TV15MI - TV25MI Projet TER : Réunion du 15/09/23

Compte-rendu

Présents : Sophie LÈBRE, Thomas AYRIVIÉ, Mehdi BELKHITER, Jamila CHERKAOUI, Dina EL HIJJAWI, Magatte LO.

Prochain rendez-vous : Vendredi 22 septembre à 11h00, Université Paul Valéry, bâtiment B.

Travail à effectuer et répartition :

- Installer sur nos ordinateurs les données JASPAR (Motif vertébrés, redondants, format MEME, single batch), les données RNA seq (NCBI Gêne Expression Omnibus). + Installer le package GEOquery sur R et télécharger les données directement avec R. + Installer logiciel FIMO (de la suite MEME) pour scanner les séquences. *Everyone.*
- Commencer à lire les articles scientifiques. *Dina (anglais), everyone (français).*

- Bien définir le sujet. *Jamila.*
- Les prochaines étapes seront la description de la liste des variables, l'analyse exploratoire du jeu de données et de la matrice des X avec notamment les cours d'analyse multidimensionnelle (ACP, AFC…). *Everyone pour regarder comment les données s'articulent et tester les commandes GEOquery.*

- Finir ce compte-rendu. *Thomas.*
- Noter nos questions pour le prochain rendez-vous et envoyer l'ordre du jour avant mercredi soir. *Tout le groupe.*
- Commencer à rédiger/répondre aux questions proposées du 1er rapport intermédiaire à rendre le 15 octobre. *Mehdi avec l'appui de tout de monde.*

À aborder au prochain rendez-vous :

- Mme Lèbre va voir avec Mathilde pour se procurer les séquences ADN associées à chaque gène.
- Explications sur l'autre approche d'un modèle linéaire, les arbres et forêts.
- Bien expliquer les scores avec l'appui des formules mathématiques.

Points abordés :

- Explication de la problématique grâce aux slides de l'équipe. Mme Lèbre est en collaboration depuis 2015 avec Mathilde R, Laurent B et Charles L (biologie, bioinfo).

- Dans les cancers, certains gènes seront plus actifs que d'autres, puisqu'à la réplication, ils se copient pas identiquement. On peut récupérer toutes ces copies pour les

séquencer/les compter (entre 0 et des milliers). La technologie s'appelle le **séquençage à haut débit**.

- Les protéines s'appellent les **facteurs de transcription**. Un k-**mère** est une sous-chaîne, un morceau de séquence/combinaison de nucléotides.

- On a un **problème de classification**. On essaie de comprendre la séquence ADN des patientes (Gènes à l'intérieur qui codent les protéines).

- Nous avons plusieurs **types de données**. Des données de comptage, RNA seq : Compter pour chaque gènes le nombre de copies (20 000 codes de gènes différents). L'avantage de ce dataset c'est que les données sont déjà le log10 ratio of experimental (treatment) to reference (baseline) sample (c'est du 2 color agilent).

  Données de transcription, ChiP-seq : Liste des positions où le facteur s'est fixé. + Autres données expérimentales, épigénétiques…

- Sophie Lèbre nous également transmis un code R de Mathilde Robien qui importe les données RNA Seq :

  > data <- GEOquery::getGEO("GSE130787", GSEMatrix = TRUE, getGPL = TRUE) : Import automatique des données RNA Seq
  >
  > data@featureData@data : donne accès aux noms des gènes correspondants aux identifiants (hyper pratique!!)
  >
  > data@phenoData@data : donne accès aux info sur les patientes, leur traitement et leur réponse (RD = residual disease ; PCR = pathological complete response)
  >
  > data@assayData$exprs : donne accès aux données d'expression

- Le **logiciel Fimo** scanne les séquences, calcule de score de chaque séquence d'ADN (Données Jaspar). Lors de nos recherches, nous devrons prendre les maximums des scores.

- Comme en biologie généralement, les **variables sont très corrélées**. Nous utiliserons **Glmnet (Lasso)** pour l'approche modèle linéaire, afin de pouvoir "sélectionner correctement les variables optimales", ajuster des modèles de régression linéaire généralisée (GLM).

- D'autre part, nous avons pris en photo quelques schémas dessinés sur le tableau pendant la réunion pour la compréhension du travail de recherche.

  Nous disposerons de données RNA-seq avant et après traitement. Grâce à un **FoldChange** de ces valeurs, nous pourrons apercevoir les éventuels changements génétiques. À partir de cela, 3 groupes de valeurs en seront déduits. Un premier groupe avec peu ou pas de changement avant/après traitement (valeur 0). Un deuxième avec un changement significatif positif (valeur 1). Et un dernier avec un changement significatif négatif (valeur -1). *fig. 3*

  D'autre part, grâce à des séquences connues, nous cherchons les combinaisons de nucléotides ayant les scores les plus élevés dans les données biologiques à notre disposition. *fig. 1 et 2*

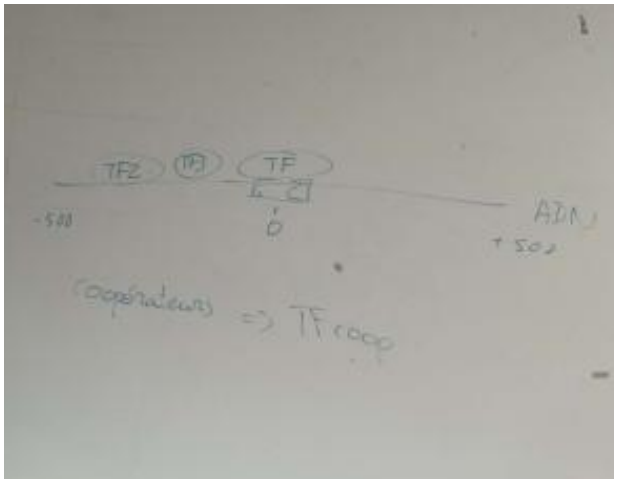

Figure 1

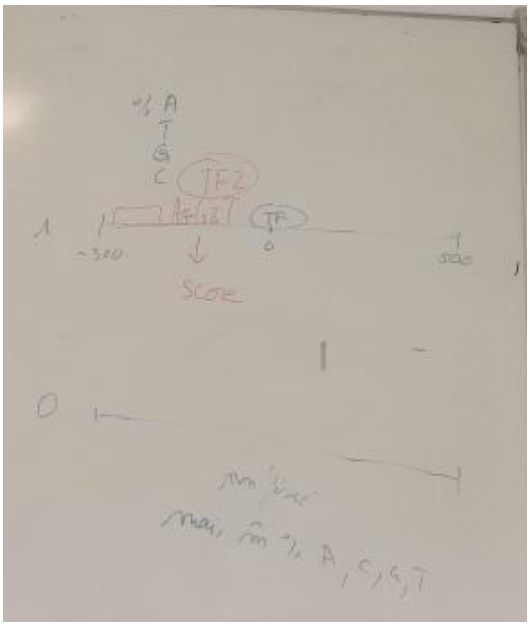

Figure 2

Figure 3

Définition du sujet :

Le sujet de recherche se concentre sur l'identification des voies de signalisation activées en réponse aux traitements du cancer du sein. Au fil du temps, les cellules cancéreuses ont la capacité de s'adapter et de développer des mécanismes de résistance, y compris face à la chimiothérapie. Comprendre ces mécanismes est essentiel pour améliorer les stratégies de traitement.

Pour ce faire, nous utilisons une analyse approfondie des séquences génétiques afin de déterminer quelles combinaisons de nucléotides obtiennent les scores les plus élevés. Ces scores reflètent la similitude ou la pertinence d'une séquence donnée par rapport à un modèle de référence.

Préalablement, un modèle de régression a été développé pour expliquer l'activité des gènes en utilisant uniquement la séquence d'ADN. Cependant, nous visons maintenant à mettre en place un nouveau modèle de classification, dont les détails seront abordés ultérieurement au cours du projet.

Pour mener à bien notre recherche, nous utilisons la base de données Jaspar, une ressource spécialisée dans les matrices de séquences d'ADN liées aux sites de liaison des facteurs de transcription. Ces matrices, également appelées matrices de poids de position (PWM), offrent une représentation numérique des motifs de liaison des facteurs de transcription. Jaspar fournit des informations détaillées sur la fréquence de chaque base (A, C, G, T) à chaque position au sein de ces motifs de liaison. Cette base de données est essentielle pour développer des modèles plus précis dans le but de mieux comprendre la régulation génique et les réponses cellulaires aux traitements du cancer du sein.

En complément, nous utilisons pareillement la base de données GEO (Gene Expression Omnibus). GEO stocke une grande variété de données d'expression génique provenant de diverses expériences et technologies, incluant les puces à ADN et les séquençages ARN. Les chercheurs déposent leurs données d'expression génique dans GEO, permettant ainsi à d'autres scientifiques d'y accéder, de les partager et de les analyser. Cette base de données facilite l'exploration et la comparaison des profils d'expression génique dans différents contextes biologiques et expérimentaux.

Jaspar nous aide à analyser les motifs de liaison des facteurs de transcription, tandis que GEO nous permet d'accéder à une vaste gamme de données d'expression génique pour une exploration approfondie des mécanismes biologiques et des réponses aux traitements.