Avis de Soutenance

Monsieur David HIRST

Biologie-Santé - Spécialité Bioinformatique et Génomique

Soutiendra publiquement ses travaux de thèse intitulés

Amélioration de l'inférence de variables latentes à partir de données omiques grâce à l'utilisation d'information a priori

dirigés par Madame Anaïs BAUDOT

Soutenance prévue le ***lundi 25 novembre 2024*** à 14h00
Lieu : 163 Avenue de Luminy, Case 901, Marseille, 13009
Salle : HEXAGONE - AUDITORIUM

Composition du jury proposé

Mme Anaïs BAUDOT	Aix Marseille Université, Marseille Medical Genetics,	Directrice de thèse
M. Arthur TENENHAUS	CentraleSupélec, Le laboratoire des signaux et systèmes (L2S)	Rapporteur
Mme Andrea RAU	INRAE, GENETIQUE ANIMALE et BIOLOGIE INTEGRATIVE (GABI)	Rapporteure
M. Jacques VAN HELDEN	Aix Marseille Université	Président
M. Carl HERRMANN	Heidelberg University, Institute for Pharmacy and Molecular Biotechnology	Examinateur
M. Matthieu VIGNES	Massey University, School of Mathematical and Computational Sciences	Invité

Mots-clés : Bioinformatique/Biologie Computationelle,Biologie des Systèmes,Intelligence artificielle,Maladies rares,Reduction de dimension,Apprentissage par transfert

Résumé :

Une grande variété de processus biologiques sont nécessaires au développement et à la survie des organismes vivants. Ces processus biologiques dépendent d'un grand nombre de molécules biologiques. Les technologies à haut débit permettent d'établir les profils des molécules dans les échantillons, générant ainsi les données dites "omiques" qui permettent de mieux comprendre les cellules, les organes et les organismes. Les données omiques sont généralement de haute dimension. Elles se composent d'un grand nombre de molécules biologiques, les variables observées, qui peuvent ne pas être informatives pour la compréhension des conditions biologiques d'intérêt. Dans ce contexte, les approches de réduction de dimension non-supervisée sont populaires pour l'analyse des données omiques. Ces approches permettent de déduire un nombre réduit de variables latentes représentant potentiellement les processus biologiques et aidant à l'identification des variables observées informatives. De plus, différents types de données omiques

permettent de charactériser les processus biologiques à différentes échelles à l'intérieur et entre les cellules. Ainsi, il est attendu que l'analyse de données multi-omiques permette de mieux comprendre les systèmes biologiques par rapport à l'analyse de données omiques simples. Une méthode efficace pour déduire les variables latentes des données multi-omiques est la factorisation conjointe de matrices multi-omique, une méthode non supervisée de réduction des dimensions. Cependant, dans certains cas, en raison du petit nombre d'échantillons disponible, les méthodes de factorisation de matrice peuvent avoir du mal à séparer les signaux biologiques latents en une représentation pertinente. Il s'agit d'un défi particulier dans l'étude des maladies rares, pour lesquelles il peut être difficile d'obtenir des données multi-omiques sur une nombre suffisant d'échantillons. Dans le cadre de mon doctorat, j'ai développé MOTL, une approche d'apprentissage par transfert pour la factorisation de matrices multi-omique. MOTL effectue une factorisation de matrices sur un ensemble de données cible, généré à partir d'un petit nombre d'échantillons, en incorporant des informations déduites de la factorisation multi-omique effectuée sur un ensemble de données d'apprentissage, généré à partir d'un grand nombre d'échantillons caractérisés par diverses conditions biologiques. Dans les évaluations de MOTL, j'ai démontré que l'apprentissage par transfer peut améliorer la capacité de la factorisation de matrices multi-omique à découvrir des processus associés à une condition biologique d'intérêt. Au cours de l'un de mes autres projets de doctorat, j'ai également exploré l'utilisation d'informations a priori pour améliorer l'inférence de variables latentes à partir de données omiques. J'ai mis en œuvre et évalué une modification d'une architecture d'autoencodeur variationnel, un autre type de méthode de réduction de dimension non supervisée. Cette modification a permis d'inclure des poids contrôlant la contribution relative des variables observées à la fonction de perte. J'ai observé que l'inclusion de poids basés sur des informations préalables permettait d'obtenir des valeurs de variables latentes qui différenciaient mieux les groupes d'échantillons. Enfin, j'ai participé à une collaboration portant sur l'analyse de différents types de données omiques provenant de patients atteints de deux maladies génétiques rares différentes et de témoins sains. Dans ce projet, différents types de données omiques ont été analysés séparément avec des tests différentiels, ce qui a permis d'obtenir des sous-ensembles de variables observées potentiellement associées à ces maladies. Avec mes collaborateurs, nous avons utilisé des informations préalables, sous la forme de ressources d'annotation, pour convertir ces listes de variables observées pertinentes en listes de processus biologiques pertinents.

LE DOYEN

Georges LEONETTI