

Projet stage 2027

Kevin Dradjat, IBISC, Evry University
kevin.dradjat@univ-evry.fr

2026

Title: Apprentissage auto-supervisé de représentations biologiquement informées pour les données bulk RNA-seq.

Contexte

Les données d'expression de gènes issues du séquençage RNA-seq (bulk RNA-seq) permettent de caractériser l'état biologique de patients et de tissus à partir de l'expression de plusieurs milliers de gènes. Cependant, la forte dimensionnalité et le nombre limité d'échantillons annotés pour certaines tâches rendent l'utilisation de modèles d'apprentissage profond difficile pour l'analyse de ces données. Pour pallier à ce manque d'annotations, les approches récentes d'apprentissage auto-supervisé cherchent à exploiter de grandes collections de données RNA-seq non annotées afin d'apprendre des représentations pouvant être transférées vers différentes tâches biologiques [1]. Des modèles récents adoptent ainsi ces approches avec des architectures de type Transformer afin d'apprendre des représentations informatives et généralisables [3, 2]. Cependant, la forte dimensionnalité de ces données entraîne un coût computationnel important pour le pré-entraînement et l'utilisation de ces modèles Transformer. La conception de méthodes permettant d'exploiter efficacement l'information transcriptomique avec ces modèles, tout en limitant le coût computationnel constitue donc un enjeu important.

Objectif

Une approche consiste à représenter les données à un niveau biologique plus abstrait, en regroupant les gènes au sein de voies biologiques. Cette représentation permet de réduire la dimensionnalité des données tout en intégrant des connaissances biologiques. Ce projet vise ainsi à développer un modèle de fondation de type Transformer entraîné de façon auto-supervisée et exploitant ces voies biologiques afin de produire des représentations latentes généralisables pour les profils transcriptomiques. La transférabilité de ces représentations latentes sera ensuite évaluée sur des tâches biologiques disposant d'une faible quantité de données annotées. L'objectif de ce projet est donc d'étudier dans quelle mesure une représentation au niveau des voies biologiques constitue une alternative pertinente à une représentation au niveau des gènes dans un contexte d'apprentissage auto-supervisé, notamment en terme de qualité des représentations latentes apprises et d'efficacité de transfert.

Méthodes

Le projet débutera par la construction de représentations au niveau des voies biologiques à partir de profils d'expression transcriptomiques. Ces représentations se baseront sur des scores d'activité des voies calculés à l'aide de méthodes telles que ssGSEA, puis utilisées comme entrées du modèle. Le modèle sera pré-entraîné de manière auto-supervisée sur une large collection de données bulk RNA-seq (ARCHS4, GTEx). Plusieurs stratégies de pré-entraînement pourront être étudiées, telles que les

tâches de reconstruction, l'apprentissage contrastif ou encore des stratégies intégrant des connaissances biologiques. Les représentations latentes obtenues seront ensuite transférées vers des tâches supervisées sur des données de la base TCGA (prédiction de cancer, survie). L'évaluation sera réalisée avec différentes proportions de données annotées afin d'étudier les performances dans des conditions de faible disponibilité des données. Les résultats seront comparés à ceux de modèles utilisant une représentation au niveau des gènes, notamment BulkFormer [3], sur les aspects de performances de transfert et de coût computationnel.

Résultats attendus

Ce projet cherche à déterminer si la réduction de dimension, avec les représentations au niveau des voies biologiques, permet de diminuer la complexité et le coût computationnel du modèle tout en conservant des représentations latentes suffisamment informatives. Les expériences réalisées permettront d'étudier l'impact du niveau de représentation des données et de la stratégie de pré-entraînement sur la capacité de transfert.

References

- [1] Kevin Dradjat, Massinissa Hamidi, Pierre Bartet, and Blaise Hanczar. Self-supervised representation learning on gene expression data. *Bioinformatics*, page btaf533, 09 2025.
- [2] Shuai Jiang and Saeed Hassanpour. Transformer-based representation learning for robust gene expression modeling and cancer prognosis. *Scientific Reports*, 15(1):37581, Oct 2025.
- [3] Boming Kang, Rui Fan, Meizheng Yi, Chunmei Cui, and Qinghua Cui. A large-scale foundation model for bulk transcriptomes. *bioRxiv*, 2025.