TV15MI - TV25MI Projet TER : Réunion du 20/10/23

Compte-rendu

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

Prochain rendez-vous : Vendredi 10 Novembre 2023 9h15, Université Paul Valéry, bâtiment B.

Travail à effectuer :

- Créer les classes : up, down et neutre en considérant un seuil à 2 pour le up (resp. 1/2 pour les down) pour le ratio des comptages de read ARNm ("fold change")

  => attention les ratio d'expression sont donnés en log10(fold change), il vous faudra donc adapter le seuil (log10(2) et -log10(2) resp.).

- Démarrer la classification.
- Répartition des tâches.
- Analyse exploratoire des données.

Points abordés :

Utilisation de Python => pas de soucis. Mais en statistiques utiliser R.

On utilise un seuil = 2 ⇔ 2 fois plus => différence significative.
Sélection des gènes sur exprimés, gènes sous exprimés.

> Regarder la distribution de Y avant les seuils et après les seuils.

> Analyse descriptive de la matrice des scores et de la matrice des X.

Faire le rang et le ratio des valeurs propres, ACP puis regarder l'inertie.
La partie descriptive intéresse beaucoup Madame lèbre pour les X.

Démarrer la classification. 800 modèles car 800 variables, on garde le meilleur modèle. (For Word ?)

BIEN DÉCRIRE X > Non supervisé = on sait qu'on va avoir des classes. Clustering.

OBJECTIF FINAL > Supervisé = on sait qu'on aura une réponse (connaitre les gènes up ou down). Classification.

Ce semestre on attend plus l'analyse des données que la classification.

QUESTION : Construction de X : Voir s'il y a des doublons entre séquence name (motif ID et motif alt ID)

Sortie FIMO : Regarder pourquoi il y a des Nan, et faire en sorte de ne pas les avoir
Vérifier avec Fimo que nous n'avons plus de nan avec les nouveaux paramètres (élever le seuil de pvaleurs 10^-3 on passe à 10^-2)

Remettre en dur tout le pipeline de constitution des données !!

Suite du travail

1) Jointure avec Y
2) Combien de Nan
3) Distribution des p-valeurs (entre 0 et 10^-3)
4) Considérer de refaire tourner fimo avec un seuil à 10^-2 sur les 17000 séquences de la matrice seulement
   Ou remplacer les nan par une valeur supérieure à 10^-2