Proposition de Thèse de Doctorat

Machine learning based approaches for multi-omics data in personalized treatment of sepsis

Approches basées sur l'apprentissage automatique des données multi-omiques pour le traitement personnalisé du sepsis

Université de Versailles Saint-Quentin – Université Paris-Saclay

Description

Contexte général :

Ce sujet de thèse d'inscrit dans le projet RHU RECORDS (Rapid Recognition of Corticosteroid Resistant or Sensitive Sepsis) qui vise à identifier et à valider des biomarqueurs de prédiction de la réponse thérapeutique aux corticoïdes dans le cadre du sepsis. Parmi ces biomarqueurs, les "omiques" ont actuellement une place particulière car elles permettent l'exploration approfondie et non biaisée, libre de toute hypothèse préalable, d'une couche (ou "layer") d'un système vivant.

Les "omiques" parmi les plus analysées en médecine sont :

- la génomique, c'est-à-dire dans le contexte de cette discussion, l'étude des variants héréditaires de l'ADN,
- la transcriptomique, ou l'étude des transcrits, qui sont les premiers produits de l'expression de l'ADN au niveau des gènes,
- la protéomique, ou l'étude des protéines, obtenues après la traduction des transcrits,
- la métabolomique, qui étudie les métabolites résultant de l'activité des protéines.

Ces omiques, dont la liste présentée n'est pas limitative, correspondent en fait au flux d'information génétique qui part de l'ADN et aboutit au phénotype final et contribue, avec les facteurs de l'environnement, à sa variabilité inter-individuelle. La compréhension des mécanismes de cette variabilité est essentielle car c'est elle qui détermine la susceptibilité individuelle aux maladies et la réponse des patients aux traitements de ces maladies.

La contrepartie de la puissance des méthodes omiques et du caractère massif des données qu'elles génèrent réside dans la complexité de leur analyse. Si l'analyse des omiques considérées une par une (single omics) a fait des progrès considérables sur le plan méthodologique, il peut arriver, par exemple pour des raisons de puissance statistique, que les résultats obtenus soient peu concluants.

Il faut aussi constater que chacune des couches omiques, même si elle est étroitement connectée à celle qui la précède et à celle qui lui succède dans le flux de l'information génétique et du métabolisme, possède sa propre dynamique et ne livre qu'une vision partielle du système vivant global.

Enfin, de nouveaux transcrits, gènes, métabolites continuent d'être découverts et la compréhension de leur signification devrait être significativement enrichie par l'analyse non seulement de la couche omique à laquelle ils appartiennent, mais aussi des autres couches.

Motivations et objectifs du projet :

Dans les différentes situations citées ci-dessus, l'analyse simultanée des différentes couches omiques, ou analyse multi-omique, devrait présenter un intérêt considérable. Il est en effet vraisemblable que la caractérisation des relations ou interactions entre les différentes couches éclairent chacune des couches individuelles d'un jour nouveau voire inattendu. Toutefois, la méthodologie de ces approches multi-omiques reste encore largement du domaine de la recherche. Les méthodes classiques d'inférence statistique qui sont bien souvent suffisantes pour les analyses uni-omiques sont rarement opérationnelles pour les analyses multi-omiques. Il y a, de fait, relativement peu de librairies disponibles dans ce but. Parmi elles, on cite Cluster+, MOFA et mixOmics qui présentent les librairies les plus utilisées dans l'analyse multi-omiques, d'où l'intérêt suscité par les méthodes issues de l'intelligence artificielle (IA), et notamment l'apprentissage automatique, qui ont fait leur preuve dans de nombreux domaines où des données massives sont à interpréter. Ainsi, l'objectif principal de ce sujet de thèse est d'étudier les méthodes existantes de l'IA dans l'analyse de données omiques obtenues dans le cadre du RHU RECORDS, et de proposer de nouvelles méthodes en fonction des limites qui seront identifiées.

Spécifiquement, le projet RECORDS s'appuie sur trois de ces omiques, la génomique, la transcriptomique et la métabolomique, qui sont parmi les plus robustes et de loin les plus utilisées dans l'exploration des maladies humaines. Dans ce cadre, la recherche de biomarqueurs (pas seulement omiques) de prédiction de la réponse aux corticoïdes dans le sepsis peut se faire avec deux perspectives :

Dans un premier cas, on recherche des marqueurs différentiellement représentés entre les différents groupes de patients (par exemple, patients sensibles ou non au traitement par les corticoïdes). On parle d'apprentissage supervisé. Dans ce premier cas toujours, une situation particulière est celle où l'on dispose d'une liste de marqueurs trop longue pour être exploitée dans la pratique médicale quotidienne. Il faut alors réduire cette liste en sélectionnant les variables les plus pertinentes qui permettront de développer des tests rapides, faits au lit du patient, pour une prise de décision thérapeutique dans les heures suivant son hospitalisation. La réduction de dimensions est un problème classique en analyse de données et en apprentissage. Cependant, il existe des méthodes spécifiques pour les données multi-omiques, lesquelles intègrent les différentes couches. Un benchmark de ces méthodes a été publié récemment dans le contexte du cancer (Cantini et al., 2021).

Dans un second cas de figure, l'objectif est de regrouper les patients sans connaissance préalable des variables pertinentes. Dans ce type d'apprentissage dit non supervisé, on cherche à reconnaître l'hétérogénéité des patients. Par exemple, on peut chercher à identifier les mécanismes de corticorésistance qui sont certainement multiples et hétérogènes en analysant les différentes couches omiques, de manière individuelle ou combinée. Là, il existe des méthodes statistiques dédiées pour la tâche de regroupement (de clustering) mais qui sont d'efficacité limitée lorsque les données sont multimodales et clairsemées et l'apprentissage automatique devrait constituer une approche très fructueuse.

Méthodologie de travail :

Comme précédemment indiqué, l'objectif principal de cette thèse est d'étudier les méthodes existantes de l'apprentissage automatique dans l'analyse de données omiques obtenues dans le cadre du RHU RECORDS, et de proposer de nouvelles méthodes en fonction des limites qui seront identifiées. La méthodologie de travail qui sera suivie par le doctorant(e) est comme suit :

1. Étudier les méthodes de l'état de l'art dédiées à la réduction de dimensions pour les données multi-omiques de RHU RECORDS.
2. Identifier les limites de celles-ci et proposer des nouvelles méthodes qui permettent de les pallier.
3. Étudier les méthodes de l'état de l'art dédiées au regroupement afin de découvrir des relations ou interactions entre les patients.
4. Identifier les limites de celles-ci et proposer des nouvelles méthodes qui permettent d'y remédier.
5. Proposer et implémenter, après avoir fait une étude de l'existant, un pipeline bio-informatique permettant l'intégration et la visualisation des données multi-omiques.
6. Résumer le travail réalisé sur chaque partie dans un document de recherche scientifique qui sera soumis à une conférence/une revue internationale.

Références bibliographiques :

1. Krassowski, Michal, et al. "State of the field in multi-omics research: From computational needs to data mining and sharing." Frontiers in Genetics 11 (2020).
2. Cantini L, Zakeri P, Hernandez C, Naldi A, Thieffry D, Remy E, Baudot A. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nature Commun. 2021 Jan 5;12(1):124. doi: 10.1038/s41467-020-20430-7.
3. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Briefings in bioinformatics. 2016 Jul 1;17(4):628-41.
4. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: A review. Biotechnology Advances. 2021 Mar 29:107739.

Profil et expertises recherchés :

La candidate ou le candidat doit être titulaire d'un Master en informatique ou bio-informatique. Elle/il devrait avoir :

- Solides compétences en programmation objet et système et bases de données
- Bonne expérience en statistiques
- Bonne expérience en apprentissage automatique (*machine learning*)
- Bonne expérience en bio-informatique (souhaitable mais non obligatoire)
- Bonne communication orale en anglais, compétences techniques en lecture et en écriture ;
- La maîtrise du français est souhaitable mais pas obligatoire.

Procédure de candidature :

Prendre contact avec les encadrants au plus tôt en joignant :

- un CV détaillé
- une lettre de motivation
- les relevés de notes des deux dernières années universitaires
- des lettres de recommandation s'il y en a.

Date limite de candidature :

Avant de soumettre la candidature sur le site ADUM, il est impératif d'envoyer les documents par mail aux contacts donnés ci-dessous et ce **au plus tard le 30 Avril 2022.**

Encadrements & contacts :

- Laboratoire DAVID : Prof. Karine ZEITOUNI, Karine.Zeitouni@uvsq.fr, Directrice; Dr. Zaineb CHELLY DAGDIA, zaineb.chelly-dagdia@uvsq.fr, co-encadrante.
- Prof. Henri-Jean Garchon, henri-jean.garchon@uvsq.fr, Prof. Stanislas Grassin Delyle, stanislas.grassin-delyle@uvsq.fr, co-encadrants.

Laboratoires d'accueil :

- DAVID/équipe ADAM, Campus de sciences à Versailles, Université de Versailles St-Quentin UVSQ / Université Paris-Saclay.
- Inserm UMR 1173 Laboratoires II et LARENE, UFR Simone Veil, Montigny-le-Bretonneux, UVSQ / Université Paris-Saclay