

Stage de 6 mois (M2): Intégration de données hétérogènes

Lieu de stage: Hub de bioinformatique / biostatistique, Institut Pasteur, Paris, France

Contexte du stage

Avec la multiplication des technologies de séquençage et d'acquisition de données biologiques, de plus en plus d'études proposent plusieurs points de vue sur une même problématique en mesurant des caractéristiques différentes -comme par exemple sur le génome, le transcriptome, le protéome ou l'épigénome- sur le même échantillon. Cela entraîne la production de jeux de données complexes et souvent multimodaux. En parallèle, les résultats de nombreuses expériences sont rassemblés dans des bases de données publiques, ce qui permet de mettre en relation des éléments biologiques en fonction de critères variés. Par exemple, les bases de données KEGG ou Hallmark permettent, entre autres, de créer des groupes de gènes en fonction de leur fonction biologique.

Dans le cadre de ce stage, le/la candidat-e participera de façon active au développement de nouvelles méthodes permettant l'intégration de structures de données complexes (comme des groupes de gènes) dans des méthodes classiques d'analyse de données multivariées : l'Analyse en Composantes Principales, l'Analyse des Corrélations Canoniques, la régression PLS etc. De plus, RGCCA [1] étant la méthode de choix pour l'intégration de données comportant plusieurs blocs, les développements effectués pendant ce stage pourraient également lui être appliqués.

Les données qui seront analysées incluent des données publiques issues du projet TCGA ([The Cancer Genome Atlas](#)), les données du projet [Milieu Intérieur](#) ou les données du projet [Afribiota](#).

Le stage se déroulera à l'Institut Pasteur, dans les locaux du département de biologie computationnelle (<https://c3bi.pasteur.fr/>).

Profil recherché

Statisticien/ne, data analyst ou data scientist, le/la candidat/e doit avoir de solides compétences en analyse de données multivariées et de bonnes compétences en programmation R et maîtriser l'utilisation de R Markdown.

Le/la candidat/e doit également avoir un intérêt pour l'application des méthodes statistiques à l'analyse de données biologiques, et pouvoir faire preuve d'autonomie.

Contact

Envoyez CV et lettre de motivation à vincent.guillemot@pasteur.fr avant le 15 janvier 2020.

[1] Tenenhaus, A., & Tenenhaus, M. (2014). Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. *European Journal of Operational Research*. <https://doi.org/10.1016/j.ejor.2014.01.008>