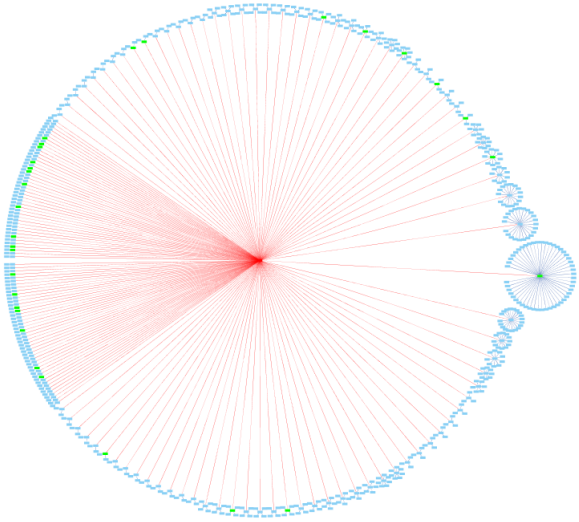

FIGURE 6 – Visualisation des clusters de familles de protéines avec Cytoscape

On observe que la plupart des familles de protéines regroupent à la fois des protéines de l'homme (représentées par des carrés bleus) et de la souris (représentées par des losanges bleus) : ce sont des protéines semblables selon le critère de la famille de protéine (Figure 7, page 11).

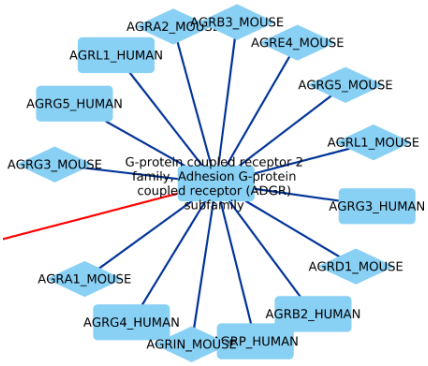

FIGURE 7 – Visualisation de la famille ADCR regroupant des protéines de l'homme et de la souris avec Cytoscape

Cependant il existe quelques familles de protéines qui sont spécifiques à l'une des deux espèces, par exemple, la famille **MHC** (Figure 8, page 11) est spécifique à l'Homme, ce qui est tout a fait normal car il s'agit des protéines du Complexe Majeur d'Histocompatibilité du système immunitaire spécifique à l'homme (mais il est présent sous un autre nom, le Complexe H-2 chez la souris).

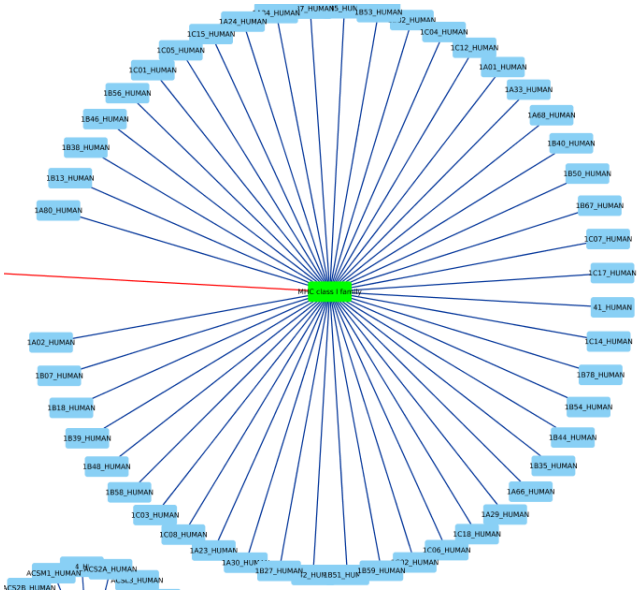

FIGURE 8 – Visualisation de la famille de protéines MHC chez l'Homme avec Cytoscape

Le même exemple est applicable à la souris avec la famille **AF4** (Figure 9, page 12) qui regroupe des protéines également impliqué dans le système immunitaire de la souris.

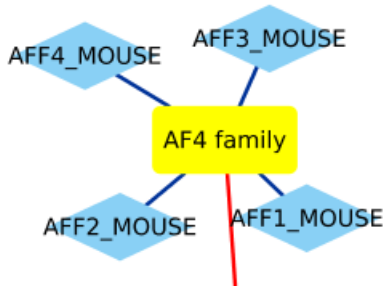

FIGURE 9 – Visualisation de la famille de protéines AF4 chez la Souris avec Cytoscape

3.4 Clustering hiérarchique final

Notre clustering hiérarchique final aboutit au graphe suivant : Figure 10, page 12

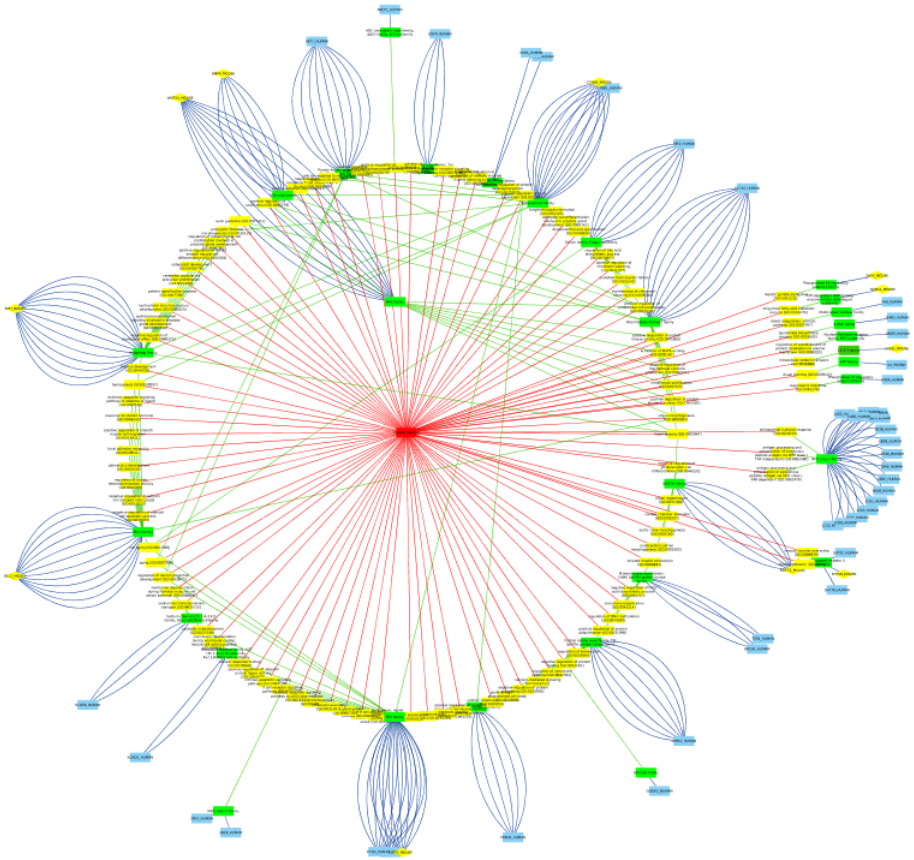

FIGURE 10 – Visualisation des processus biologiques (GO) et des familles de protéines

Premièrement, le cluster général (représenté par le noeud rouge central) est divisé en clusters "processus biologiques" (représentés par des noeuds jaunes). Puis, les clusters processus biologiques sont subdivisés en clusters familles de protéines (représentés par des noeuds verts). Enfin, les clusters familles de protéines sont divisés en toutes les protéines appartenant à une famille spécifique au sein d'un processus biologique spécifique (représentés par des noeuds bleus).

Cela permet d'observer par exemple que pour les processus biologiques "contractraction du muscle lisse" et "comportement d'alimentation" la famille de protéine "G-protein-coupled Receptor 1" est impliquée. Cette famille comprend deux protéines humaines et une protéine de souris que l'on pourrait qualifier de semblables car elles sont impliquées dans les mêmes processus biologiques et appartiennent à la même famille de protéines (Figure 11, page 13).

Figure 11, page 13

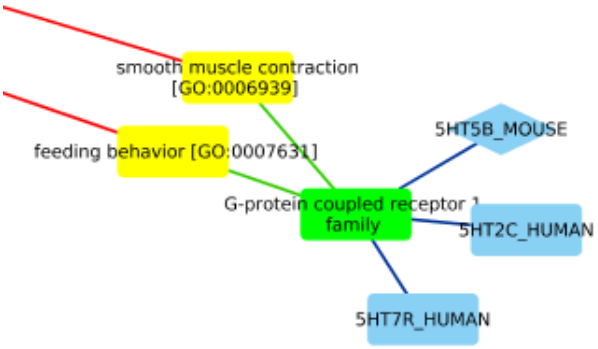

FIGURE 11 – Exemple de protéines de l'Homme et de la souris de la famille des protéines G impliqués dans les processus de contraction musculaire et dans le comportement alimentaire visualisé avec Cytoscape

Bilan

A l'issu de notre projet nous avons pu répondre à la problématique suivante : **quels sont les protéines semblables entre l'Homme et la Souris ?**
Nous avons pu regrouper les protéines semblables entre ces deux espèces en différents clusters selon l'ordre des critères que nous avons appliqué à nos clusters. Nous nous sommes basés sur la méthode de clustering hiérarchique DIANA qui nous a permis d'obtenir les protéines semblables entre les deux espèces qui partagent le même processus biologique et qui appartiennent à la même famille de protéines.
Notre analyse des protéines semblables est basée sur une approche fonctionnelle qui résulte des critères que nous avons choisis (processus biologiques de la Gene Ontology, famille de protéines et organisme).

La recherche de protéines semblables représente un enjeu important, cela permet d'extrapoler sur les propriétés des protéines qui ont été peu ou pas étudiées et ainsi de faire des avancées considérables par exemple dans la recherche de nouveaux traitements pour l'Homme testés chez la souris.

Une utilisation possible de notre clustering pourrait être pour les chercheurs de cibler un processus biologique qui les intéresse pour leur recherche et de visualiser grâce à notre script dans Cytoscape les familles de protéines et les protéines impliquées dans ce processus biologique d'intérêt.

Pour une analyse plus globale des protéines semblables, nous pourrions envisager de clusteriser nos protéines avec plus de critères. Cependant certains critères sont difficiles à traiter par rapport à la masse d'informations qu'ils regroupent et sont en l'occurrence difficiles à visualiser sous un format qui permet de regrouper toutes leurs informations sans être dépassés par la quantité de ces dernières.