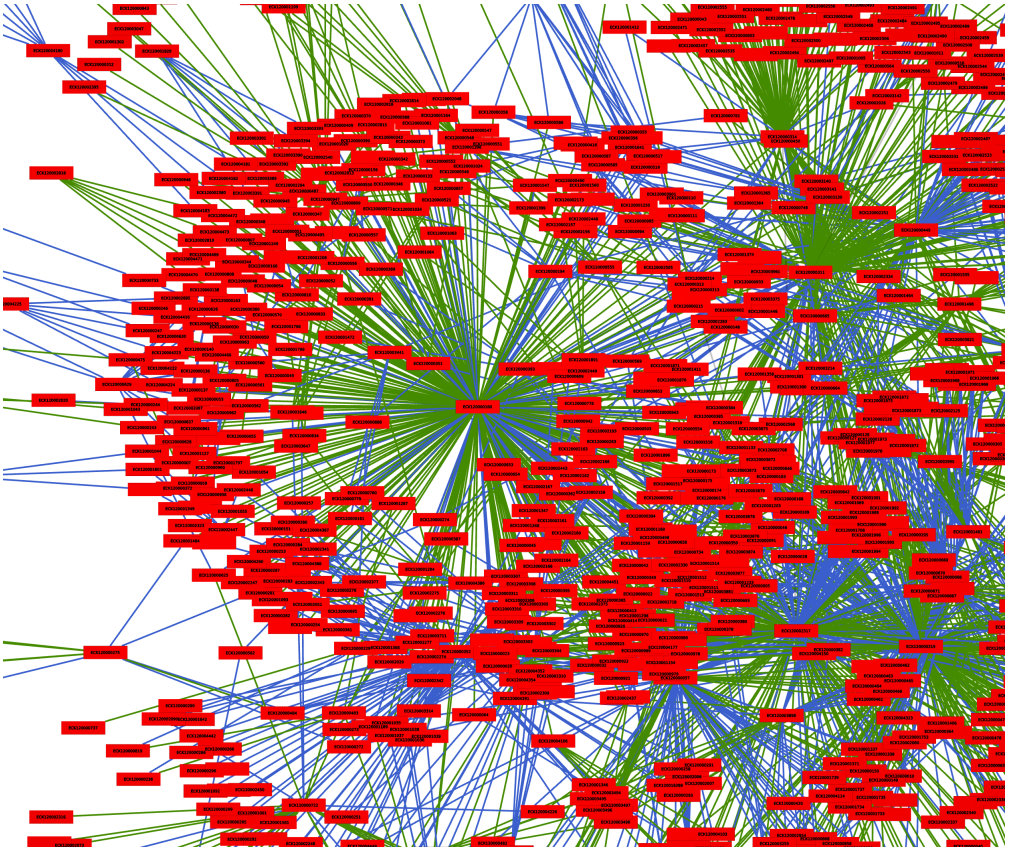
Master Bioinformatique

Analyse visuelle de données d'expression génique



Jean MAINGUY & Mauriac RAZAFIMAFONTY

Table des matières

Introduction

La compréhension des mécanismes régissant l'expression des gènes est un enjeu important dans le domaine de la biologie. L'utilisation des données «omiques» permet de mieux appréhender cette problématique cependant au vu l'importance des données, il est souvent pas très évident de faire l'exploration de celles-ci.

Une modélisation en réseau de gènes permet de minimiser cette difficulté d'exploration en choisissant la meilleure technique de visualisation correspondante. L'étude des niveaux d'expression est une approche pour comprendre le métabolisme d'un organisme, pour pallier à des problèmes de prévention, dans le cas d'un organisme pouvant engendrer une maladie, ou tout simplement comprendre son fonctionnement.
Ici, on souhaiterai visualiser les niveaux d'expressions de E. coli à partir des données d'expressions provenant de 17 temps, durant lesquelles E. coli a poussé dans un mélange de glucose et de lactose. Cette bactérie utilise préférentiellement le glucose comme source d'énergie, mais une fois que celle-ci est épuisé, elle arrête de se croître pour pouvoir ajuster son développement en utilisant le lactose. Cette phase de substitution de source d'énergie est appelée : la DIAUXIE. Ce phénomène intervient au bout du sixième temps. Au quatorzième temps, la phase stationnaire intervient ou la bactérie arrête complètement de pousser en raison d'une insuffisance de nutriments dans le milieu. Durant cette phase, pour économiser de l'énergie, la bactérie cesse plusieurs de ses processus métaboliques.

On va donc implémenter une première visualisation à partir d'un graphe des différents locus des gènes. Une méthode de partitionnement des gènes selon leurs niveau d'expression.

Les différents groupes seront par la suite analysés grâce à la création d'une carte de chaleur qui permettra de mettre en évidence les différents niveaux d'expression au sein de chaque groupe.

Chapitre 1

Description des algorithmes implémentés

1.1 Description des données analysées

Les données que nous analysons sont issus d'une expérience sur l'E coli au cours duquel la bactérie pousse sur un milieu contenant deux types de sucres : du glucose et du lactose. Nous avons donc à notre disposition les données d'expression des gènes de la bactérie, mesurer à différents temps, allant de tp1 à tp17. Les gènes sont identifiés par leurs codes ECK. Ce code désigne un régulon, qui est un ensemble fonctionnel de gènes organisés en opéron ou non, dont l'expression est co-régulée.

A chaque temps, les niveaux d'expression peuvent différer, d'un locus à un autre. Certains locus présentent même des niveaux d'expression nuls pour l'ensemble des 17 temps. Cette absence apparente d'expression pourrait être expliquée soit par un artefact dû à l'expérience ou bien tout simplement ces locus désignent des gènes qui n'ont pas été exprimés lors de l'expérience. Les données fournies contiennent également les régulations entre les gènes. La régulation opérée par un gène régulateur sur un gène régulé peut être de deux sortes : positive ou négative. Dans le cas d'une régulation positive, le régulateur, lui lorsqu'il est exprimé favorise l'expression du gène régulé et à l'inverse une régulation négative entraîne une répression du gène régulé par le régulateur.

L'analyse de ces données est réalisée en utilisant le logiciel Tulip version 5.1.0 [1] en réalisant une pré-visualisation des données, un partitionnement multi-échelle des locus basé sur une mesure de la similarité de leurs niveaux d'expression pour enfin visualiser les groupes de gènes obtenus à l'aide d'une carte de chaleur. La carte de chaleur permet de distinguer l'évolution des niveaux d'expression au cours du temps au sein des groupes.

1.2 Pré-traitement et première visualisation

Un graphe permet de visualiser des données sous forme de nœuds et d'arêtes. Dans notre cas, les gènes sont représentés par les nœuds du graphe et les arêtes correspondent aux régulations entre les gènes. La pré-visualisation des données nous permet de faire une première visualisation des locus. Une étiquette rectangulaire a été choisie pour afficher les nœuds permettant ainsi une visualisation des noms des gènes de façon optimale sans perte d'espace.

Nous avons choisi des couleurs différentes pour différencier les types de régulation entre deux gènes, verte pour les régulations positives, et bleue pour celles qui sont négatives. Un modèle de représentation par le modèle de force a été adopté pour déterminé la position relative des sommets du graphe. Ce modèle fournit une visualisation général et harmonieuse des données sans avoir d'à priori sur sa structure.

1.3 Partitionnement des gènes

1.3.1 Choix de la mesures de distance/similarité entre les gènes

Le partitionnement est basé sur une mesure de similarité ou de dis-similarité des gènes basés sur leurs niveaux d'expression. Pour ce faire, il faut donc prendre en considération la distance entre ces objets. Un grand nombre de mesures existent pour déterminer une distance ou une similarité entre deux objets basé sur une série de valeurs numérique. En biologie le challenge de déterminer des groupes de gènes selon leurs niveaux d'expression est bien connu. Plusieurs mesures peuvent être utilisées et préférées selon le type de données biologiques à disposition et la méthode de regroupement par la suite utilisée pour les regrouper. Un article de Jaskowiak, P. et all [2] répertorie et teste les différentes mesures utilisées en Biologie lors d'un groupement de gènes basé sur des données d'expression génique. Ces mesures peuvent être des corrélations : corrélation de Pearson (PE), Goodman-Kruskal (GK), Spearman (SP), Kendall (KE) par exemple ou bien des mesures de distance plus traditionnelle telles que la distance euclidienne, de similarité de Cosine, ou de distance de Manhattan. Enfin ils ont également testé des mesures spécialement créées pour l'étude d'expression de gène dont notamment la mesure Jackknife, Local Shape-based Similarity, YS1 et YR1 notamment. Les auteurs préconisent certaines mesures adaptées aux études de niveau d'expression géniques au cours du temps. Ainsi il ressort que YS1, YR1 jackknife et Cosine semblent assez appropriés alors que d'autre mesures sont déconseillées dont notamment la corrélation de Pearson. [2]. Par conséquent nous avons choisis d'utiliser Jacknife et Cosine ainsi que la correlation Pearson dans notre étude de manière à avoir plus de liberté.

1.3.2 Implémentation de Cosine et Jackknife

Nous avons fait le choix de coder nous même les calculs de mesures pour d'une part assurer une meilleur portabilité de notre programme et d'autre part les calculs de distance que nous avons choisis sont relativement simple à implémenter. La mesure de Jackknife dépend directement des mesures de Pearson donc nous avons également implémenté Pearson. Cependant nous avons néanmoins comparé les résultats de Pearson et Cosine grâce au module Scipy et numpy. Nous nous sommes basés sur les formules décrites dans l'article [2].

— La formule de la mesure de similarité de Cosine (cos) :

$$cos_{sim}(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\sqrt{\sum_{i=1}^{n}(x_i)^2}}$$

— La formule de la mesure de Jackknife (JK) correspond à la plus petite valeurs de correlation de Pearson sur l'ensemble des cas considérant le retrait d'une valeur de la liste :

$$JK(x,y) = min_{(0\leq i\leq n)} PE^{i}(x,y))$$

PE^{i} correspond à la corrélation de Pearson sans la ième valeur.

1.3.3 Algorithme de mesure et première filtration

L'algorithme compare un a un tous les nœuds ensemble en prenant soin de pas comparer deux fois le même nœuds. Une première étape de filtration est effectué sur les nœuds présentant un niveau d'expression de 0 à tous les temps. Ces gènes sont enlevés car ils sont soit dus à un artefact de l'expérience ou bien à une absence totale d'expression lors de l'expérience et donc n'apportent pas d'information pertinente pour le reste de l'analyse. Les mesure de similarité sont enregistrées dans une propriété du graphe "similarity" au niveau des arêtes reliant les pairs de nœuds. Pour la mesure de Jackknife, il est possible d'avoir des valeurs négatifs ce qui indique entre deux gènes leurs anti-corrélation. Les mesures négatives sont également pas prise en compte dès l'étape de la mesure. Bien que l'anti-correlation est informative et pourrait être intéressante à étudier et incorporée dans l'étude, nous préférons garder seulement des valeurs corrélées. Ces premières filtrations retirent 213 nœuds qui ont un niveau d'expression à 0 et également les arêtes reliant deux gènes anti-corrélés.

1.3.4 Choix de l'algorithme de partitionnement

Pour réaliser l'étape de partitionnement nous avons choisis d'utiliser un plugin tulip MCL clustering pour Markov Cluster Algorithm. Cet algorithme va chercher à regrouper les nœuds qui sont très connectés entre eux et il ne prend pas seulement le nombre d'arête mais également la valeur de similarité associée. Il est cependant nécessaire de réaliser une filtration sur le valeur de corrélation de telle sorte à réduire le nombre d'arête. Une trop faible filtration entraîne un faible partitionnement et une trop grande filtration entraîne un grand nombre de petit groupe inutilisable.

1.3.5 Partitionnement multi-échelle

Nous avons choisis d'implémenter un algorithme permettant de générer plusieurs niveaux. Cependant le nombre de niveaux maximum possible est paramétrable donc il est tout à fait possible de réaliser un partitionnement à un seul niveau avec notre implémentation. La stratégie adoptée pour réaliser un partitionnement multi-échelle est l'utilisation de fonctions récursives et de dictionnaire imbriqué gardant en mémoire les nœuds et leur appartenance à un niveau de partitionnement.

Avant chaque partitionnement les nœuds sont filtrés en fonction d'un percentile donné en paramètre (par défaut le percentile est égal à 0.5 donc correspond à la médiane) de l'ensemble des valeurs de similarité du groupe de nœuds à partitionner. Toutes les arêtes portant une valeur inférieure au percentile donné sont éliminées. Cette approche de filtration permet de s'adapter à chaque niveau aux groupes partitionnés. En effet un seuil de filtration fixe déterminé au début du partitionnement ne sera plus en adéquation avec les groupes des partitionnements des niveaux suivants.

Les étapes du partitionnement sont les suivantes :

1. Un groupe de gènes donnés à la fonction récursive pour se faire partionner
2. Un sous-graphe vide du graphe ancestral est créé et intègre le groupe de gènes à partitionner
3. Les arêtes du groupe de gènes est filtré selon le percentile
4. Les gènes sont partionnés par MCL clustering et les groupes sont récupérés dans un dictionnaire
5. S'il y a pas de nouveau groupe créer par MCL ou bien que la taille des groupes est plus petit que le seuil du nombre de gène minimum requis pour partitionner, la récursion se termine et le dictionnaire est renvoyé
6. S'il y par contre des groupes qui satisfont toutes les conditions précédentes alors la fonction de partitionnement est lancée récursivement sur chacun d'entre eux. (retour au point 1.)

Le dictionnaire imbriqué retourné par la fonction de partitionnement récursive permet d'avoir une variable facile à manipuler pour créer la carte de chaleur à plusieurs niveaux.

1.4 Carte de chaleur

Après le partitionnement des gènes, une carte de chaleur est créée pour permettre une visualisation les gènes regroupés avec les gènes dont l'expression varie similairement aux différents temps de l'expérience.

L'approche multi-échelle pour la création de la carte de chaleur utilise également la récursivité.

Nous avons veillé lors de la phase d'implémentation des algorithmes plus particulièrement pour la carte de chaleur, de coder le moins possible en "dur" pour permettre à la table de chaleur de s'adapter à n'importe quelle taille de jeu de données. En effet les dimensions de la table de chaleur sont régi par un ratio largeur/hauteur de 0.5. A partir de ce ratio, et du nombre de gène ainsi que le nombre de temps dans l'expérience, les pas pour les coordonnées X et Y sont établis. Ces pas correspondent à la distance entre les centres de deux nœuds adjacents. Il correspond donc à la distance à laquelle un nouveau nœud est disposé dans l'arbre.

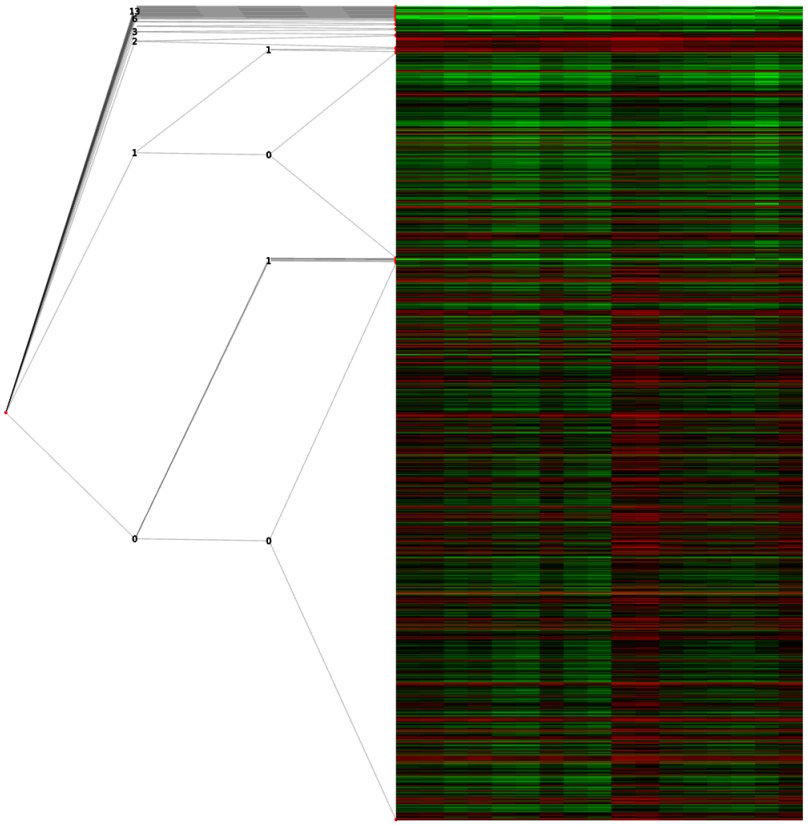


FIGURE 1.1 – Arbre et carte de chaleur permettant la visualisation des groupes et le niveau de partitionnement dont ils sont issus

Chapitre 2

Résultats et interprétation

Pour le graphe de pré-visualisation, basé sur le modèle de force, on peut voir sur la représentation qu'on a des gènes qui sont très connectés à d'autres , ils pourraient donc être potentiellement des régulateurs importants. D'autres gènes ne sont pas du tout connectés au réseau, ils constituent des petits groupes à part.
Les différents niveaux de partitionnement permettent de visualiser les différents groupes englobant d'autre groupes. Quelque groupes sont composés d'un nombre importants de gènes alors que d'autres non. On a donc décidé de se focaliser sur le groupe qui contenait le plus de gènes à l'intérieur (562 gènes) 2.2.

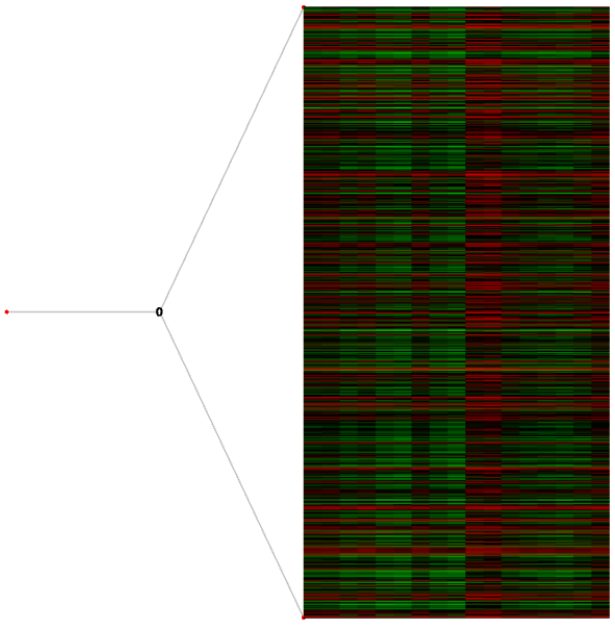


FIGURE 2.1 – Groupe de gènes issu du 3ème niveau de partitionnement calculer à partir de mesure de Jackknife. Correspondant au groupe le plus en bas de la Figure 2.2

On a donc en largeur les différents temps d'expérience durant lesquels on a effectué la culture de l'E.Coli et en hauteur les gènes contenus dans le cluster. On peut donc voir sur cette carte de chaleur que à certains temps (à temps 5 et 6), la plupart des gènes sont sous-exprimés (couleur verte dominante). A temps 7 par contre, on a une sur-expression temporaire des gènes qui se sous-expriment juste après. Mais la plus importante des activités des gènes surviennent au bout du 10 ème temps (couleur rouge dominante).

On a également effectué un test d'enrichissement sur Panther [1] du cluster en question pour aller plus loin dans l'interprétation en tenant en prenant comme référence le **Biological Process** du **Gene Ontology**. L'étude statistique des gènes montrent que certains gènes sont enrichis au sein de la liste de référence de l'E Coli alors que d'autres non. On a pu constater que les gènes qui sont sur-représentés sont ceux qui sont impliqués fonctionnellement dans les processus de métabolisme cellulaire. On a pu constaté cela également grace à une présentation en camembert. Le premier niveau montre la grosse par des gènes dans le métabolisme cellulaire. Le deuxième niveau a montré leur implication dans le processus de métabolisme primaire. Cela pourrait être expliqué par le fait que la cellule étant dans une phase de croissance, mobilise donc tous les gènes qui sont essentiels à sa survie. Ces gènes là auront donc une activité importante lors de la phase exponentielle de croissance avant la phase stationnaire.
En contraste avec cela, on peut également constater des gènes qui sont très sous-représentés au sein du cluster. Ce sont des gènes qui sont impliqués dans tous des processus de transcription, de réparation de l'ADN, processus de développement. Ce sont probablement ces gènes là qui sont sous-exprimés lors de la phase stationnaire.

1. http ://www.pantherdb.org/

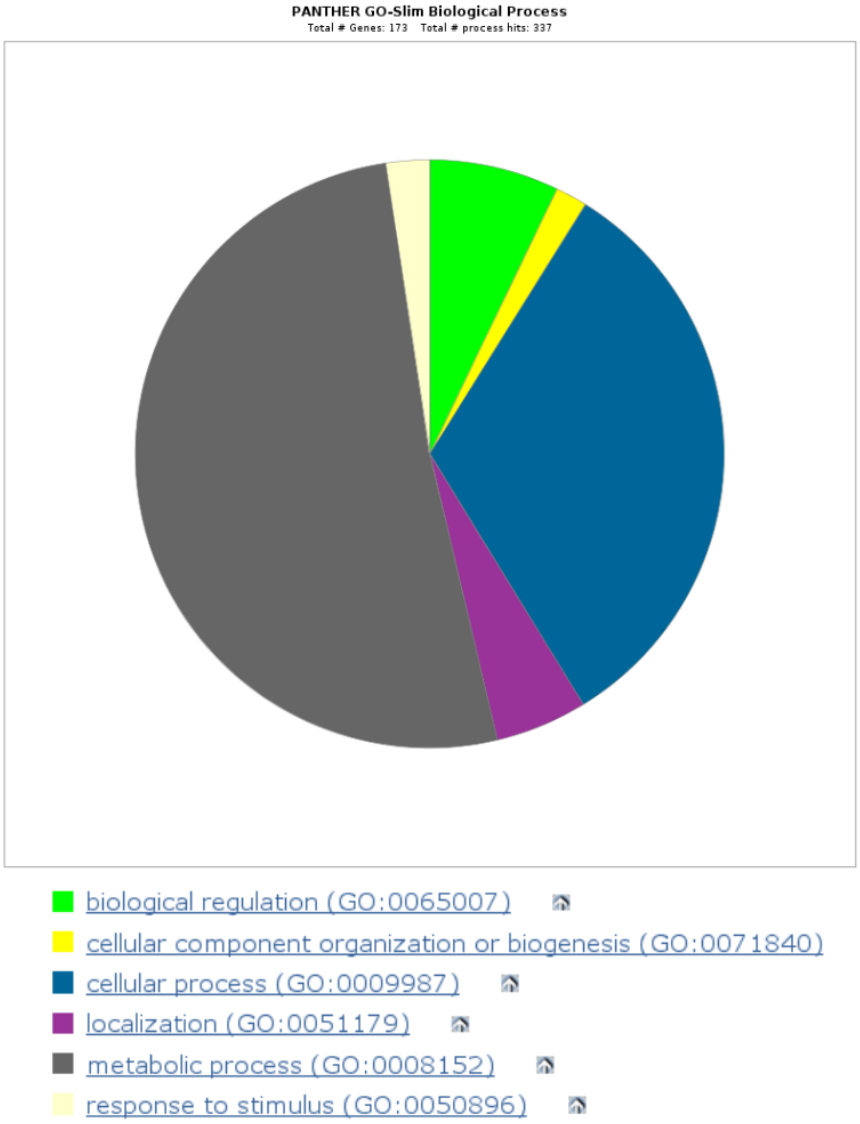


FIGURE 2.2 – Représentation de la part d'implication des gènes dans le processus Biologique

Conclusion et Perspectives

La compréhension des phénomènes biologiques d'une cellule peut être faite par l'analyse des données d'expression de cette cellule. Cette analyse de données peut alors être faite de différente manière, mais une des façons les plus commode est de groupé les gènes selon leurs niveau d'expression dans les différent temps ou conditions de l'expérience. Ces groupes de gènes ensuite analysé par enrichissement et visualisé à l'aide de graphe et de carte de chaleur. Notre analyse visuelle a permis de visualiser les interactions de régulation entre les gènes et observer la présence d'une poignée de gènes régulateur affectant l'expression de nombreux autres gènes. Notre carte de chaleur quant à elle a permis de mettre en évidence que certains groupe de gènes sont plus ou moins exprimé à différent stade de de l'expérience. Et finalement l'étape l'enrichissement a permis de mettre en évidence que ces gènes sont effectivement impliqués dans des processus de métabolisme biologique, qui est une étape non négligeable lors de la croissance d'une cellule.

Les algorithmes que nous avons implémentés lors de ce projet ont permis de mener à bien l'analyse visuelle. Cependant les groupes obtenus lors du partitionnement sont assez disproportionnés avec un grand nombre de petit groupe et 1 voire 2 groupes géants regroupant la majorité des gènes. Il serait donc intéressant soit de tester d'autres algorithmes de partitionnement tels que des algorithmes de clustering hiérarchiques.

Enfin les algorithmes implémentés pour ce projet constituent un socle solide pour mener à bien des analyse biologiques plus poussées et ne sont pas spécifiques aux données analysées. De plus les fonctions sont très paramétrables. Il est par exemple possible de choisir entre trois fonctions de mesure de similarité et il serait également très facile d'en rajouter des nouvelles.

Bibliographie

[1] David Auber. Tulip—a huge graph visualization framework. *Graph drawing software*, pages 105–126, 2004.

[2] Pablo A Jaskowiak, Ricardo JGB Campello, and Ivan G Costa. On the selection of appropriate distances for gene expression data clustering. *BMC bioinformatics*, 15(2) :S2, 2014.