Stages en Deep Learning
pour la médecine personnalisée

Mots-clés : Deep learning, GAN, Data augmentation, Données génomiques, Diagnostique médicale.

Sujet : Génération de données transcriptomiques combinées à partir de modèles génératifs d'apprentissage profond

Les données issues de lignées cellulaires sont utilisées en recherche biomédicale pour la prédiction de phénotypes. Elles sont notamment utilisées pour la prédiction du cancer à partir du niveau d'expression des gènes (analyse transcriptomique). Cependant, les lignées cellulaires de cellules cancéreuses sont très hétérogènes et cette complexité n'est capturée qu'en agrégeant l'information au niveau d'une population de cellules (données *bulk*). Ces deux types de données renferment l'information d'expression des gènes dans plus de 20 000 variables, mais restent en quantité limitée (coût des méthodes de séquençage, manque de patients, etc.).

L'objectif de ce stage est de développer une méthode de génération de données à partir de modèles génératifs d'apprentissage profond, en combinant les données au niveau cellulaire et au niveau du tissu. Les modèles génératifs de l'état de l'art, tels que les GANs [1] et les modèles de diffusion [2,3], peuvent être conditionnés par un type de tissu ou un type de cancer pour générer un échantillon transcriptomique correspondant [4]. Ils sont basés sur des architectures de *deep learning* qui sont capables d'apprendre et d'extraire l'information complexe et non linéaire entre les gènes.

Les enjeux de concevoir un générateur de données transcriptomiques sont triples :
1) Les jeux de données sont généralement de petite taille car chers à produire, un générateur permettrait de faire de **l'augmentation de données** et donc d'améliorer les performances des modèles de prédiction.
2) En fournissant des données artificielles grâce au générateur, les chercheurs pourront développer des modèles sans avoir besoin d'accéder aux données réelles et donc préserver la **confidentialité** des données patients.
3) A l'aide d'un générateur il sera possible de **simuler des interventions** sur l'expression des gènes et d'identifier de potentielles cibles thérapeutiques pour certaines maladies.

Profil recherché :

- Étudiant Master 2 Recherche ou en dernière année d'école d'ingénieur de formation informatique ou mathématiques appliquées.
- Connaissances des librairies de machine learning (Scikit-learn) et de deep learning (Pytorch)

- De solides compétences en apprentissage profond sur une diversité d'architectures
- Autonomie et curiosité pour la recherche scientifique

Votre candidature :

- Type de contrat : Stage 5-6 mois
- Lieu : Évry
- Début du contrat : 2024
- Envoyez votre CV et les relevés de notes de vos deux derrières années à blaise.hanczar[at]univ-evry.fr

Références :

[1] Goodfellow, Ian, et al. "Generative adversarial networks." (2014)
[2] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models." *Advances in Neural Information Processing Systems* 33 (2020)
[3] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents.

[4] Lacan A., Sebag M., Hanczar B., GAN-based data augmentation for transcriptomics: survey and comparative assessment, *Bioinformatics*, 2023