Journée thématique du groupe « Statistique & Génomique »

du Réseau Interdisciplinaire autour de la Statistique

« APPRENTISSAGE AUTOMATIQUE SUR DONNEES OMIQUES »

à l'Institut des Systèmes Complexes.

Adresse complète : ISC-PIF. 113 rue Nationale, 75013, Paris.

Salles 1.1 et 1.2

JEUDI 6 OCTOBRE 2022

10h – 10h30 : **Accueil & café**

10h30 – 11h30 : "RESEAUX DE NEURONES INTERPRETABLES POUR LA PREDICTION DE PHENOTYPE A PARTIR D'EXPRESSION DE GENES"

Blaise Hanczar, labotatoire 'Informatique, BioInformatique, Systèmes Complexes',
Département Computer Science, Université Evry-Paris-Saclay

L'apprentissage profond est une avancée majeure de l'intelligence artificielle de ces dernières années et l'un de ses principaux enjeux est son application à la médecine de précision. Dans ce contexte, nous montrons comment utiliser les réseaux de neurones profonds pour la prédiction de diagnostic, prognostique ou réponse aux traitements de patients à partir de leur profil d'expression de gènes. Les deux principaux verrous scientifiques pour cette tâche sont la petite taille des jeux de données disponibles et le manque de transparence des réseaux de neurones. Par une série d'expériences, nous montrons que l'apprentissage par transfert est une piste très prometteuse pour pallier le premier problème. Pour le deuxième, nous proposons d'intégrer de la connaissance du domaine (Gene Ontology) dans le modèle afin de rendre ses prédictions interprétables.

11h30 – 12h15 : " IDENTIFICATION DES BIOMARQUEURS PROGNOSTIQUES DANS LE CANCER A PARTIR DE DONNEES D'EXPRESSION DE GENES"

Ekaterina Flin, Institut pour l'Avancée des Biosciences, Université Grenoble Alpes

Les mesures de niveaux d'expression de certains gènes dans les cancers permettent de prédire le pronostic vital des patients atteints de ces maladies. Grâce à ces gènes dit "biomarqueurs", il devient donc possible d'identifier les formes particulièrement agressives de cancer à des stades encore précoces et d'adapter le traitement en conséquence. Les gènes candidats "biomarqueurs" doivent répondre à plusieurs exigences dont les plus importantes sont leur efficacité, la robustesse et la reproductibilité des résultats dans différentes populations de patients. De plus, leur utilisation doit être compatible avec des technologies applicables en routine en clinique et les tests effectués à des coûts raisonnables.
Le groupe EpiMed de l'Institut pour l'Avancée des Biosciences (IAB) mène depuis plus de 10 ans des travaux sur l'épigénétique et le cancer. Il a développé une approche originale qui permet d'identifier des biomarqueurs pronostiques dans les cancers à partir de données d'expression de gènes. Cette approche se base sur l'existence de dérégulations majeures de l'expression des gènes dans les cellules cancéreuses, qui résultent elles-mêmes d'anomalies multiples affectant le système de balisage, l'épigénome, qui normalement contrôle le niveau d'expression des gènes. Le groupe EpiMed a notamment démontré que certains gènes normalement exprimés dans un nombre restreint de cellules, et silencieux dans toutes les cellules adultes non germinales, peuvent être anormalement activés dans la tumeur. De plus, l'activation anormale de certains de ces gènes peut être associée à un pronostic vital défavorable (survie globale ou survie sans rechute).
La méthode proposée intègre ces connaissances biologiques dans un pipeline d'apprentissage automatique pour définir d'une façon robuste des seuils d'activation des gènes biomarqueurs potentiels et leur impact sur la probabilité de survie de patients. Les meilleurs gènes candidats sont ensuite combinés dans un outil pronostic et validés dans plusieurs cohortes indépendantes. L'outil final s'adapte à plusieurs technologies (par exemple, RNA-seq, microarrays, RT-qPCR et immunohistochimie) et peut être utilisé en clinique.

12h15 - 13h30 : Pause déjeuner

13h30 – 14h30 : " INFERRING THE 3D STRUCTURE OF THE GENOME FROM HI-C DATA / L'INFERENCE DE LA STRUCTURE TRI-DIMENSIONNELLE DU GENOME "

Nelle Varoquaux, Laboratoire 'Recherche Translationnelle et Innovation en Médecine et Complexité', Equipe Translational microbiology, Evolution and Engineering

The spatial and temporal organization of the 3D structure of chromosomes is thought to have an important role in genomic function, but is poorly understood. Advances in chromosome conformation capture (3C) technologies, initially developed to assess interactions between specific pairs of loci, allow one to simultaneously measure multiple contacts on a genome scale, paving the way for more systematic and genome-wide analysis of the 3D architecture of the genome. These new Hi-C techniques result in a genome-wide contact map, a matrix indicating the contact frequency between pairs of loci. This matrix can be used to analyze the three-dimensional structure of the genome.

However, despite extensive research, inferring a three-dimensional model from this contact map remains a fundamental problem.
I will discuss here statistical based methods to infer a consensus 3D model of the structure.

14h30 – 15h30 : "METHODES D'APPRENTISSAGE POUR LA MODELISATION DE SEQUENCES REGULATRICES"

Laurent Bréhélin, *Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier* - Campus St Priest, Montpellier

L'étude des liens entre l'ADN et l'expression des gènes a une longue histoire en bioinformatique. Notamment, de nombreuses approches ont été proposées dans le passé pour identifier les motifs associés aux sites de fixation des facteurs de transcription. Ces travaux se sont toutefois généralement limités à la modélisation de sites de fixation uniques, sur la base de matrices poids-positions (PWM) ou de modèles proches. Bien que très largement utilisés aujourd'hui, ces modèles ne permettent cependant pas de modéliser toute la complexité de la régulation transcriptionnelle, et notamment la spécificité cellulaire de cette régulation. Ces dernières années, de nouvelles méthodes d'apprentissage ont été proposées pour modéliser des régions régulatrices entières, en allant bien au-delà des sites de fixation individuels. Dans cet exposé, je présenterai des travaux de notre groupe sur cette thématique, et notamment une méthode d'apprentissage qui a pour objectif d'identifier les déterminants génomiques impliqués dans la spécificité cellulaire de la fixation des facteurs de transcription.
https://www.biorxiv.org/content/10.1101/2022.08.16.504098

15h30 – 16h15 : " HOW TO SAVE AN AWFUL-DESIGNED PROJECT? - WITH THE RECOURSE OF MACHINE LEARNING APPROACHES FOR MULTI-OMICS ANALYSIS "

Lijiao Ning, (Epi)génomique Fonctionnelle et Physiologie Moléculaire Du Diabète et Maladies Associées, Université de Lille

What to do if the single omics analysis does not or cannot give any "significant" evidence related to the study design? The machine learning approaches provide different perspectives on data exploitation in an integrated way. This talk will cover the application of unsupervised clustering and multi-omics analysis (especially the usage of the R-package "mixOmics") in a real-live project.

16h15 – 16h30 : Mot de la fin

Nous remercions l'ensemble des intervenants pour la richesse de cette journée, l'Institut des Systèmes Complexes pour son accueil et la Mission pour les Initiatives Transverses et Interdisciplinaires pour son soutien.