

Intégration de jeux de données protéomiques et PTM-omiques (stage)

Entreprise : Institut Pasteur, Paris

Période : 6 mois, printemps 2025

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.
- Stage de fin d'étude.
- 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 :

Contexte : Le protéome est l'ensemble des protéines présentes dans des échantillons biologiques (cellules, organismes, etc.). La plateforme protéomique de l'Institut Pasteur a pour vocation l'étude à grande échelle 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 à la plateforme 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 volumineuses qui nécessitent plusieurs étapes d'analyses pour être interprétées.

D'un point de vue statistique, les jeux de données produits possèdent certaines caractéristiques particulières. Ainsi, on mesure des dizaines de milliers de peptides dans seulement quelques échantillons (données de type « large p, small n »). Il en résulte des matrices d'intensités mesurées qui possèdent des dizaines de milliers de lignes (les peptides) et seulement quelques colonnes (les échantillons). Elles peuvent être entachées de valeurs manquantes (survenant lorsque la protéine n'a pas été détectée dans un échantillon par exemple).

Un sujet d'étude particulier concerne les modifications post-traductionnelles (« Post-translational modifications », ou PTM) de protéines. Ces modifications, notamment la phosphorylation, jouent un rôle clé dans la communication cellulaire. Leurs variations d'abondance peuvent perturber les signaux cellulaires, entraînant des désordres et des maladies. Comparer l'évolution d'un protéome non modifié à celle d'un phospho-protéome (ou autre PTM-ome) sous différentes conditions biologiques permet de déterminer si l'abondance de la modification est liée à celle de la protéine d'origine au fil des conditions. En cas d'absence de différence, on peut conclure que les conditions biologiques n'affectent pas

l'abondance des modifications, et donc les signaux cellulaires. Ce type d'analyse aide à la compréhension du fonctionnement de maladies infectieuses et, in fine, à identifier des processus cellulaires à cibler pour améliorer le traitement de patients infectés.

Contenu du stage : Sous la supervision d'un ingénieur de recherche statisticien de l'Institut, vous contribuerez au développement d'un programme en R, ainsi que de son interface en Shiny, pour simplifier ce type d'analyses au sein de la plateforme protéomique. Une première version de l'application est déjà en cours d'élaboration, mais plusieurs défis restent à relever, notamment la mise en place de méthodes statistiques pour plus de deux conditions biologiques et pour la visualisation des résultats. Ces améliorations sont essentielles pour rendre l'application pleinement opérationnelle dans le cadre des activités quotidiennes de la plateforme. Une comparaison exhaustive de méthodes d'intégration de ce type de données est à mener pour déterminer quelle approche statistique est la plus optimale à utiliser.

Nous recherchons un(e) candidat(e) motivé(e) par cette problématique biologique concrète, par le développement de méthodes statistiques en R et par la création d'interfaces intuitives en Shiny, permettant d'exporter des résultats sous forme de tableaux Excel, fichiers PowerPoint ou Word. Idéalement, le/la candidat(e) possède de solides connaissances en mathématiques/statistiques, notamment en modèles de régression, clustering et analyses multivariées, et des connaissances en R, qu'il/elle pourra approfondir durant le stage.