Intitulé de l'offre de stage

CUP classifieur

Stagiaire en	Informatique (Deep Learning)
Affectation	Équipe commune CRCT/IRIT
Durée	5 à 6 mois
Rémunération	environ 670 € net mensuel
Date de la publication	1/12/2023
Date d'embauche prévue	mars / avril 2024 suivant disponibilités
Lieu	CRCT, 2 Avenue Hubert Curien – 31100 Toulouse

Le Centre de Recherche en Cancérologie de Toulouse

Le CRCT est une unité de recherche conjointe entre l'Inserm et l'Université Toulouse III Paul Sabatier. Au cœur de l'Oncopole de Toulouse, le CRCT, avec tous ses partenaires (institutionnels, universitaires, cliniques, industriels, caritatifs…), stimule l'innovation en termes de recherche et d'enseignement dans la lutte contre le cancer. Le CRCT conduit une approche intégrée entre la recherche, les soins et l'enseignement, dans une logique transversale et multidisciplinaire.

L'Institut de Recherche en Informatique de Toulouse

L'Institut de Recherche en Informatique de Toulouse (IRIT), une des plus imposantes Unité Mixte de Recherche au niveau national, est l'un des piliers de la recherche en Occitanie avec ses 700 membres, permanents et non-permanents. De par son caractère multi-tutelle (CNRS, Universités toulousaines), son impact scientifique et ses interactions avec les autres domaines, le laboratoire constitue une des forces structurantes du paysage de l'informatique et de ses applications dans le monde du numérique, tant au niveau régional que national.

L'équipe

L'informatique et l'intelligence artificielle sont appelés à prendre une place croissante dans le monde de la recherche médicale, et en particulier dans le monde de la recherche contre le cancer. L'équipe commune IRIT/CRCT, co-localisée sur le site de l'Oncopole de Toulouse, a pour but de faire travailler ensemble et sur le même site chercheurs en informatique, chercheurs dans le domaine du cancer et médecins. Elle peut s'appuyer sur le plateau technique du CRCT (analyses biologiques, séquençages génomique,...) et sur les moyens de calcul de la région Occitanie (CALMIP).

Objet du stage

Les tumeurs humaines sont très variées. Par exemple, derrière le terme grand public « cancer du sein » se cachent plus d'une vingtaine d'entités tumorales différentes, fréquentes ou rares. L'identification de chaque entité est capitale, car elle a des implications pronostiques et thérapeutiques. L'identification des types tumoraux repose sur l'analyse au microscope de la morphologie des cellules anormales, effectuée par des médecins spécialistes, les pathologistes. Ces dernières années, l'OMS a intégré des mutations de l'ADN tumoral au diagnostic microscopique, rendant la classification plus fiable et plus précise (exemple de diagnostic intégré : astrocytome IDH muté). Plus récemment encore, l'OMS a intégré une technique de classification des tumeurs cérébrales (brain classifier), développée par l'Université de Heidelberg, se basant sur le profil de méthylation de l'ADN tumoral et utilisant le Random Forest.

Le principe du méthylome est le suivant : sur les milliards de bases (A,T, C, G) de notre ADN, 28 millions de Cytosine (C), toujours à côté d'une Guanine (G) (on parle de doublets CpG) peuvent-être méthylées ou non. Cette méthylation régule la transcription des gènes. En particulier, si plusieurs CpG au niveau d'un promoteur (séquence d'ADN qui peut initier l'expression d'un gène) portent des cytosines méthylées, un blocage de la transcription du gène est généralement observé. Les cellules tumorales détournent ce système pour inactiver des anti-oncogènes et activer des oncogènes. A partir d'un set de gènes choisis pour leur implication dans le cancer, des puces d'hybridation pour étudier la méthylation des cellules tumorales ont été développées. Le pattern de méthylation de 450 000 CpG suffit amplement à identifier un type tumoral. Voici le lien vers une conférence de vulgarisation que nous avions faite sur ce sujet et qui est disponible en replay sur inscription.

Cette démarche de classification par le méthylome est en train de se dévélopper pour d'autres tumeurs (ORL, tissus mous etc...). Des fichiers informatiques (.idat) correspondant aux données brutes du pattern de méthylation sont disponibles en open source pour de nombreux cancers. Certaines tumeurs sont complètement indifférenciées et les pathologistes sont en échec pour les typer, en particulier dans le contexte de patients multi-métastatiques dont la tumeur primitive n'est pas retrouvé (Carcinoma of Unknown Primary : CUP syndrome). Le traitement ne peut alors être que probabiliste. Nous voulons développer un CUP classifier pour pouvoir attribuer ces tumeurs indifférenciées à un organe et un type tumoral. Nous avons commencé lors d'un premier stage à recueillir 6 000 fichiers .idat (par webscrapping) de différents cancers épithéliaux situés dans différents organes. A partir de cette base de données organisée en dataframe , nous avons commencé à évaluer l'efficience de différents algorithmes (Random forest, LOGREG, SVM) pour prédire l'organe d'origine et le type de tumeur. Il faudra :

> tester d'autres algorithmes : XGBoost, réseaux de neurones en privilégiant des méthodes permettant une évaluation/interprétation des résultats (Shapley Value...)

- optimiser le fonctionnement (i) en améliorant la base de données : équilibrer les classes, (ii) paralléliser l'utilisation des CPU avec Spark (iii) enrichir la dataframe avec les anomalies chromosomiques.
- valider les résultats en validation croisée et comparer les résultats obtenus avec la clusterisation en t-SNE, u-MAP ou ACP.
- identifier les CpG les plus importants en utilisant l'ACP, le clustering hiérarchique et la sélection de features (Boruta...) pour répondre à des questions biologiques et également permettre une utilisation de cet outil avec des technologies plus légères et plus rapides.

Formation

École d'ingénieur, de préférence avec spécialisation en informatique ou mathématiques (en année de césure ou stage long). Master 2 informatique ou mathématiques appliquées.

Compétences attendues

Des connaissances en science des données, apprentissage et réseaux de neurones seront appréciées.

Modalité de candidature

- CV à envoyer à: cup@stages-medecine-numerique.fr
- Date limite de candidature : 31/01/2024
- Encadrants: Professeur Emmanuelle Uro-Coste (CHU de Toulouse/IUCT-Oncopole/CRCT), Sandrine Mouysset (IRIT), Daniel Ruiz (IRIT)

Une première sélection sera effectuée sur la base des CVs reçus. Les candidats seront immédiatement informés du résultat, et ceux qui seront retenus à l'issue de la première sélection seront invités à un ou deux entretiens individuels en téléconférence avant sélection définitive.

L'ensemble des offres de stage est disponible sur http://www.stages-medecine-numerique.fr.