Méthodes de deep learning pour la prédiction des ARNs longs non-codants. Application au cancer

Mots clés : Bioinformatique, deep-learning, intégration de données, ARN longs non-codants, génomique, cancer, médecine de précision

Contexte

Les ARN, et plus précisément les ARN non-codants (ARNncs, ARN non traduits en protéines), suscitent depuis quelques années un intérêt croissant auprès de la communauté scientifique internationale, de par leur implication avérée dans de nombreux processus biologiques et le rôle important qu'ils peuvent jouer dans des processus pathologiques comme le cancer. Ils sont ainsi de plus en plus considérés comme de potentiels cibles thérapeutiques ou biomarqueurs (marqueurs diagnostiques et pronostiques).

Récemment, de nombreux longs ARNncs (ARNlncs), de taille supérieure à 200 nucléotides, ont été identifiés comme de potentiels régulateurs. Mais contrairement aux petits ARNncs, leur caractérisation par leur structure et leur fonction sont loin d'être établies. La détermination de la structure, 2D ou 3D d'un ARNlnc par des méthodes expérimentales (cristallographie, RMN) ou bioinformatiques est un challenge majeur, puisque cela contribue à élucider sa fonction. Les ARN d'une même famille partagent en effet la même structure, leur conférant la même fonction, la structure guidant notamment les interactions de cet ARN avec des protéines ou d'autres ARN.

Objectifs

Dans ce projet, nous proposons de développer des méthodes computationnelles basées sur du Deep Learning pour prédire et caractériser er les ARNlncs en intégrant différentes sources de données : la séquence, la structure 2D et 3D, l'interaction avec des gènes codants ou non-codants et les altérations génétiques et épigénétiques. Le développement de méthodes pour prédire la structure 3D des ARN, telles que celles développées par DeepMind (la filiale IA de Google), pourra également être envisagé.

Les méthodes développées seront appliquées au cancer et permettront de mieux comprendre l'implication des ARN dans cette pathologie. Un cancer dans un tissu donné est une maladie hétérogène ; plusieurs sous-types de cancers peuvent être identifiés. Les traitements et le diagnostic doivent être adaptés à chaque sous-type. Dans ce projet, nous nous intéresserons aux ARNlncs dans un cancer fréquent, le cancer de vessie (4ème cancer en termes d'incidence chez l'homme) ainsi que dans un cancer pédiatrique, le rétinoblastome. Un petit nombre d'ARNlncs prédits comme potentiellement impliqués seront validés fonctionnellement par l'équipe de biologistes. Nous espérons in fine pouvoir proposer aux cliniciens de nouveaux marqueurs diagnostiques ou pronostiques et leur permettre de mieux comprendre les causes biologiques de la maladie afin d'optimiser les traitements.

L'objectif final du projet sera de mettre en œuvre des méthodes et des outils génériques pour la prédiction des ARNlncs. Les outils développés seront mis à disposition de la communauté scientifique via notre plateforme EvryRNA : http://EvryRNA.ibisc.univ-evry.

Encadrement du doctorant

Ce projet sera réalisé en collaboration étroite entre (i) l'équipe d'accueil, l'équipe AROBAS du laboratoire IBISC, dont l'un des principaux thèmes de recherche concerne la bioinformatique des ARN et le deep learning (ii) la société Adlin, société experte dans l'analyse des données et (iii) l'équipe Oncologie Moléculaire de l'Institut Curie, équipe reconnue internationalement pour l'étude des cancers de vessie et composée de biologistes, cliniciens et pathologistes s'intéressant à ce cancer.

Contact : Fariza Tahi, Pr. IBISC, Université d'Evry, Université Paris-Saclay. **Email : fariza.tahi@univ-evry.fr**

Deep learning methods for long non-coding RNA prediction. Application to cancer

Keywords: Bioinformatics, deep-learning, data integration, long non-coding RNA, genomics, cancer, precise medicine

Context

RNAs, and more precisely non-coding RNAs (ncRNAs, RNA untranslated into proteins), have aroused growing interest in the international scientific community in recent years, due to their proven involvement in many biological processes and the important role they can play in pathological processes such as cancer. They are thus increasingly considered as potential therapeutic targets or biomarkers (diagnostic and prognostic markers).

Recently, many long ncRNAs (lncRNAs), larger than 200 nucleotides, have been identified as potential regulators. But unlike small ncRNAs, their characterization by structure and function is far from established. The determination of the structure, 2D or 3D, of an lncRNA by experimental methods (crystallography, NMR) or bioinformatics methods is a major challenge, since it helps to elucidate its function. RNAs from the same family indeed share the same structure, giving them the same function, the structure guiding in particular the interactions of this RNA with proteins or other RNAs.

Objectives

In this project, we propose to develop computational methods based on Deep Learning to predict and characterize lncRNAs by integrating different data: sequence, 2D and 3D structure, interaction with coding or non-coding genes. and genetic and epigenetic alterations. The development of methods to predict the 3D structure of RNAs, such as those developed by DeepMind (the AI subsidiary of Google), could also be considered.

The methods developed will be applied to cancer and will provide a better understanding of the involvement of RNAs in this pathology. Cancer in a given tissue is a heterogeneous disease; several cancer subtypes can be identified. Treatments and diagnosis should be tailored to each subtype. In this project, we will be interested in lncRNAs in a frequent cancer, bladder cancer (4th cancer in terms of incidence in men) as well as in pediatric cancer, retinoblastoma. A small number of lncRNAs predicted to be potentially involved will be functionally validated by the team of biologists. We hope ultimately to be able to offer clinicians new diagnostic or prognostic markers and enable them to better understand the biological causes of the disease in order to optimize treatments.

The final objective of the project will be to implement generic methods and tools for the prediction of lncRNAs. The tools developed will be made available to the scientific community via our EvryRNA platform: http://EvryRNA.ibisc.univ-evry.

Supervision of the doctoral student

This project will be carried out in close collaboration between (i) the host team, the AROBAS team of the IBISC laboratory, which main research themes concern RNA bioinformatics and deep learning (ii) the company Adlin, an expert company in data analysis and (iii) the Molecular Oncology team at the Institut Curie, an internationally recognized team for the study of bladder cancer and made up of biologists, clinicians and pathologists interested in this cancer.

Contact : Pr. Fariza Tahi, IBISC, UEVE, Université de Paris-Saclay. **Email :** fariza.tahi@univ-evry.fr