Titre du stage :

Clustering de données protéomiques de grandes dimensions

Entreprise :

Institut Pasteur , Paris

Période : 6 mois, printemps 2019

Candidatures :

Envoyer CV et lettre de motivation aux adresses suivantes :
quentin.giaigianetto@pasteur.fr
mariette.matondo@pasteur.fr

Profil recherché :

• Master en mathématiques appliquées avec une spécialisation en statistiques, analyse de données.
• Débutant (stage de fin d'étude), H/F.
• Maîtrise courante du français ou de l'anglais.
• Compétences en statistiques, analyse de données et programmation en R.
• Intérêt pour le travail en équipe dans un environnement multidisciplinaire.

Sujet :

La protéomique est l'étude à grande échelle de l'ensemble des protéines (appelé « protéome ») présentes dans des échantillons biologiques (cellules, organismes, etc.). La plateforme protéomique de l'Institut Pasteur a pour vocation l'étude des protéomes de pathogènes (virus, bactéries, etc.) et des interactions de ces pathogènes avec leur hôte. Dans ce but, elle travaille en collaboration avec différents laboratoires de recherche de l'Institut Pasteur. L'expérience utilisée le plus couramment se base sur la spectrométrie de masse et consiste à réaliser des analyses « bottom-up » (les protéines sont préalablement digérées en peptides pour une meilleure détection). Ces expériences produisent des données complexes et volumineuses qui nécessitent plusieurs étapes d'analyses pour être interprétées.

D'un point de vue statistique, les jeux de données résultant de ces analyses possèdent certaines caractéristiques particulières. Ainsi, on mesure des dizaines de milliers de peptides à partir de seulement quelques échantillons. Il en résulte des matrices d'intensités mesurées qui possèdent des dizaines de milliers de lignes (les peptides) et seulement quelques colonnes (les échantillons). D'autre part, il peut arriver que ces matrices soient entachées de nombreuses valeurs manquantes (allant de 10% à 50%). On est donc dans un cadre statistique « large p, small n » avec généralement une forte proportion de valeurs manquantes.

Un cas particulier d'étude concerne l'évolution d'un protéome dans diverses conditions biologiques. Par exemple, on peut être amené à étudier l'évolution d'un protéome au fur et à mesure du temps. Dans ce cadre, on cherche à classifier les protéines évoluant de manière similaire entre les différentes conditions. L'approche « classique » utilisée dans le domaine est assez simple. On remplace les valeurs manquantes par des petites valeurs, on somme les intensités des peptides associés à une même protéine, puis on réalise un clustering hiérarchique « classique » sur ces intensités sommées [1]. Cette approche ne prend pas en compte l'aspect « large p, small n » [2] [3]. De plus, l'inférence des abondances de protéines à partir de peptides ainsi que le problème des valeurs manquantes sont des problématiques qui complexifient le clustering.

Encadré par un statisticien de l'Institut, vous devrez :

1) Réaliser une revue de littérature sur le sujet du clustering de données protéomiques dans le cadre de plusieurs conditions biologiques.
2) Réfléchir et proposer des solutions nouvelles à la problématique.
3) Implémenter et tester les solutions proposées en R. A partir de vrais jeux de données, comparer les solutions proposées à des approches classiques.
4) Développer une petite interface Shiny « clique-bouton » à partir de R permettant de réaliser les clustering proposés.
5) Présenter les résultats obtenus de manière intelligible à un public « non-statisticien ».

Au cours de ce stage, vous aurez l'occasion d'améliorer vos compétences en statistiques/analyse de données de grandes dimensions, vos compétences en programmation R et notamment dans la création de petites interfaces utilisables par des « non-statisticiens ».

[1] Tyanova et al., *The Perseus computational platform for comprehensive analysis of (prote)omics data,* Nature methods (2016), 731-740
[2] Pan W., Shen X., *Penalized Model-Based Clustering with Application to Variable Selection*, Journal of Machine Learning Research 8 (2007), 1145-1164

[3] Wang S., Zhu J., *Variable Selection for Model-Based High-Dimensional Clustering and Its Application to Microarray Data,* Biometrics (2008), 440-448