Analyses statistiques de données single-cell RNA-seq
Application à l'étude des cellules stromales mésenchymateuses

Contexte du stage :

L'équipe « Plasticité des tissus adipeux » du STROMALab s'intéresse à l'étude de la dynamique cellulaire des tissus adipeux. En utilisant le tissu adipeux comme modèle, l'objectif est d'identifier les signaux issus du stroma susceptibles d'induire la régénération de tissus et/ou d'organes, et plus précisément de déterminer comment les cellules stromales mésenchymateuses dérivées du tissu adipeux (ASC) pourraient participer à la régénération. Cette équipe souhaite maintenant profiter de la nouvelle technologie de séquençage, le single cell RNA-seq (scRNAseq) pour poursuivre ses recherches.

Cette nouvelle technologie consiste à isoler les différentes cellules d'un échantillon (par ex morceau de tissu sur un organisme) et de les séquencer séparément (c'est-à-dire mesurer l'expression des gènes à partir des ARN). Un tableau typique de données scRNAseq se compose de comptages qui mesurent l'expression de p gènes pour n cellules séquencées. On se retrouve alors avec un tableau de très grande dimension ($n << p$ ou $n \sim p$) et sparse au sens qu'au moins 80% des données sont des zéro (principalement des non-détections d'expression de gènes dits dropouts). A partir de ces données, on souhaite mettre en évidence des groupes de cellules ayant les mêmes évolutions d'expression, des gènes signatures permettant de différencier ces groupes de cellules et de reconstruire un « chemin » d'évolution des cellules (pseudo-time).

Objectifs du stage :

Dans un premier temps, l'étudiant(e) devra s'approprier les données de scRNAseq, les méthodes statistiques déjà existantes ainsi que les packages associés.

Dans un second temps, l'étudiant(e) s'intéressera à la question de la classification non supervisée des cellules et à la détermination de gènes signature, en tenant compte des spécificités de ces données massives et sparses. Au travers de cette question, l'étudiant(e) sera amené(e) à mettre en place des méthodes de sélection de variables, de classification non supervisée, d'apprentissage, de biclustering, ...

Dans un troisième temps, on s'intéressera au problème dit du « pseudo-time ». L'étudiant(e) devra comparer des méthodes existantes afin de déterminer celles qui répondent le mieux à la problématique des biologistes du STROMALab.

L'implémentation des méthodes et le travail d'analyse seront effectués à l'aide du logiciel R. Il sera peut-être nécessaire de travailler en Python pour certains points. La création d'une interface graphique type Shiny pourra être envisagée pour aider les biologistes dans la prise en main des outils développés.

Remarque :

Ce sujet de stage étant très riche, il est possible de le subdiviser en deux stages pour niveau M2 ou M1. Les stagiaires pourront travailler en collaboration sur certains points.

Conditions du stage :

- Durée : 4 à 6 mois pour niveau M2 ; 3 à 4 mois pour niveau M1
- Rémunération : taux légal (environ 550 euros par mois)
- Encadrement : Cathy Maugis-Rabusseau, Sandrine Laguerre et Marielle Ousset
- Localisation : INSA de Toulouse

Profil recherché :

- Master 2 en mathématiques appliquées ou équivalent en école d'ingénieur, possibilité pour niveau Master 1
- Bonnes connaissances en statistiques théoriques et appliquées
- Langages en programmation : R maitrisé ; Python apprécié
- Aucune connaissance en biologie n'est nécessaire mais un goût pour ce domaine d'application sera apprécié

Pour candidater :

Envoyer un CV et une lettre de motivation à cathy.maugis@insa-toulouse.fr