Stage M2 Informatique ou Bioinformatique

Titre : Deep-learning pour la prédiction de structures 3D des ARN

Sujet :
Les ARN non codants sont des macro-molécules du vivant dont la fonction est liée à leur structure (résultat du repliement de la séquence de nucléotides dans l'espace). La prise de conscience lors de la dernière décennie de la grande variété de ces ARN et des rôles importants qu'ils jouent à différents niveaux de la vie de la cellule, ainsi que de leur implication dans un grand nombre de maladies telles que le cancer donne lieu à un regain d'intérêt pour leur étude structurale. Par exemple, ils sont maintenant envisagés comme de possibles cibles thérapeutiques, comme le sont déjà différentes classes de protéines. Ces ARN sont classés par famille, au sein desquelles les structures secondaires et tertiaires (structures 3D) sont assez similaires.

Notre équipe de recherche s'intéresse à la prédiction *in-silico* de la structure des ARN non codants. Ces dernières années, les méthodes d'apprentissage se sont développées en bioinformatique structurale, en particulier pour les structures protéiques [1], et nous tentons d'adapter certaines idées d'algorithmes à l'ARN.
Plutôt que de réaliser des prédictions de structure 3D à partir de la seule séquence, les méthodes les plus performantes capturent souvent le repliement global des molécules d'une même famille.
Nous avons pour ceci constitué un jeu de données de structures d'ARN en 3D, réalignées avec des alignements multiples de séquences homologues de la même famille. Il s'agit du dataset RNANet [2] disponible sur la plateforme EvryRNA (http://EvryRNA.ibisc.univ-evry.fr).

En 3D, on exprime la forme d'une chaine de nucléotides repliée dans l'espace par 3 mesures géométriques : la distance entre chaque nucléotide et le suivant, les angles plans que forment chaque triplet de nucléotides consécutifs, et les angles de torsion formés par chaque quadruplet de nucléotides consécutifs. On sait que les valeurs de ces angles forment des clusters bien identifiés.

L'objectif du stage est le développement d'un algorithme pour la prédiction des structures 3D des ARN basé sur :

- un réseau de neurones profond apprenant et prédisant les valeurs des angles de torsion à partir du contexte de séquence et de variabilité de séquence au sein de la famille d'ARN
- et un réseau de neurones profond apprenant la matrice des distances entre nucléotides au sein d'une famille d'ARN.

Bibliographie
[1] M. AlQuraishi. *End-to-End Differentiable Learning of Protein Structure*, Cell Systems, 2019 https://doi.org/10.1016/j.cels.2019.03.006
[2] L. Becquey, E. Angel et F. Tahi. *RNANet: An automatically built dual-source dataset integrating homologous sequences and RNA structures*, , Bioinformatics, 2020. https://doi.org/10.1093/bioinformatics/btaa944

Contacts
Fariza Tahi, Professeur des Universités : fariza.tahi@univ-evry.fr
Louis Becquey, doctorant : louis.becquey@univ-evry.fr

Equipe AROB@S (Algorithmique, Recherche Opérationelle, Bioinformatique et Apprentissage Statistique)
Laboratoire IBISC (Informatique, Biologie Intégrative et Systèmes Complexes)
Université Paris-Saclay, Université d'Evry.

Lieu du stage : IBGBI, 23 bv. de France, 91000 Evry.