Université de Bordeaux
Master 2 Bio-Informatique

Année 2019- 2020

Visualisation d'Information
Projet – Analyse visuelle de données d'expression génique –

Contexte : Lorsque nous nous intéressons à l'activité des gènes, nous recherchons des régions régulatrices à proximité des gènes, or il a été montré que la distance entre le gène et celles-ci n'était pas une information suffisante, voire trompeuse. Nous étudions ici l'entrée dans le cycle cellulaire de cellule T et cherchons à identifier des régions régulatrices de l'expression des gènes différentiellement exprimés entre des cellules dormantes (quiescentes) et activées. Pour cela, nous avons utilisé des données de type capture-HiC, grâce auxquelles avons identifié des régions de l'ADN étant proches dans le noyau de la cellule mais éloignées dans l'espace linéaire. L'expression des gènes a été acquise à travers une expérience de RNA-seq.

Objectif : L'objectif de ce projet est de réaliser une analyse d'un jeu de données qui vous est fourni. Pour cela, vous trouverez ci-dessous plusieurs sous-objectifs à atteindre (le barème est donné à titre indicatif). Le résultat sera remis sous la forme d'un fichier contenant le code python (compatible avec l'IDE Python de Tulip) permettant à partir d'un graphe vide d'atteindre ces différents sous-objectifs. Vous devrez aussi fournir un rapport d'une dizaine de pages expliquant les algorithmes, choix que vous aurez faits et informations tirées de l'analyse que vous ferez des données. Enfin, vous exposerez votre travail au cours d'une soutenance de 15 min. Ce jeu de données est un jeu de données réel, toujours en cours d'analyse. Il se peut donc qu'il soit difficile d'en tirer des informations. Vous serez donc aussi bien évalués sur la démarche que vous mettrez en œuvre que sur les résultats que vous obtiendrez.

Modalités : Projet à réaliser en groupe de 3 ou 4 étudiants

Dates importantes : remise du code et du rapport au plus tard le 12 février à 23h59 et soutenance le 13 février de 9h à 12h

Fichiers fournis : Vous trouverez à l'adresse :
http://www.labri.fr/perso/bourqui/downloads/cours/Master/2019/Projet/
les fichiers suivants :

- interactions_chromosome6.csv : liste des interactions probables entre fragments d'ADN ainsi que la présence de chaque interaction dans chaque condition
- chromosome6_fragments_expression.csv : décrit le type de fragment ainsi que son expression différentielle (données qualitatives)
- KEGG.symbols.csv : liste de voies métaboliques issue de KEGG et les gènes qui y sont impliqués
- REACTOME.symbols.csv : liste de voies métaboliques issue de Reactome et les gènes qui y sont impliqués

Un descriptif du format de chacun de ces fichiers est donné en fin d'énoncé.

Partie 1 : Chargement du réseau d'interaction (/3 points)

En utilisant la description fournie ci-dessous, écrivez un algorithme qui à partir d'un chemin d'accès au fichier interaction_chromosome6.csv permet de construire le graphe des interactions.

Partie 2 : Chargement des informations relatives aux fragments (/3 points)

En utilisant la description fournie ci-dessous, écrivez un algorithme qui à partir d'un chemin d'accès au fichier chromosome6_ fragments_expression.csv permet d'ajouter au graphe obtenu précédemment les informations contenues dans ce fichier.

Partie 3 : Chargement des voies métaboliques (/3 points)

En utilisant la description fournie ci-dessous, écrivez un algorithme qui à partir de chemins d'accès aux fichiers *.symbols.csv permet d'ajouter au graphe obtenu précédemment les informations contenues dans ce fichier.

Partie 4 : Visualisation du graphe d'interaction (/3 points)

Ecrivez un algorithme permettant de dessiner le ou les graphes que vous utiliserez dans la Partie 5 et d'affecter des position, couleur, forme, taille, *etc.* à chaque sommet et/ou arête.

Partie 5 : Analyse des données (/8 points)

Cette partie est laissée relativement libre, l'objectif est d'identifier des groupes de gènes fortement associés et de mettre en évidence les fonctions biologiques régulées.

Vous pouvez utiliser n'importe quelle source de données externe dans la mesure où cela est fait de manière programmatique (vous pourrez aussi utiliser vos propres connaissances). Vous pourrez aussi programmer n'importe quelle autre représentation, partitionnement (le partitionnement fournit permet-il une interprétation ?), extraction de sous-parties sur/sous exprimées, *etc*. Afin de pouvoir réaliser cette partie, il faudra éventuellement revenir sur la modélisation ou la visualisation des données que vous aurez faites (*cf* Parties 1, 2, 3 et 4).

Les résultats de cette partie devront être décrits dans le rapport. Il est important qu'apparaissent dans votre rapport les questions auxquelles vous avez essayé de répondre (ou répondu) ainsi que les moyens mis en œuvre pour y arriver.

Description des fichiers fournis :

interactions_chromosome6.csv contient toutes les interactions stables / gagnées / perdues lors de l'activation de la cellule T.

- chromosome // Une seule valeur car les interactions sont intra-chromosomales
- ID_locus1 // identifiant de la première région (nom de gène ou position chromosomique)

- ID_locus2 // identifiant de la seconde région (nom de gène ou position chromosomique)
- interaction_status // stable / gain / loss
- distance // distance linéaire sur l'ADN entre les 2 régions

chromosome6_fragments_expression.csv

- chromosome
- IDs // Correspond aux identifiants dans le fichier précédent (ID_locus1,ID_locus2)
- expression // expression des gènes (intergenic: non genic regions, stable, up, down, NA: valeur non-available)

[db].symbols.csv

- identifiant de la voie métabolique
- url décrivant la voie métabolique
- gènes (LOCUS) impliqués dans la voie métabolique