Title project: Deep learning algorithms for the prediction, classification and analysis of long non-coding RNAs

Titre du projet : Algorithmes de deep learning pour la prédiction, la classification et l'analyse des ARN longs non-codants

Laboratory: IBISC (Informatique, Bioinformatique et Systèmes Complexes), Université d'Evry-Val d'Essonne, Université de Paris-Saclay

Project coordinator and thesis supevisor: Fariza TAHI, Professeure des universités, IBISC, UEVE.

Collaborations:
Farida Zehraoui, Maitre de Conférence, IBISC.
Nicolas Fortunel, Directeur de recherche, LGRK, CEA Evry.
Michèle Martin, Directeur de recherche, LGRK, CEA Evry.

Summary of the thesis project:

In this project we want to propose new deep learning methods for the prediction and analysis of non-coding RNAs. Non-coding RNAs (ncRNAs) are RNAs that do not code for proteins and constitute the largest part of genomes) are increasingly identified as playing important roles in deregulation processes leading to pathologies such as cancer. We will focus more particularly on long non-coding RNAs (lncRNAs), larger than 200 nucleotides, which have been identified as potential regulators. But unlike small ncRNAs, their characterization and classification by structure and function are far from established. Determining the structure of a lncRNA is a difficult problem, both by experimental (crystallography, NMR) and bioinformatic methods. In this project, we propose to develop original computational methods based on Deep Learning (DL) to predict, classify and identify the function of ncRNAs, by integrating different characteristics: sequence, structure (especially secondary), genomic position and chromosomal, interaction with coding or non-coding genes and genetic and epigenetic alterations. Two methodological challenges are to be considered: (i) making it possible to take into account heterogeneous characteristics (multi-source approach); (ii) predicting known classes of ncRNAs while being able to predict new classes, and this by combining a supervised approach with an unsupervised approach. we will apply these *in silico* approaches to the biomedical domain of radiation-induced fibroses, which are severe sequelae that can develop after treatment of tumors by radiotherapy. This will involve predicting and analyzing the ncRNAs involved in these fibroses, for a better understanding of this pathology, and ultimately for appropriate treatments.

Résumé du projet de thèse :

Dans ce projet nous souhaitons proposer de nouvelles méthodes de deep learning pour la prédiction et l'analyse de séquences génomiques particulières que sont les ARN non-codants. Les ARN non-codants ou ARNncs (ARN ne codant pas pour des protéines et constituant la grande majorité des génomes) sont de plus en plus identifiés comme jouant des rôles importants dans les processus de dérégulation entrainant des pathologies comme le cancer. Nous nous intéresserons plus particulièrement aux ARN longs non-codants (ARNlncs), de taille supérieure à 200 nucléotides, qui ont été identifiés comme de potentiels régulateurs. Mais contrairement aux petits ARNncs, leur caractérisation et classification par leur structure et leur fonction sont loin d'être établies. Déterminer la structure d'un ARNlnc est un

problème difficile, aussi bien par des méthodes expérimentales (cristallographie, RMN) que bioinformatiques. Dans ce projet, nous proposons de développer des méthodes computationnelles originales basées sur le Deep Learning (DL) pour prédire, classer et identifier la fonction des ARNlncs, en intégrant différentes caractéristiques : la séquence, la structure (notamment secondaire), la position génomique et chromosomique, l'interaction avec des gènes codants ou non-codants et les altérations génétiques et épigénétiques. Deux défis méthodologiques sont à relever : (i) permettre de prendre en compte des caractéristiques hétérogènes (approche multi-sources) ; (ii) prédire des classes connues d'ARNlncs tout en étant capable de prédire de nouvelles classes, et ce en combinant une approche supervisée avec une approche non-supervisée. Nous appliquerons ces approches *in silico* à une thématique biomédicale, les fibroses radio-induites, qui sont des séquelles sévères qui peuvent se développer après traitement de tumeurs par radiothérapie. Il s'agira de prédire et d'analyser les ARNlncs impliqués dans ces fibroses, pour une meilleure compréhension de cette pathologie, et à terme à des traitements adéquats.

Présentation du projet (English version below) **:**

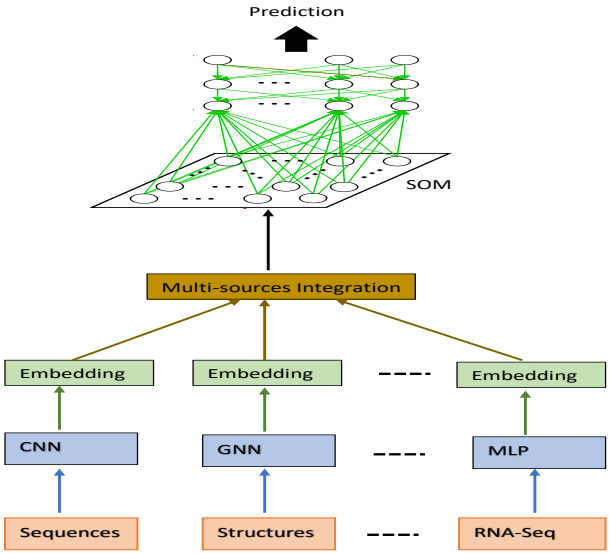
Ces dernières années, les méthodes d'apprentissage automatique, notamment de deep learning ou apprentissage profind, ont connu un essor considérable, et ont montré leur efficacité dans un grand nombre de domaines, y compris en biologie et en médecine. De plus en plus de méthodes et d'outils bioinformatiques basés sur du deep learning sont en effet proposées dans la littérature, pour répondre à diverses questions biologiques et biomédicales. Dans ce projet nous souhaitons proposer de nouvelles méthodes de deep learning pour la prédiction et l'analyse de séquences génomiques particulières que sont les ARN non-codants, et ce dans un contexte biomédical et plus précisément de médecine de précision.

Les ARN non-codants ou ARNncs (ARN ne codant pas pour des protéines et constituant la grande majorité des génomes) sont de plus en plus identifiés comme jouant des rôles importants dans les processus de dérégulation entrainant des pathologies comme le cancer. Ils sont ainsi considérés comme de potentiels marqueurs diagnostiques et cibles thérapeutiques. Leur identification et la détermination de leur fonction sont des enjeux importants, et avec les séquençages de nouvelles générations (NGS) qui génèrent des volumes considérables de données omiques, leur prédiction et leur caractérisation par des méthodes *in silico* est indispensable pour permettre d'orienter les études expérimentales ultérieures. Récemment, les ARN longs non-codants (ARNlncs), de taille supérieure à 200 nucléotides, ont été identifiés comme de potentiels régulateurs. Mais contrairement aux petits ARNncs, leur caractérisation et classification par leur structure et leur fonction sont loin d'être établies. Déterminer la structure d'un ARNlnc est un problème difficile, aussi bien par des méthodes expérimentales (cristallographie, RMN) que bioinformatiques. Déterminer sa fonction est encore moins facile, d'autant plus que contrairement aux protéines, des ARNlncs avec des fonctions similaires manquent souvent d'homologie de séquence (les séquences d'ARN présentent des mutations compensatoires maintenant une conservation au niveau structural). Des tentatives de classification des ARNlncs ont été proposées, basées sur des critères différents : longueurs des transcrits, localisation, association avec des gènes codant des protéines. Un résumé de ces classifications a été proposé dans (St.Laurent et al., 2015). Dans (Kopp and Mendell, 2018), les auteurs suggèrent une étude des ARNlncs selon leur localisation, en expliquant que celle-ci est souvent liée la fonction. Mais une grande majorité des travaux sont dédiés à l'étude d'un ARNlnc précis. Par exemple une étude récente (Uroda et al., 2019) révèle l'importance de la présence d'un pseudonœud (motif particulier de la structure secondaire) dans le mécanisme de régulation de l'ARNlnc MEG3 dans la voie biologique de p53, gène impliqué dans de nombreux cancers.

D'un point de vue computationnel, quelques méthodes ont été proposées dans la littérature pour la classification d'ARNncs bien caractérisés et dont la structure est bien connue. Ces méthodes, basées sur l'apprentissage automatique supervisé, souvent de type Deep Learning, proposent un modèle construit sur un dataset composé de 13 classes de petits ARNncs. Nous pouvons citer RNAcon (Panwar et al.,

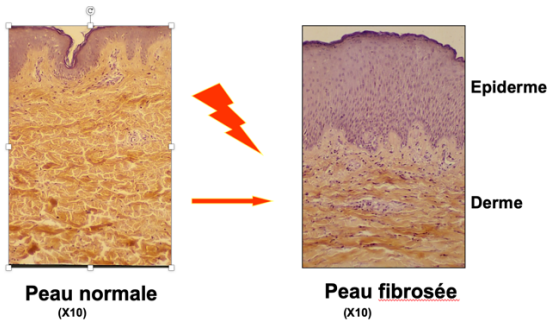
2014) basé sur le modèle des forêts aléatoires et nRC (Fiannaca et al., 2017) basé sur les réseaux de convolution (CNN), où la structure secondaire est utilisée pour la classification ; puis plus récemment ncRDeep (Chantsalnyam et al., 2020) basé sur les CNN et ncRFP (Wang et al., 2020) basé sur des réseaux de neurones récurrents (RNN), tous deux considérant uniquement des caractéristiques de séquence. De très rares méthodes s'intéressent plus spécifiquement à la classification des ARNlncs. Par exemple SEEKR (Kirk et al., 2018) se base sur la séquence, plus précisément les profils de k-mers, pour regrouper les transcrits qui se ressemblent le plus et forment une classe fonctionnelle, en utilisant un algorithme de clustering basé sur une corrélation de Pearson (apprentissage non-supervisée). LncADeep (Yang et al., 2018) utilise un réseau de neurones profond (DNN) pour identifier des interactions entre des ARNlncs et des protéines, en se basant sur la séquence et la structure secondaire. L'outil utilise ensuite l'annotation des protéines associées à un ARNlnc pour décrire les fonctions biologiques dans lesquelles il est potentiellement impliqué. Bien que ces méthodes permettent de préciser la grande catégorie des ARNlncs, elles restent limitées. De plus, les classes résumées dans (St.Laurent et al., 2015) n'y sont pas toutes identifiées. Nous pensons qu'il pourrait être possible de classifier plus finement les ARNlncs en prenant en compte d'autres caractéristiques.

Dans ce projet, nous proposons de développer des méthodes computationnelles originales basées sur le Deep Learning (DL) pour prédire, classer et identifier la fonction des ARNlncs, en intégrant différentes caractéristiques : la séquence, la structure (notamment secondaire), la position génomique et chromosomique, l'interaction avec des gènes codants ou non-codants et les altérations génétiques et épigénétiques. Deux défis méthodologiques sont à relever : (i) permettre de prendre en compte des caractéristiques hétérogènes (approche multi-sources) ; (ii) prédire des classes connues d'ARNlncs tout en étant capable de prédire de nouvelles classes, et ce en combinant une approche supervisée avec une approche non-supervisée. Un point que nous jugeons également important concerne la partie visualisation des résultats, pour une meilleure compréhension et interprétation par l'utilisateur. Les cartes auto-organisatrices (« self organizing maps » ou SOM) sont des réseaux de neurones non supervisés capables de regrouper et de visualiser des données de grandes dimensions. En utilisant un algorithme d'apprentissage compétitif non supervisé, cette technique est capable de produire une carte, représentant l'espace d'entrée, dans laquelle les données proches sont localisées dans des régions proches de la carte. Afin de représenter les sources hétérogènes, nous proposerons des approches multimodales originales basées sur le DL qui permettraient de fusionner les différentes sources de données. La fusion peut être effectuée en utilisant trois stratégies principales (Ramachandram and Taylor, 2017) : fusion précoce, fusion conjointe et fusion tardive. La fusion précoce consiste à combiner les caractéristiques d'entrée des différentes sources avant l'utilisation d'un seul modèle DL. La fusion conjointe fait référence au processus de combinaison des représentations des entrées apprises au niveau des couches intermédiaires des différents réseaux de neurones qui représentent les modalités. La fusion tardive permet de combiner les décisions de plusieurs réseaux de neurones qui traitent les modalités pour fournir une décision finale. Nous nous intéresserons particulièrement à la fusion conjointe pour la classification des ARNlncs et l'identification de leurs fonctions biologiques. Afin de prendre en compte les différentes sources hétérogènes, chaque source de données sera traitée par un modèle DL adéquat, comme les CNNs, les « Graph Neural Networks » (GNNs) et les perceptrons multi-couches (MLPs), ce qui permettra de mieux extraire les caractéristiques de haut niveau de cette source. Pour permettre la découverte de nouvelles classes, nous allons étudier l'association de différentes options de rejet (Geifman and El-Yaniv, 2019) au modèle multimodal. La combinaison de ce modèle avec les SOMs (Platon et al. 2018) permettra la visualisation des nouvelles classes d'ARNncs longs. Nous allons également nous intéresser à l'identification des sources de données et des caractéristiques qui ont mené aux prédictions (Platon et al. 2018bis). Ceci permettra d'expliquer les prédictions et de découvrir de nouvelles propriétés qui pourront être associées aux ARNncs longs.



Le but à terme de notre projet est de mettre en œuvre une méthodologie générique permettant, étant donnée une problématique biologique donnée, et plus particulièrement une pathologie, d'identifier les ARNlncs impliqués, prédire leur structure, leurs interactions avec d'autres ARN ou avec des protéines, pour enfin déterminer le rôle qu'ils jouent dans le processus étudié. Il s'agira donc d'utiliser les algorithmes développés dans ce projet mais aussi des outils de bioinformatique des ARN développés (ou en cours de développement) dans l'équipe AROBAS (et disponibles sur la plateforme EvryRNA : http://EvryRNA.ibisc.univ-evry, tels que RNANet (Becquey et al., 2020), Biorseo (Becquey et al., 2020), RCPred (Legendre et al, 2019), ou IRSOM (Platon et al., 2018)), et également proposés dans la littérature.

Dans le cadre du projet de thèse, nous appliquerons ces approches *in silico* à une thématique biomédicale, les fibroses radio-induites, qui sont des séquelles sévères qui peuvent se développer après traitement de tumeurs par radiothérapie. Le rôle des ARNlncs dans le développement des fibroses a fait l'objet de publications récentes, comme par exemple l'article concernant H19X (Pachera JCI 2020), mais aucune étude n'a été dédiée aux fibroses après radiothérapie. Cette application sera développée *via* une collaboration avec le laboratoire CEA de 'génomique et radiobiologie de la kératinopoïèse' (LGRK, Evry).



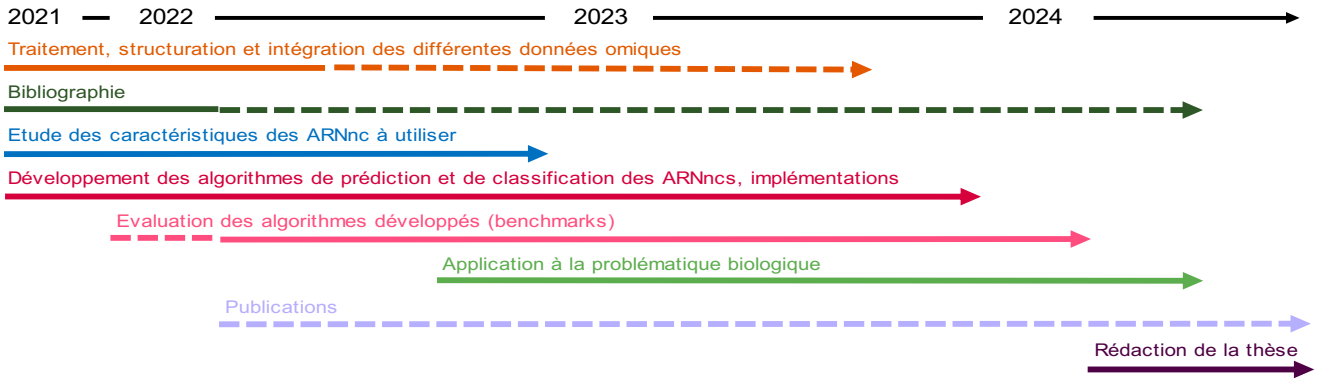
Le traitement de tumeurs profondes par radiothérapies occasionne une exposition du tissu sain environnant à des fortes doses de rayonnements, qui peuvent induire des réactions aigües et des pathologies chroniques sévères telles que des fibroses. Or, les patients présentent une susceptibilité variable à ces complications, 5 à 10% étant hyper-sensibles, ce qui pose la problématique de la personnalisation des traitements de radiothérapie. Le LGRK s'est doté d'une échantillothèque de cellules de peau issues de patients ayant présenté une hyper-radiosensibilité lors d'une radiothérapie, qui été caractérisée au niveau exome et transcriptome par séquençage. Avec pour socle ce matériel, les

recherches développées visent à acquérir une compréhension des réseaux moléculaires effecteurs de la susceptibilité à la fibrose, intégrant le génome et ses variants, ainsi que l'épigénome. Les ARNlncs sont particulièrement ciblés (Martin, 2019). Ces recherches ont pour but d'identifier de nouveaux effecteurs de la physiopathologie de la fibrose et des biomarqueurs originaux qui permettraient un pronostic de l'hyper-radiosensibilité. L'analyse cherchera notamment à positionner les ARNlncs issus de l'approche prédictive *in silico* au sein des réseaux fonctionnels déjà identifiés et validés par le LGRK pour leur implication dans la régulation du processus de fibrose ((Martin et al., 2000) et données originales récentes non publiées). Les candidats les plus pertinents constitueront un vivier pour l'initiation d'une stratégie de validation expérimentale.

Dans ce projet de thèse il s'agira donc, en plus du développement d'algorithmiques originaux de deep learning et de méthodes bioinformatiques dédiées aux ARN, tout aussi originales, d'aider, grâce aux méthodes que nous développerons, à l'analyse d'une problématique de santé, pour une meilleure réponse thérapeutique.

Les différents algorithmes et méthodes informatiques et bioinformatiques que développera le doctorant seront présentés à différents séminaires et groupes de travail, ainsi qu'a des conférences nationales et internationales de bioinformatique telles ISMB, ECCB, JOBIM, ou d'informatique telles que ICONIP, ICML, WSOM donnant lieu à des publications d'actes. Ils intéresseront également les congrès de radiothérapie, en tant que nouvelle approche de radiothérapie personnalisée (ESTRO). Ils seront soumis à publication dans des revues scientifiques internationales d'informatique ou de bioinformatique telles que Bioinformatics, BMC Bioinformatics, Plos One, Plos Computational Biology. L'application à la collection de patients de radiothérapie pourra être présentée à des journaux comme I J Radiat Oncol Biol Phys. Nous prévoyons une à deux publications par an. Par ailleurs, les algorithmes et outils dédiés aux ARN qui seront développés durant le projet seront mis à disposition de la communauté scientifique sous forme de web server via EvryRNA (http://EvryRNA.ibisc.univ-evry), plateforme logicielle de bioinformatique d'IBISC labélisée par Genopole et faisant partie des plateformes de l'Université de Paris-Saclay (Plug in Labs).

La thèse se déroulera selon le planning suivant. Des réunions bimensuelles seront organisées regroupant les partenaires, et un rapport semestriel sera rédigé par l'étudiant.



Références :

-Becquey L, Angel E, Tahi F. RNANet: an automatically built dual-source dataset integrating homologous sequences and RNA structures. Bioinformatics. 2020 Nov 2:btaa944. doi: 10.1093/bioinformatics/btaa944.

-Becquey L, Angel E, Tahi F. BiORSEO: a bi-objective method to predict RNA secondary structures with pseudoknots using RNA 3D modules. Bioinformatics. 2020 Apr 15;36(8):2451-2457. doi: 10.1093/bioinformatics/btz962. PMID: 31913439.

-Brademan DR, Miller IJ, Kwiecien NW, Pagliarini DJ, Westphall MS, Coon JJ, Shishkova E. Argonaut: A Web Platform for

Collaborative Multi-omic Data Visualization and Exploration. Patterns (N Y). 2020 Oct 9;1(7):100122. doi: 10.1016/j.patter.2020.100122..

-Chantsalnyam, T., Lim, D. Y., Tayara, H., & Chong, K. T. (2020). ncRDeep: Non-coding RNA classification with convolutional neural network. In *Computational Biology and Chemistry* (Vol. 88). Elsevier Ltd. https://doi.org/10.1016/j.compbiolchem.2020.107364

-Fiannaca, A., La Rosa, M., La Paglia, L., Rizzo, R., Urso, A. (2017). NRC: Non-coding RNA Classifier based on structural features. *BioData Mining*, *10*(1), 27. https://doi.org/10.1186/s13040-017-0148-2

-Geifman, Y. & El-Yaniv, R.. (2019). SelectiveNet: A Deep Neural Network with an Integrated Reject Option. *Proceedings of the 36th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 97:2151-2159.

-Kirk, J. M., Kim, S. O., Inoue, K., Smola, M. J., Lee, D. M., Schertzer, M. D., Wooten, J. S., Baker, A. R., Sprague, D., Collins, D. W., Horning, C. R., Wang, S., Chen, Q., Weeks, K. M., Mucha, P. J., Calabrese, J. M. (2018). Functional classification of long non-coding RNAs by k-mer content. *Nature Genetics*, *50*(10), 1474–1482. https://doi.org/10.1038/s41588-018-0207-8

-Kopp, F., & Mendell, J. T. (2018). Functional Classification and Experimental Dissection of Long Noncoding RNAs. In Cell (Vol. 172, Issue 3, pp. 393–407). https://doi.org/10.1016/j.cell.2018.01.011

-Legendre A, Angel E, Tahi F. RCPred: RNA complex prediction as a constrained maximum weight clique problem. BMC Bioinformatics. 2019 Mar 29;20(Suppl 3):128. doi: 10.1186/s12859-019-2648-1.

-Martin MT. Long non-coding RNAs: new mechanisms regulating sensitivity to ionizing radiation. European Commission Proceedings, 2019, Epigenetic effects, potential impact on radiation protection, Radiation Protection n°189, November, 32-39. ISSN 1681-6803.

-Martin, M., Lefaix, J.-L., and Delanian, S. TGF-β1 and radiation fibrosis: a master switch and a specific therapeutic target. Int J Radiat Oncol Biol Phys. 2000 May 1;47(2):277-90. doi: 10.1016/s0360-3016(00)00435-1. PMID: 1080235

-McGowan T, Johnson JE, Kumar P, Sajulga R, Mehta S, Jagtap PD, Griffin TJ. Multi-omics Visualization Platform: An extensible Galaxy plug-in for multi-omics data visualization and exploration. Gigascience. 2020 Apr 1;9(4):giaa025. doi: 10.1093/gigascience/giaa025.

-Netanely D, Stern N, Laufer I, Shamir R. PROMO: an interactive tool for analyzing clinically-labeled multi-omic cancer datasets. BMC Bioinformatics. 2019 Dec 26;20(1):732. doi: 10.1186/s12859-019-3142-5. PMID: 31878868; PMCID: PMC6933892.

- Pachera E, Assassi S, Salazar GA, Stellato M, Renoux F, Wunderlin A, Blyszczuk P, Lafyatis R, Kurreeman F, de Vries-Bouwstra J, Messemaker T, Feghali-Bostwick CA, Rogler G, van Haaften WT, Dijkstra G, Oakley F, Calcagni M, Schniering J, Maurer B, Distler JH, Kania G, Frank-Bertoncelj M, Distler O. Long noncoding RNA H19X is a key mediator of TGF-β-driven fibrosis. J Clin Invest. 2020 Sep 1;130(9):4888-4905. doi: 10.1172/JCI135439. PMID: 32603313

-Panwar, B., Arora, A., & Raghava, G. P. S. (2014). Prediction and classification of ncRNAs using structural information. *BMC Genomics*, *15*(1), 127. https://doi.org/10.1186/1471-2164-15-127

-Platon L, Zehraoui F, Bendahmane A, Tahi F. IRSOM, a reliable identifier of ncRNAs based on supervised self-organizing maps with rejection. Bioinformatics. 2018 Sep 1;34(17):i620-i628. doi: 10.1093/bioinformatics/bty572.

-Platon L, Zehraoui F, Tahi F. Localized Multiple Sources Self-Organizing Map. ICONIP (3) 2018: 648-659

-Ramachandram D. and Taylor, G. W. "Deep Multimodal Learning: A Survey on Recent Advances and Trends," in *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96-108, Nov. 2017, doi: 10.1109/MSP.2017.2738401.

-St.Laurent, G., Wahlestedt, C., & Kapranov, P. (2015). The Landscape of long noncoding RNA classification. In *Trends in Genetics* (Vol. 31, Issue 5, pp. 239–251). Elsevier Ltd. https://doi.org/10.1016/j.tig.2015.03.007

-Uroda, T., Anastasakou, E., Rossi, A., Teulon, J. M., Pellequer, J. L., Annibale, P., Pessey, O., Inga, A., Chillón, I., Marcia, M. (2019). Conserved Pseudoknots in lncRNA MEG3 Are Essential for Stimulation of the p53 Pathway. *Molecular Cell*, 75(5), 982-995.e9. doi.org/10.1016/j.molcel.2019.07.025

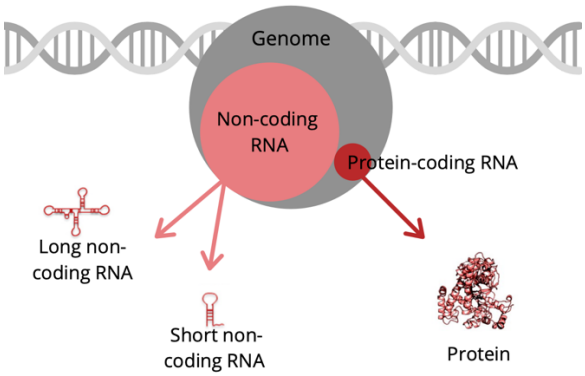
-Wang, L., Zheng, S., Zhang, H., Qiu, Z., Zhong, X., Liu, H., Liu, Y. (2020). ncRFP: A novel end-to-end method for non-coding RNAs family prediction based on Deep Learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1–1. https://doi.org/10.1109/tcbb.2020.2982873

-Yang, C., Yang, L., Zhou, M., Xie, H., Zhang, C., Wang, M. D., Zhu, H. (2018). LncADeep: An ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics*, *34*(22), 3825–3834. https://doi.org/10.1093/bioinformatics/bty428.

-Yu SH, Ferretti D, Schessner JP, Rudolph JD, Borner GHH, Cox J. Expanding the Perseus Software for Omics Data Analysis With Custom Plugins. Curr Protoc Bioinformatics. 2020 Sep;71(1):e105. doi: 10.1002/cpbi.105. PMID: 32931150.

Presentation of the project:

In recent years, machine learning methods, particularly deep learning, have grown considerably, and have shown their effectiveness in a large number of fields, including biology and medicine. More and more bioinformatic methods and tools based on deep learning are indeed proposed in the literature, to answer various biological and biomedical questions. In this project, we want to propose new deep learning methods for the prediction and analysis of a specific genomic sequences that are non-coding RNAs, in a biomedical context, and more precisely in precision medicine.

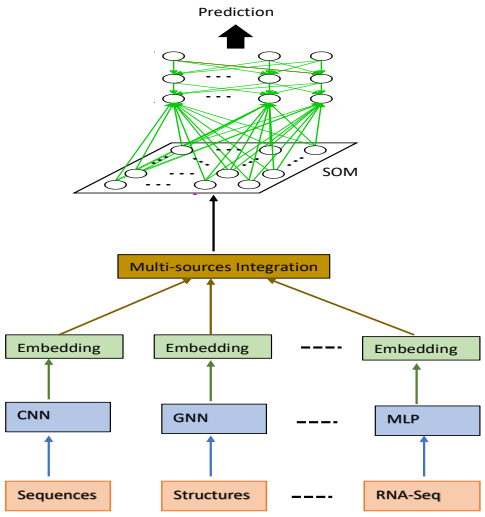


Non-coding RNAs or ncRNAs (RNAs which do not code for proteins and constitute the largest part of genomes) are increasingly identified as playing important roles in the deregulation processes leading to pathologies such as cancer. They are thus considered as potential diagnostic markers and therapeutic targets. Their identification and the determination of their function are important issues, and with the next generation sequencing (NGS) which generate considerable volumes of omics data, their prediction and their characterization by in silico methods is essential to make it possible to orient the experimental studies. Recently, long non-coding RNAs (lncRNAs), larger than 200 nucleotides, have been identified as potential regulators. But unlike small ncRNAs, their characterization and classification by structure and function are far from established. Determining the structure of a lncRNA is a difficult problem, both by experimental (crystallography, NMR) and bioinformatic methods. Determining its function is even more difficult, especially since unlike proteins, ncRNAs with similar functions often lack sequence homology (RNA sequences show compensatory mutations maintaining structural conservation). Attempts to classify ncRNAs have been proposed, based on different criteria: length of transcripts, location, association with genes encoding proteins. A summary of these classifications has been proposed in (St. Laurent et al., 2015). In (Kopp and Mendell, 2018), the authors suggest a study of ncRNAs according to their location, explaining that this is often linked to function. But a large majority of the work is dedicated to the study of a precise lncRNA. For example, a recent study (Uroda et al., 2019) reveals the importance of the presence of a pseudoknot (particular pattern of the secondary structure) in the mechanism of regulation of the MEG3 lncRNA in the biological pathway of p53, gene involved in many cancers.

From a computational point of view, a few methods have been proposed in the literature for the classification of well characterized ncRNAs whose structure is well known. These methods, based on supervised machine learning, often of the Deep Learning type, offer a model built on a dataset composed of 13 classes of small ncRNAs. We can cite RNAcon (Panwar et al., 2014) based on the model of random forests and nRC (Fiannaca et al., 2017) based on convolutional networks (CNN), where the secondary structure is used for classification; then more recently ncRDeep (Chantsalnyam et al., 2020) based on CNNs and ncRFP (Wang et al., 2020) based on recurrent neural networks (RNN), both considering only sequence characteristics. Very few methods are specifically interested in the classification of lncRNAs. For example, SEEKR (Kirk et al., 2018) uses the sequence, more precisely the profiles of k-mers, to group the transcripts which are most similar and form a functional class, using a clustering algorithm

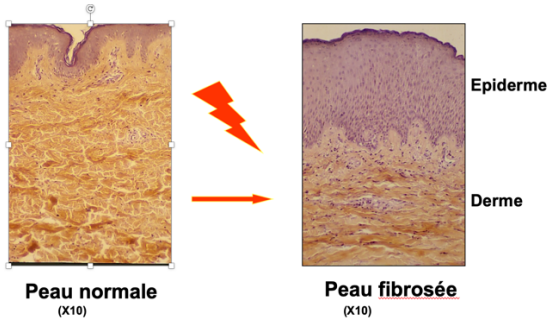
based on a Pearson correlation (unsupervised learning). LncADeep (Yang et al., 2018) uses a deep neural network (DNN) to identify interactions between lncRNAs and proteins, based on sequence and secondary structure. The tool then uses the annotation of proteins associated with a lncRNA to describe the biological functions in which it is potentially involved. Although these methods make it possible to specify the broad category of lncRNAs, they remain limited. In addition, the classes summarized in (St. Laurent et al., 2015) are not all identified by the existing tools. We believe that it might be possible to more finely classify lncRNAs by taking into account other characteristics.

In this project, we propose to develop original computational methods based on Deep Learning (DL) to predict, classify and identify the function of lncRNAs, by integrating different characteristics: sequence, structure (especially secondary), genomic and chromosomal position, interaction with coding or non-coding genes, and genetic and epigenetic alterations. Two methodological challenges are to be considered: (i) making it possible to take into account heterogeneous characteristics (multi-source approach); (ii) predicting known classes of lncRNAs while being able to predict new classes, and this by combining a supervised approach with an unsupervised approach. An important point that we also consider concerns the visualization part of the results, for a better understanding and interpretation by the user. Self-organizing maps (SOM) are unsupervised neural network capable of grouping and visualizing large-scale data. Using an unsupervised competitive learning algorithm, this technique is able to produce a map, representing the input space, in which nearby data is located in regions close to the map. In order to represent heterogeneous sources, we will propose original multimodal approaches based on DL which would allow to merge the different data sources. Fusion can be performed using three main strategies (Ramachandram and Taylor, 2017): early fusion, joint fusion and late fusion. Early merging involves combining the input characteristics of different sources before using a single DL model. Joint fusion refers to the process of combining representations of inputs learned at the intermediate layers of different neural networks that represent modalities. Late fusion allows the decisions of several neural networks that process modalities to be combined to provide a final decision. We will be particularly interested in joint fusion for the classification of lncRNAs and the identification of their biological functions. In order to take into account the different heterogeneous sources, each data source will be processed by an adequate DL model, such as CNNs, “Graph Neural Networks” (GNNs) and multi-layer perceptrons (MLPs), which will allow better extraction of high level features from this source. To allow the discovery of new classes, we will study the association of different rejection options (Geifman and El-Yaniv, 2019) to the multimodal model. The combination of this model with SOMs (Platon et al. 2018) will allow the visualization of new classes of lncRNAs. We will also be interested in identifying the data sources and the characteristics that led to the predictions (Platon et al. 2018bis). This will make it possible to explain the predictions and to discover new properties that could be associated with lncRNAs.



The ultimate goal of our project is to implement a generic methodology allowing, given a biological problem, and more particularly a pathology, to identify the lncRNAs involved, predict their structure, their interactions with other RNAs or with proteins, to finally determine the role they play in the process studied. It will therefore be a question of using the algorithms developed in this project but also RNA bioinformatics tools developed (or under development) in the AROBAS team (and available on the EvryRNA platform (http://EvryRNA.ibisc.univ-evry.fr), such as RNANet (Becquey et al., 2020), Biorseo (Becquey et al., 2020), RCPred (Legendre et al, 2019), or IRSOM (Platon et al., 2018)), and also proposed in the literature.

As part of the thesis project, we will apply these *in silico* approaches to the biomedical domain of radiation-induced fibroses, which are severe sequelae that can develop after treatment of tumors by radiotherapy. The role of lncRNAs in the development of fibroses has been the subject of recent publications, such as the article on H19X (Pachera JCI 2020), but no study has been dedicated to fibrosis after radiotherapy. This application will be developed through a collaboration with the CEA laboratory of "genomics and radiobiology of keratinopoiesis" (LGRK, Evry).



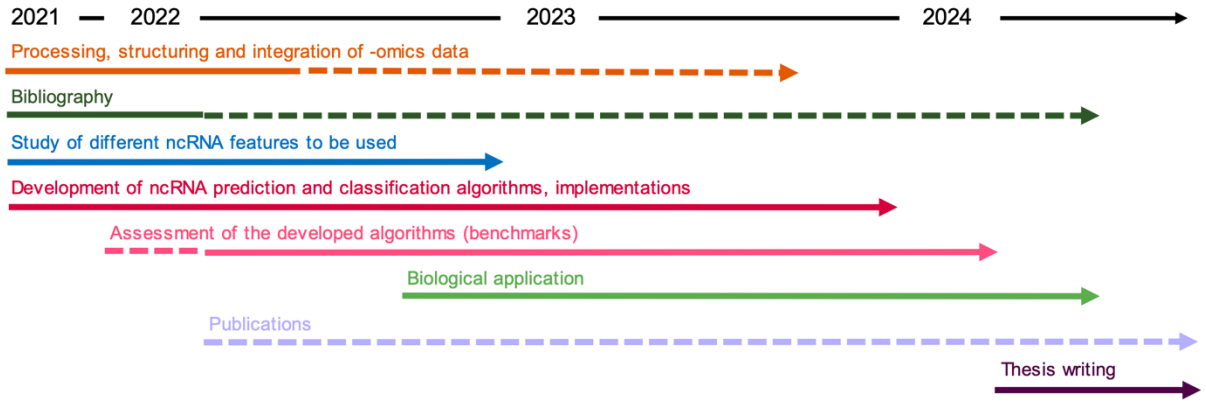
The treatment of deep tumors by radiotherapy causes exposure of the surrounding healthy tissues to high doses of radiation, which can induce acute reactions and severe chronic pathologies such as fibroses. However, patients have a variable susceptibility to these complications, 5 to 10% being hyper-sensitive, which poses the problem of radiotherapy treatment personalization. The LGRK has generated a sample library of skin cells from patients who have presented hyper-radiosensitivity during radiotherapy, which has been characterized at the exome and transcriptome level by sequencing. Based on this material, the research developed aims at understanding the molecular effector networks that control susceptibility to fibrosis, integrating the genome and its variants, as well as the epigenome. LncRNAs are particularly targeted (Martin, 2019). This research aims to identify new effectors of the pathophysiology of fibrosis and original biomarkers that would allow a prognosis of hyper-radiosensitivity. The analysis in particular position the lncRNAs resulting from the *in silico* predictive approach within the functional networks already identified and validated by the LGRK for their involvement in the regulation of the fibrosis process (Martin et al., 2000 and recent original unpublished data). The most relevant candidates will constitute a background for the initiation of an experimental validation strategy.

In this thesis project the aim will therefore be, in addition to the development of original deep learning algorithms and original bioinformatics methods dedicated to RNAs, to help, thanks to the methods that we will develop, in the analysis and understanding of a health issue, for a better therapeutic response.

The various computer and bioinformatics algorithms and methods that the doctoral student will develop will be presented at various seminars and working groups, as well as at national and international conferences in bioinformatics such as ISMB, ECCB, JOBIM, or in computer science such as ICONIP, ICML, WSOM, giving rise to publications of acts. They will also be of interest to radiotherapy congresses, as a new approach to personalized radiotherapy (ESTRO). They will be submitted for publication in international scientific journals of computer science or bioinformatics such as Bioinformatics, BMC Bioinformatics, Plos One, Plos Computational Biology. The application to the radiotherapy patient collection may be featured in journals such as I J Radiat Oncol Biol Phys.

We expect one to two publications per year. In addition, the algorithms and tools dedicated to RNAs which will be developed during the project will be made available to the scientific community via EvryRNA (http://EvryRNA.ibisc.univ-evry.fr), a bioinformatics software platform of IBISC labeled by Genopole and part of the platforms of the University of Paris-Saclay (Plug in Labs).

The thesis will take place according to the following schedule. Bi-monthly meetings will be organized bringing together the partners, and a biannual report will be written by the student.



References:

-Becquey L, Angel E, Tahi F. RNANet: an automatically built dual-source dataset integrating homologous sequences and RNA structures. Bioinformatics. 2020 Nov 2:btaa944. doi: 10.1093/bioinformatics/btaa944.

-Becquey L, Angel E, Tahi F. BiORSEO: a bi-objective method to predict RNA secondary structures with pseudoknots using RNA 3D modules. Bioinformatics. 2020 Apr 15;36(8):2451-2457. doi: 10.1093/bioinformatics/btz962. PMID: 31913439.

-Brademan DR, Miller IJ, Kwiecien NW, Pagliarini DJ, Westphall MS, Coon JJ, Shishkova E. Argonaut: A Web Platform for Collaborative Multi-omic Data Visualization and Exploration. Patterns (N Y). 2020 Oct 9;1(7):100122. doi: 10.1016/j.patter.2020.100122..

-Chantsalnyam, T., Lim, D. Y., Tayara, H., & Chong, K. T. (2020). ncRDeep: Non-coding RNA classification with convolutional neural network. In *Computational Biology and Chemistry* (Vol. 88). Elsevier Ltd. https://doi.org/10.1016/j.compbiolchem.2020.107364

-Fiannaca, A., La Rosa, M., La Paglia, L., Rizzo, R., Urso, A. (2017). NRC: Non-coding RNA Classifier based on structural features. *BioData Mining*, *10*(1), 27. https://doi.org/10.1186/s13040-017-0148-2

-Geifman, Y. & El-Yaniv, R.. (2019). SelectiveNet: A Deep Neural Network with an Integrated Reject Option. *Proceedings of the 36th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 97:2151-2159.

-Kirk, J. M., Kim, S. O., Inoue, K., Smola, M. J., Lee, D. M., Schertzer, M. D., Wooten, J. S., Baker, A. R., Sprague, D., Collins, D. W., Horning, C. R., Wang, S., Chen, Q., Weeks, K. M., Mucha, P. J., Calabrese, J. M. (2018). Functional classification of long non-coding RNAs by k-mer content. *Nature Genetics*, *50*(10), 1474–1482. https://doi.org/10.1038/s41588-018-0207-8

-Kopp, F., & Mendell, J. T. (2018). Functional Classification and Experimental Dissection of Long Noncoding RNAs. In Cell (Vol. 172, Issue 3, pp. 393–407). https://doi.org/10.1016/j.cell.2018.01.011

-Legendre A, Angel E, Tahi F. RCPred: RNA complex prediction as a constrained maximum weight clique problem. BMC Bioinformatics. 2019 Mar 29;20(Suppl 3):128. doi: 10.1186/s12859-019-2648-1.

-Martin MT. Long non-coding RNAs: new mechanisms regulating sensitivity to ionizing radiation. European Commission Proceedings, 2019, Epigenetic effects, potential impact on radiation protection, Radiation Protection n°189, November, 32-39. ISSN 1681-6803.

-Martin, M., Lefaix, J.-L., and Delanian, S. TGF-β1 and radiation fibrosis: a master switch and a specific therapeutic target. Int J Radiat Oncol Biol Phys. 2000 May 1;47(2):277-90. doi: 10.1016/s0360-3016(00)00435-1. PMID: 1080235

-McGowan T, Johnson JE, Kumar P, Sajulga R, Mehta S, Jagtap PD, Griffin TJ. Multi-omics Visualization Platform: An extensible Galaxy plug-in for multi-omics data visualization and exploration. Gigascience. 2020 Apr 1;9(4):giaa025. doi: 10.1093/gigascience/giaa025.

-Netanely D, Stern N, Laufer I, Shamir R. PROMO: an interactive tool for analyzing clinically-labeled multi-omic cancer datasets. BMC Bioinformatics. 2019 Dec 26;20(1):732. doi: 10.1186/s12859-019-3142-5. PMID: 31878868; PMCID: PMC6933892.

- Pachera E, Assassi S, Salazar GA, Stellato M, Renoux F, Wunderlin A, Blyszczuk P, Lafyatis R, Kurreeman F, de Vries-Bouwstra J, Messemaker T, Feghali-Bostwick CA, Rogler G, van Haaften WT, Dijkstra G, Oakley F, Calcagni M, Schniering J, Maurer B, Distler JH, Kania G, Frank-Bertoncelj M, Distler O. Long noncoding RNA H19X is a key mediator of TGF-β-driven fibrosis. J Clin Invest. 2020 Sep 1;130(9):4888-4905. doi: 10.1172/JCI135439. PMID: 32603313

-Panwar, B., Arora, A., & Raghava, G. P. S. (2014). Prediction and classification of ncRNAs using structural information. *BMC Genomics*, *15*(1), 127. https://doi.org/10.1186/1471-2164-15-127

-Platon L, Zehraoui F, Bendahmane A, Tahi F. IRSOM, a reliable identifier of ncRNAs based on supervised self-organizing maps with rejection. Bioinformatics. 2018 Sep 1;34(17):i620-i628. doi: 10.1093/bioinformatics/bty572.

-Platon L, Zehraoui F, Tahi F. Localized Multiple Sources Self-Organizing Map. ICONIP (3) 2018: 648-659

-Ramachandram D. and Taylor, G. W. "Deep Multimodal Learning: A Survey on Recent Advances and Trends," in *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96-108, Nov. 2017, doi: 10.1109/MSP.2017.2738401.

-St.Laurent, G., Wahlestedt, C., & Kapranov, P. (2015). The Landscape of long noncoding RNA classification. In *Trends in Genetics* (Vol. 31, Issue 5, pp. 239–251). Elsevier Ltd. https://doi.org/10.1016/j.tig.2015.03.007

-Uroda, T., Anastasakou, E., Rossi, A., Teulon, J. M., Pellequer, J. L., Annibale, P., Pessey, O., Inga, A., Chillón, I., Marcia, M. (2019). Conserved Pseudoknots in lncRNA MEG3 Are Essential for Stimulation of the p53 Pathway. *Molecular Cell*, 75(5), 982-995.e9. doi.org/10.1016/j.molcel.2019.07.025

-Wang, L., Zheng, S., Zhang, H., Qiu, Z., Zhong, X., Liu, H., Liu, Y. (2020). ncRFP: A novel end-to-end method for non-coding RNAs family prediction based on Deep Learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1–1. https://doi.org/10.1109/tcbb.2020.2982873

-Yang, C., Yang, L., Zhou, M., Xie, H., Zhang, C., Wang, M. D., Zhu, H. (2018). LncADeep: An ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics*, *34*(22), 3825–3834. https://doi.org/10.1093/bioinformatics/bty428.

-Yu SH, Ferretti D, Schessner JP, Rudolph JD, Borner GHH, Cox J. Expanding the Perseus Software for Omics Data Analysis With Custom Plugins. Curr Protoc Bioinformatics. 2020 Sep;71(1):e105. doi: 10.1002/cpbi.105. PMID: 32931150.