université PARIS-SACLAY

FACULTÉ DES SCIENCES D'ORSAY

UQÀM

Université du Québec à Montréal

Mémoire de Stage

M1 Bio-informatique et biostatistique

Rémi Poulard

Mise en place de modèles de prédiction des nucléotides modifiés sur des graphes de structure d'ARN

Sous la direction de :

Vladimir Reinharz

Professeur et chercheur au département d'informatique de l'UQAM

Responsable de formation :

M. Lespinet
Mme Cohen-Boulakia

Année 2022/2023

Table des matières

Introduction

Le stage a eu lieu au sein de l'UQAM (Université du Québec à Montréal) dans le département informatique au sein de l'équipe de Vladimir Reinharz.

L'objectif du stage était de mettre en place des modèles de prédiction par l'utilisation d'algorithmes de machine learning sur des graphes d'ARN afin de prédire la présence de nucléotides modifiés. Ce travail est un préambule avec comme objectif de déterminer si prédire des caractéristiques structurelles de l'ARN en utilisant des graphes est une pratique envisageable. Après avoir réalisé un modèle avec des performances jugée suffisante, une seconde partie du stage s'est porté sur son étude dans le but d'énoncer des hypothèses sur le lien entre les nucléotides modifiés et leur contexte structurel.

Représentation de l'ARN sous forme de graphe

Les molécules d'ARN peuvent être représentées de différentes façons, par exemple de manière linéaire en se concentrant sur la séquence, ou encore comme une structure 3D en attribuant à chaque atome des coordonnées dans l'espace. Il est cependant aussi possible de représenter les structures d'ARN comme des graphes dans lesquels chaque nœud correspond à un nucléotide et des arêtes sont mises entre les nœuds lorsque les nucléotides correspondants sont liés dans la structure d'ARN.

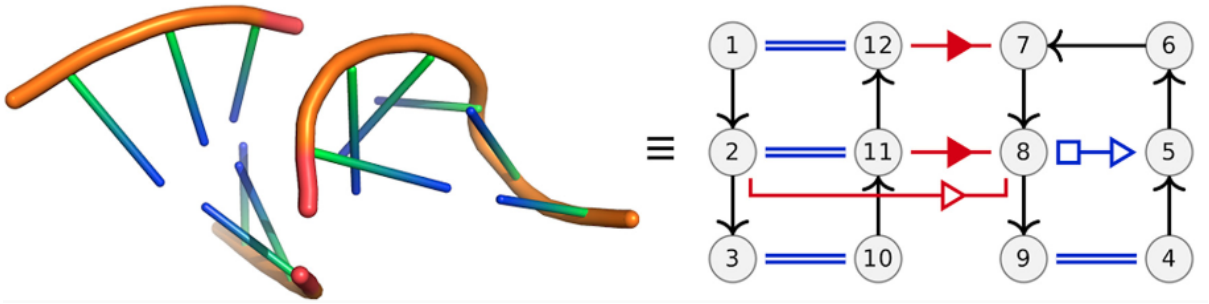


Figure 1 : Exemple de représentation 3D (à gauche) et en graphe (à droite) de la même structure d'ARN (Illustration provenant de Reinharz V et al., (2018) [1])

Cette représentation permet d'appliquer des outils provenant de la théorie des graphes pour étudier les structures d'ARN et répondre à certaines questions (par exemple la recherche de motif récurrent au sein de l'ARN ou l'étude des éléments de structures secondaires). Des bases de données de graphes existent déjà comme *RNA Structure Atlas* [2] et s'agrandissent au fur et à mesure que de nouvelles structures d'ARN sont ajoutées dans la PDB, bien que cette représentation ne soit pas aussi exploitée dans la communauté scientifique que la représentation 3D.

Subtilités biologiques des structures d'ARN

In vivo, l'ARN est replié sur lui-même et adopte une structure 3D qui possède un rôle essentiel à sa fonction. Ce repliement est permis, car les nucléotides d'une même séquence ont la capacité de former des liaisons entre eux. Il existe de nombreux types de liaisons, mais les plus étudiées sont les liaisons de type "base pair" formées par des interactions hydrogènes entre deux nucléotides. Ces liaisons sont elles-mêmes réparties en liaison canonique (liaison de Watson-Crick) et non canonique et il existe un total de 12 types de liaison "base pair" dans la nomenclature de lontis-westhof identifiées par Leontis NB et al., (2001) [3]. Les liaisons canoniques, majoritaires, ajoutent un second niveau de complexité structurelle par rapport à la séquence en elle-même en créant des hélices et des empilements. Les liaisons non canoniques sont essentielles dans la formation d'éléments de structures secondaires (par exemples des boucles ou des têtes d'épingles) ce qui rajoute un troisième niveau de complexité structurel permettant aux ARN d'adopter des structures 3D complexes et variées. Ces liaisons, minoritaires, possèdent une faible énergie de liaison, ce qui les rend durs à appréhender lorsque l'ARN est représenté comme une structure 3D. Ceci n'est pas le cas d'une représentation en graphe, où cette subtilité est illustrée par l'étiquetage des arêtes reliant les nœuds en fonction du type de liaison qui sépare les nucléotides.

Une autre subtilité des structures d'ARN réside dans la présence de nucléotides modifiés dus à des mécanismes post-transcriptionnels. Ces modifications peuvent avoir une influence sur les propriétés physico-chimiques et la taille des nucléotides. Ces modifications sont peu connues et pour le moment 340 modifications différentes ont été identifiées [4].
De plus, il semblerait que la fonction de certaines structures soit dépendantes de la présence de ces nucléotides modifiés.

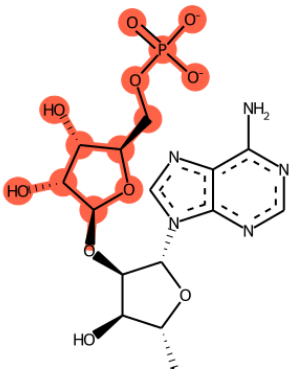


Figure 2 : Représentation de Cram de la molécule pAr(p) provenant de *Modomics* [4]. Exemple de modification chimique de l'Adénosine (la partie modifiée étant en rouge)

Utilisation de l'intelligence artificielle sur l'ARN

L'utilisation du deep-learning et du machine learning afin de créer des modèles de prédiction de l'ARN est très courant avec des applications variées concernant par exemple la structure 3D, l'attachement de ligand ou encore leur fonction. Cependant, la quasi-totalité de ces papiers scientifique se base sur une représentation 3D de l'ARN et très peu d'article de machine learning et deep-learning utilisent une représentation en graphe. C'est le cas d'une publication de l'équipe [5] dans lequel se déroule le stage et qui a fortement influé le travail réalisé au cours du stage.

Résultats

Performances du modèle de prédiction sur le jeu de données non redondant

Au cours du stage, plusieurs tentatives de modélisation ont été réalisées en modifiant le jeu de données ou le script utilisé pour la classification.
Dans un premier temps, une prédiction sur la totalité du jeu de données a été réalisé où tous les nucléotides étaient prédits par le même modèle (1er ligne de la table 1). Par la suite, les nucléotides provenant du même nucléotide originel (A, U, G ou C) ont été regroupés en quatre jeux de données différents. Un modèle de prédiction distinct a été alors mis en place pour chaque jeu de données (2d ligne de la table 1). Ceci permet au modèle de ne pas avoir à différencier les nucléotides non modifiés entre eux (devoir décider si un nucléotide est un "G" ou un "A" par exemple), ce qui n'est pas l'objectif dans notre tache de prédiction.
Nous avons ensuite émis l'hypothèse a ensuite été émise que le nombre d'attributs était trop grand et que certains d'entre eux possédaient une variabilité trop faible et que cela impactait les performances de notre modèle. Un nouveau jeu de données a alors été mis au point en regroupant les attributs concernant les types d'arrêts non canoniques, présentes en moins grand nombre (3e ligne de la table 1).

Type de prédiction		**Classification pour chaque type de nucléotide**			**Classification Binaire "Modifié" "Non-modifié"**		
Algorithme		**SVM**	**KNN**	**RF**	**SVM**	**KNN**	**RF**
Jeu de donnée total		0.29±0.02	0.27±0.01	0.26±0.01	0.60±0	0.62±0.01	0.55±0.02
Jeu de donnée globale avec séparation	**A**	0.44 0.03	**0.48±0.06**	0.43±0.03	0.67±0.01	0.59±0.01	0.54±0.01
	U	0.26±0.02	**0.42±0.08**	0.23±0.01	0.66±0.01	0.58±0.02	0.52±0
	G	**0.32±0.03**	**0.32±0.03**	0.26±0.02	0.58±0.01	0.54±0.01	0.51±0.01
	C	0.32±0.02	**0.43±0.14**	0.30±0.02	0.62±0.02	0.61±0.01	0.58±0.03
	Moy	0.33±0.08	**0.41±0.07**	0.31±0.09	0.63±0.04	0.58±0.03	0.53±0.03
Jeu de donnée réduit avec séparation	**A**	0.47±0.02	**0.48±0.05**	**0.48±0.05**	**0.69±0.02**	0.60±0.01	0.54±0.00
	U	0.30±0.05	0.33±0.4	0.23±0.02	**0.71±0.02**	0.56±0.01	0.53±0.01
	G	0.30±0.02	0.30±0.02	0.26±0.02	**0.68±0.04**	0.55±0.01	0.50±0.01
	C	0.34±0.03	0.35±0.03	0.30±0.02	**0.63±0.02**	0.57±0.03	0.52±0.02
	Moy	0.35±0.08	0.36±0.08	0.32±0.11	**0.69±0.03**	0.60±0.02	0.54±0.01

Table 1 : Mesure de la précision moyenne de chaque modèle de prédiction SVM, KNearestNeighborsClassifier (KNN) et RandomForest (RF) pour chaque tentative expliquée ci-dessus.

Dans chaque cas, une première classification a été réalisée dans lequel chaque nucléotide correspond à une classe à prédire. Par la suite, les nucléotides modifiés ont été regroupés en une seule classe, transformant le problème en une classification binaire (“Modifié” contre “Non modifié”).

On remarque que la prédiction est généralement meilleure avec une classification binaire, ce qui est contrebalancé par la perte d’information du type de modification que le nucléotide a subi. Les modèles conçus possédaient de très bonnes performances pour la prédiction des classes principales, mais avec une mauvaise prédiction des autres classes (correspondant aux nucléotides modifiés).
Il a aussi été constaté que certain nucléotide sont significativement mieux prédit que d’autre. C’est le cas de *UR3* par exemple bien prédit dans 50 % des cas où encore de *6MZ* bien prédit dans 60 % des cas.

Performances du modèle de prédiction sur le jeu de donnée hybride

Afin de compenser la sous-représentation des nucléotides modifiés, un second jeu de données a été mis au point. Les données des nucléotides non modifiés ont été récupérées à partir des structures des représentants de chaque classe de *RNA Structure Atlas* [2] et les données des nucléotides modifiés ont été récupérées sur l’ensemble des structures d’ARN (même celles considérées comme redondantes).

Type de prédiction	**Classification pour chaque type de nucléotide**		
Algorithme	**SVM**	**KNN**	**RF**
A	0.93±0.01	0.88±0.01	**0.95±0.00**
U	0.91±0.01	0.87±0.01	**0.92±0.01**
G	**0.95±0.01**	0.89±0.01	**0.95±0.01**
C	0.90±0.01	0.85±0.01	**0.92±0.01**
Moyenne	0.92±0.01	0.87±0.01	**0.94±0.01**

Table 2 : Mesure de la précision moyenne de chaque modèle de prédiction SVM, KNearestNeighborsClassifier (KNN) et RandomForest (RF) sur le second jeu de donnée

Nous pouvons voir que les performances sont vraiment très bonnes. Le modèle réussit à prédire de manière presque parfaite certains groupes de nucléotides.

Étude de l'impact des différents niveaux de contexte structurels sur les performances de prédiction

Après avoir réussi à établir un modèle performant, une question importante était de savoir si les liaisons "base pair", et plus précisément les liaisons non canoniques, jouaient un rôle important dans l'identification d'un nucléotide modifié.

Pour cela, plusieurs jeux de données ont été construits à partir de graphes dont les liaisons non canoniques ont été retirées dans un premier temps. Puis un second jeu de donnée a été mis au point où toutes les liaisons "base pair" pour ne se concentrer que sur le squelette d'ARN, donc uniquement les nucléotides adjacents dans la séquence. Le même modèle a alors été entrainé sur ces nouveaux jeux de données pour mesurer l'impact de l'information apportée par les liaisons qui manquante.

Type de prédiction	**Classification uniquement avec les arêtes du squelette**			**Classification avec les liaisons canoniques et du squelette**		
Algorithme	**SVM**	**KNN**	**RF**	**SVM**	**KNN**	**RF**
A	**0.99±0.00**	**0.97±0.01**	0.83±0.10	**0.98±0.01**	0.92±0.01	0.96±0.01
U	0.72±0.04	0.66±0.02	**0.76±0.05**	**0.74±0.06**	0.71±0.05	**0.74±0.01**
G	0.61±0.02	0.52±0.01	**0.62±0.02**	**0.55±0.08**	0.49±0.08	**0.55±0.08**
C	0.63±0.04	0.63±0.06	**0.66±0.04**	0.62±0.01	0.56±0.03	**0.64±0.03**
Moyenne	**0.73±0.17**	0.70±0.20	0.72±0.10	**0.72±0.19**	0.67±0.19	**0.72±0.18**

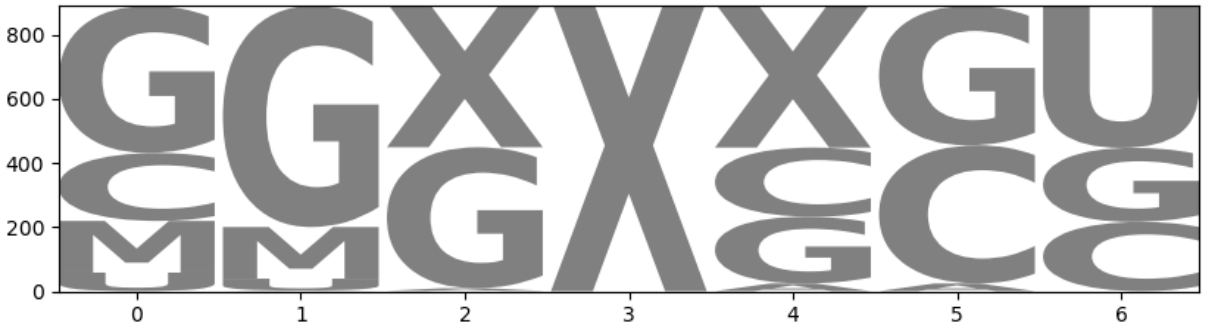
Table 3 : Mesure de la précision moyenne de chaque modèle de prédiction SVM, KNearestNeighborsClassifier (KNN) et RandomForest (RF) sur le second jeu de donnée lorsque uniquement les liaisons du squelette ont été pris en compte (colonne 1) et lorsque les liaisons du squelette ainsi que les liaisons Watson-Crick ont été pris en compte (colonne 2).

Nous pouvons voir qu'en moyenne les performance du modèle n'utilisant que le squelette sont bonnes malgré la perte d'information. De plus, le premier modèle performe aussi bien que le second.

Étude de la séquence entourant les motifs les mieux prédits

Au vu des performances du modèle de prédiction ne prenant en compte que les liaisons du squelette de l'ARN, et donc ne s'intéressant qu'aux voisins directs du nucléotide. L'hypothèse a été émise que la séquence jouait un rôle important dans la détermination des nucléotides modifiés. Des Logos autour de la séquence aux niveaux des nucléotides modifiés ont alors été générés dans l'espoir d'observer des paternes intéressant.

(1)



(2)

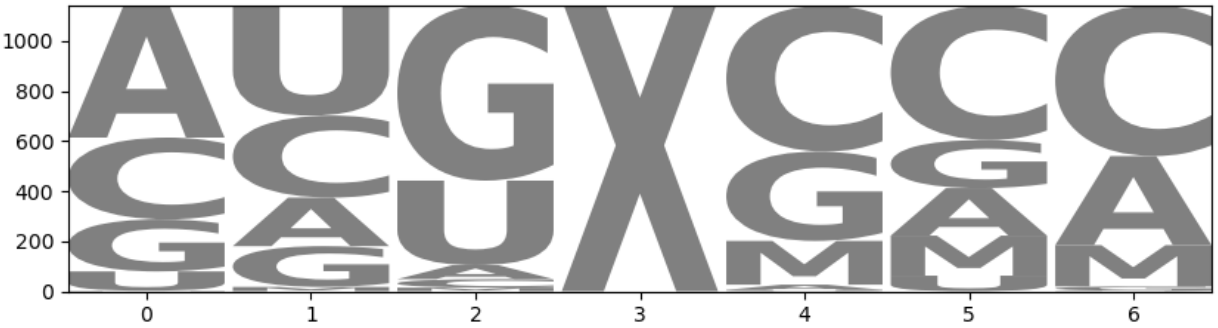


Figure 3 : Exemple de Logos intéressants générés au cours du stage pour MA6 (1) et 2MG (2). X représente le nucléotide modifié étudié et M représente un nucléotide modifié autre que celui étudié.

La figure 3.1 semble indiquer que les nucléotides MA6 sont presque tout le temps par paire dans une configuration G-MA6-MA6-(G ou C).
En observant les séquences entourant les nucléotides modifiés de type 2MG, il semblerait que ces derniers soient souvent suivis de triplet CCC ou de trois nucléotides modifiés.
Ces exemples sont parmi les plus parlants, mais ne sont pas les seuls observés et d'autre paternes peuvent être observées pour certains types de nucléotides.

Méthodes

Jeu de données

Le jeu de données utilisé durant ce stage correspond à une collection de structures d'ARN représentées en graphes qui avait déjà été établie précédemment par l'équipe de recherche pour un travail précédent.
Il est important de noter que les structures d'ARN sont très redondantes et sont répartis en classe en fonction de leurs homologies sur *RNA Structure Atlas* [2] où chaque classe possède une structure "représentante" de tous les membres de la classe. Un enjeu important est donc de prendre en compte ces redondances pour éviter des biais et un sur-apprentissage dans la modélisation en entraînant les modèles sur des graphes très similaires. De plus, toutes les liaisons autres que celle du squelette et de type "base pair" ont été enlevées des graphes n'étant pas l'objet de l'étude.

Modèle de prédiction des nucléotides modifiés

Comme dit précédemment, l'objectif était d'être capable de prédire le type d'un nucléotide présent dans le graphe. Cependant, afin d'éviter des biais, il était important de ne pas considérer les liaisons "base pair" que pourrait faire ce nucléotide, car ces informations sont en partie déterminées par le type du nucléotide supposé inconnu. Il aurait été possible de vectoriser les graphes en utilisant un réseau de neurones, ce qui est utilisé couramment dans les papiers de prédiction sur les graphes. Cependant, ce travail étant une introduction, nous voulions dans un premier temps voir si une bonne prédiction était envisageable en récupérant les propriétés connues de l'environnement de chaque nœud comme attributs. Cette approche possède l'avantage de pouvoir analyser les caractéristiques impactantes dans la prédiction, ce qui nous permet d'émettre des hypothèses, les facteurs liés à la présence de nucléotide modifiés. Pour cette raison, une vectorisation manuelle de l'ensemble des graphes a été réalisé de la manière suivante :

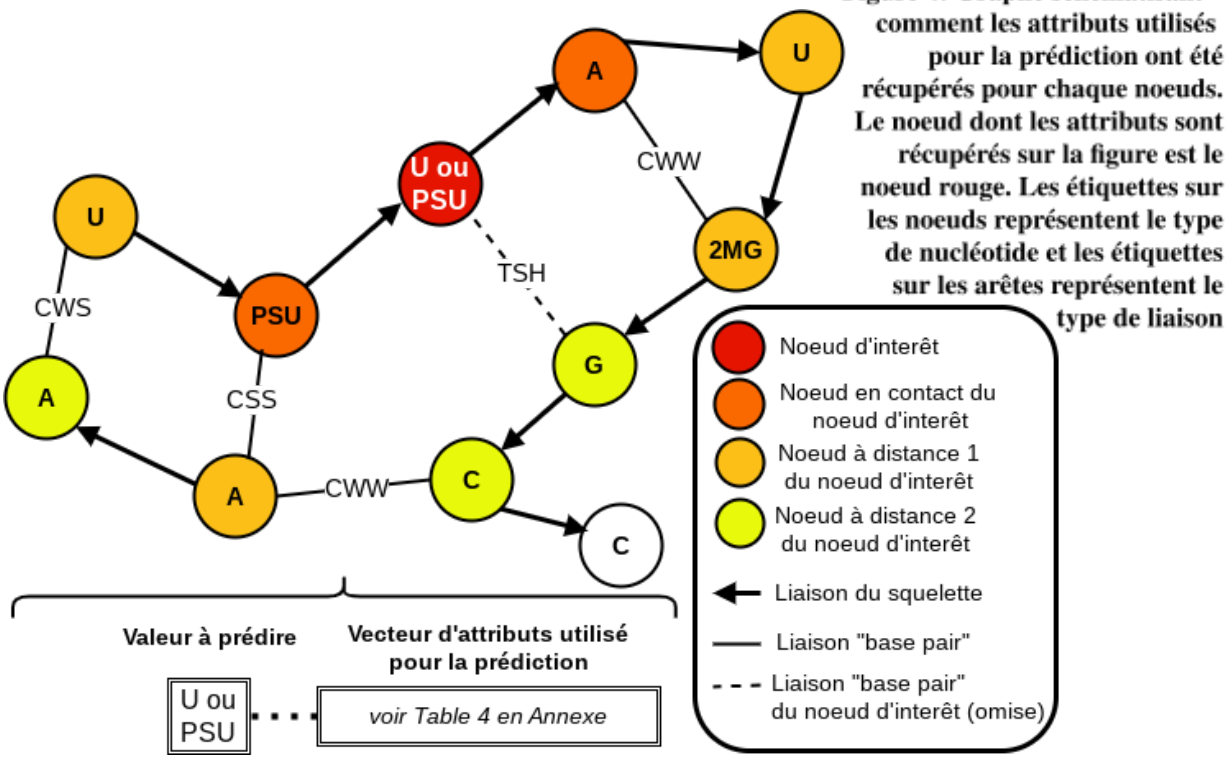


Figure 4: Graphe schématisant comment les attributs utilisés pour la prédiction ont été récupérés pour chaque noeuds. Le noeud dont les attributs sont récupérés sur la figure est le noeud rouge. Les étiquettes sur les noeuds représentent le type de nucléotide et les étiquettes sur les arêtes représentent le type de liaison

On réalise cette opération pour tous les nœuds de chaque graphe, nous permettant d'obtenir un jeu de données de 50 attributs (présent dans la table 4 en annexe, avec les valeurs récupérées pour l'exemple de la figure 2).

Première tentative de prédiction des nucléotides modifiés

Dans un premier temps, seulement les représentants de chaque classe de *RNA Structure Atlas* [2] ont été utilisés pour entraîner et évaluer le modèle afin d'être certain de ne pas avoir de biais dû à l'homologie des structures.
À partir de ces données, uniquement les 17 nucléotides les plus fréquents ont été sélectionnés car une grande majorité des nucléotides modifiés était peu fréquents et auraient mal été prédits par les algorithmes du fait du déséquilibre entre les classes. Les effectifs des nucléotides sélectionnés sont présents dans la table 5 en annexe.

Ensuite, le jeu de données a été réparti entre un jeu de données d'entraînement et un jeu de données de test. Le jeu d'entrainement à ensuite subit un sur-échantillonnage aléatoire pour équilibrer l'effectif des classes.

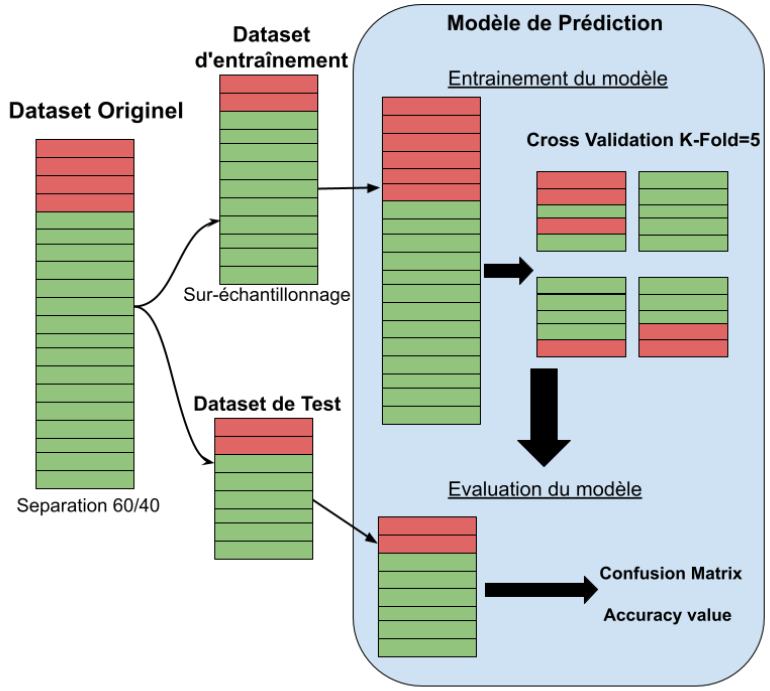


Figure 5 : Schéma représentant le pipeline utilisé pour entrainer et évaluer les modèles de machine learning

Les hyperparamètres des trois algorithmes de prédiction sont présents en annexe (table 7) et ont été établis sur un jeu de données réduit en utilisant un algorithme de GridSearch "naïf" qui permet de tester toutes les combinaisons de paramètre d'un modèle pour trouver la combinaison donnant les meilleurs résultats. De plus, les algorithmes de SVM et KNearestNeighborsClassifier peuvent prendre en compte des poids d'apprentissage pour chaque classe (illustrant à quel point le modèle va être impacté par son apprentissage sur une donnée en fonction de sa classe). Les poids des classes correspondant aux nucléotides modifiés ont été mis à 100 et les poids des classes correspondant aux nucléotides non modifiés ont été gardés à 1 pour compenser le déséquilibre entre les classes. Plusieurs algorithmes de sur-échantillonnage comme SMOTE, SMOTENN, ADASYN et Random Over-Sampler [6] sont utilisés pour tester quelles données générer pour l'entraînement permettait d'obtenir les meilleures performances des algorithmes de prédiction.

Random Over-Sampler avait de meilleurs résultats de manière systématique et a donc été retenu comme algorithme de sur-échantillonnage.

Les fonctions de traitement des données et les algorithmes de machine learning proviennent du package imblearn [6] et scklearn [7]. Le script d'entraînement et d'évaluation ainsi que le pipeline pour obtenir le jeu de données a été codé en python au cours du stage.

Seconde tentative de prédiction des nucléotides modifiés

Pour la construction de ce jeu de données, des nucléotides provenant de contexte structurel très similaires ont été utilisés, il était donc de limiter le biais entrainé par ces derniers.
Pour cette raison, la séparation des jeux de données d'apprentissage et de test a été fait en fonction de la classe de redondances d'où provenaient les nucléotides. De cette manière, les nucléotides utilisés dans le test du modèle ne proviennent pas de structures jugées comme trop homologues par RNA Structure Atlas [2] à celles qui ont servi pour son entrainement. Après avoir fait cette séparation, le même script a été utilisé pour entraîner et évaluer le modèle. La même sélection a été faite pour pouvoir comparer les performances des modèles et leurs effectifs ont été répertoriés dans la table 6 en annexe.

Étude de la séquence des ARN autours des nucléotides modifiés

Les Logos ont été générés en utilisant le package LogoMaker [8] de python, les nucléotides modifiés (dont les symboles sont constitués de trois lettres) n'était pas compatible et leur notation a dû être remplacé par des symboles à une lettre.
Les Logos ont été réalisés avec l'ensemble des séquences entourant chaque occurrence de nucléotide modifié dans l'ensemble des représentants des classes de redondance de RNA Structure Atlas [2].

Discussion et conclusion

Le premier modèle de prédiction mis en place n'a pas de très bonne performance, il est possible d'émettre l'hypothèse que ceci soit dû à là sous représentation des nucléotides modifiés. Les problèmes de déséquilibre dans les classes étant un problème récurrent dans l'exercice de prédiction en intelligence artificielle, explorer d'autre méthode connu pour y faire face pourrait augmenter la capacité prédictible du modèle. La méthode utilisée pour compenser le déséquilibre des classes lors de ce stage est le sur-échantillonnage. Cependant, il est possible que le nombre de données à sur-échantillonner est tellement faible que les nouvelles données créées soient redondantes, entrainant alors du sur-apprentissage et ne règlent pas efficacement le déséquilibre de classe.
Les nucléotides modifiés étant encore un domaine peu étudié, le jeu de données utilisé ne représentait pas l'ensemble des structures présentes dans la PDB. L'agrandissement de ce jeu de données nous donnerait plus d'effectifs dans les classes sous-représentées. Ceci aurait comme potentielle conséquence la sur-évaluation des performances des modèles

En ajoutant les nucléotides modifiés provenant des structures à forte homologie, les prédictions sont bien meilleures. Il est possible qu'un biais soit introduit par ces homologies malgré les différentes mesures pour le contenir. Les performances des modèles seraient donc possiblement surévaluées.
On remarque que certains nucléotides sont bien mieux prédit que d'autre. Ceci semblerait indiquer que les attributs recueillis ont un lien avec leur présence. Il serait alors intéressant d'étudier l'environnement de ces nucléotides pour voir si des paternes récurrents peuvent être observés. Ces informations permettraient de mieux comprendre ces modifications et de savoir par exemple si c'est l'environnement qui engendre la modification du nucléotide, ou si c'est elle qui est à l'origine de son environnement singulier. Une autre approche complémentaire serait d'étudier les propriétés physico-chimiques des nucléotides les mieux prédits, afin d'observer s'ils ont des points communs ou des différences remarquables et de lier ces propriétés à leur environnement dans la structure.

Pour commencer à répondre à ces questions, des modèles de prédictions ont été réalisés, ne prenant en compte que les liaisons canoniques et le squelette, puis uniquement le squelette. Une perte du pouvoir prédictif est observée, mais le modèle reste bon lorsque exclusivement les liaisons du squelette de l'ARN sont prises en compte. Une hypothèse pour expliquer cette observation est que, pour beaucoup de nucléotides, la séquence autour du nucléotide est la cause directe de l'apparition de la modification. Il existerait alors des motifs dans les séquences d'ARN qui engendreraient ces modifications post traductionnel. Les Logos générés à la fin du stage vont dans le sens de la présence de motifs dans la séquence autour des nucléotides modifiés. Ils ne constituent cependant pas une preuve suffisante en eux-mêmes et il est possible que les motifs observés soient en réalité un biais dû à de l'homologie entre les structures. En effet, les nucléotides modifiés dans les structures d'ARN ont seulement été identifiés pour une minorité de structures, qui sont les plus connues et les plus étudiées. Une majorité des nucléotides modifiés du jeu de donnée sont donc présents dans des structures populaires comme les ribosomes, certains ARN mitochondriaux et certains ARN de transfert.
Cette surreprésentation de structure ayant les mêmes fonctions pourraient entraîner un faux-semblant de récurrence dans les séquences autour des nucléotides modifié.

Conclusion

De manière générale, les résultats sont très encourageants pour une première tentative et ils ont montré que des paternes semblaient régir la présence de nucléotide modifié dans l'ARN.
Sur le plan méthodologique, l'utilisation d'intelligence artificiel appliquée aux graphes d'ARN est aussi une avancée qui paraît prometteuse pour répondre à certaines questions concernant la structure de l'ARN et il est possible que l'utilisation du deep-learning pour encoder les graphes rajoutent encore de nouvelles applications.

Il est possible qu'avoir plus de données, ce qui viendrait avec le temps, ou tenter d'autre implémentation, permettent d'obtenir des modèles ayant de meilleures performances en utilisant uniquement les nucléotides des structures non redondantes. Il est aussi important de noter que la durée de ce stage est de quatre mois. Ce rapport est donc écrit un peu avant la moitié du stage, les méthodes ainsi que les résultats qui y figurent sont intermédiaires et amenés à évoluer. Durant la suite du stage, vérifier les hypothèses quant à la présence de motifs autour des nucléotides modifiés sera l'objectif principal. Une approche en utilisant les réseaux bayésiens adaptés à l'ARN (approche déjà mise en place par une autre équipe partenaire [9]) et plus spécifiquement aux nucléotides modifiés est déjà en train d'être mise en place afin de voir s'il existe des éléments structuraux redondants autour des nucléotides modifiés.

Bibliographies

[1] Reinharz V, Soulé A, Westhof E, Waldispühl J, Denise A. Mining for recurrent long-range interactions in RNA structures reveals embedded hierarchies in network families. Nucleic Acids Research. 03 2018;46(8):3841-3851. doi:10.1093/nar/gky197

[2] Petrov AI, Zirbel CL, Leontis NB. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *RNA*. 2013;19(10):1327-1340. doi:10.1261/rna.039438.113

[3] Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. *RNA*. 2001;7(4):499-512. doi:10.1017/s1355838201002515

[4] Boccaletto P, Stefaniak F, Ray A, et al. MODOMICS: a database of RNA modification pathways. 2021 update. Nucleic Acids Research. 12 2021;50(D1):D231-D235. doi:10.1093/nar/gkab1083

[5] Oliver C, Mallet V, Gendron RS, et al. Augmented base pairing networks encode RNA-small molecule binding preferences. Nucleic Acids Research. 07 2020;48(14):7690-7699. doi:10.1093/nar/gkaa583

[6] Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. arXiv [csLG]. Published online 2016. http://arxiv.org/abs/1609.06570

[7] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-2830.

[8] Tareen A, Kinney JB. Logomaker: Beautiful sequence logos in python. bioRxiv. Published online 2019. doi:10.1101/635029

[9] Sarrazin-Gendron R, Yao HT, et al. Stochastic Sampling of Structural Contexts Improves the Scalability and Accuracy of RNA 3D Module Identification. Schwartz R, ed. Research in Computational Molecular Biology. Springer International Publishing; 2020:186-201

Annexes

Type du nucléotide suivant	A	**Nombre de liaisons du squelette à distance de 1**	4	**Nombre de liaisons CHH à distance de 2 nœuds**	0
Type du nucléotide précédent	Mod	**Nombre de liaisons CSH à distance de 1 nœud**	0	**Nombre de liaisons du squelette à distance de 2**	2
Nombre de nucléotides de type A à distance de 1 nœud	1	**Nombre de liaisons CHS à distance de 1 nœud**	0	**Nombre de liaisons CSH à distance de 2 nœuds**	0
Nombre de nucléotides de type U à distance de 1 nœud	2	**Nombre de liaisons THS à distance de 1 nœud**	0	**Nombre de liaisons CHS à distance de 2 nœuds**	0
Nombre de nucléotides de type G à distance de 1 nœud	0	**Nombre de liaisons TSH à distance de 1 nœud**	0	**Nombre de liaisons THS à distance de 2 nœuds**	0
Nombre de nucléotides de type C à distance de 1 nœud	0	**Nombre de liaisons CWH à distance de 1 nœud**	0	**Nombre de liaisons TSH à distance de 2 nœuds**	0
Nombre de nucléotides de type Modifié à distance de 1 nœud	1	**Nombre de liaisons CHW à distance de 1 nœud**	0	**Nombre de liaisons CWH à distance de 2 nœuds**	0
Nombre de nucléotides de type A à distance de 2 nœuds	1	**Nombre de liaisons THW à distance de 1 nœud**	0	**Nombre de liaisons CHW à distance de 2 nœuds**	0
Nombre de nucléotides de type U à distance de 2 nœuds	0	**Nombre de liaisons TWH à distance de 1 nœud**	0	**Nombre de liaisons THW à distance de 2 nœuds**	0
Nombre de nucléotides de type G à distance de 2 nœuds	1	**Nombre de liaisons CWS à distance de 1 nœud**	1	**Nombre de liaisons TWH à distance de 2 nœuds**	0
Nombre de nucléotides de type C à distance de 2 nœuds	1	**Nombre de liaisons CSW à distance de 1 nœud**	0	**Nombre de liaisons CWS à distance de 2 nœuds**	0
Nombre de nucléotides de type Modifié à distance de 2 nœuds	0	**Nombre de liaisons TWS à distance de 1 nœud**	0	**Nombre de liaisons CSW à distance de 2 nœuds**	1
Nombre de liaisons TWW à distance de 1 nœud	0	**Nombre de liaisons TWS à distance de 1 nœud**	0	**Nombre de liaisons TWS à distance de 2 nœuds**	0
Nombre de liaisons TSS à distance de 1 nœud	0	**Nombre de liaisons TWW à distance de 2 nœuds**	0	**Nombre de liaisons TWS à distance de 2 nœuds**	0
Nombre de liaisons THH à distance de 1 nœud	0	**Nombre de liaisons TSS à distance de 2 nœuds**	0		
Nombre de liaisons CWW à distance de 1 nœud	2	**Nombre de liaisons THH à distance de 2 nœuds**	0		
Nombre de liaisons CSS à distance de 1 nœud	1	**Nombre de liaisons CWW à distance de 2 nœuds**	1		
Nombre de liaisons CHH à distance de 1 nœud	0	**Nombre de liaisons CSS à distance de 2 nœuds**	0		

Table 4 : Tableau de toutes les propriétés récupérées dans les graphes pour chaque nœud et utilisées comme attribut lors de la prédiction (les valeurs présentes ici sont celles récupérées pour le nœud rouge de la figure 4)

Effectifs des différents jeux de données utilisés pour entrainer et évaluer les modèles

A	52429
A2M	32
6MZ	27

U	46244
PSU	56
OMU	24
H2U	18
5MU	15
UR3	4

G	60098
OMG	36
2MG	20
7MG	17

C	49120
5MC	43
OMC	31
4OC	4

Table 5 : Nucléotides (1re colonne) ainsi que leurs effectifs (2d colonne) sélectionnés dans le jeu de données non redondant séparés en fonction du nucléotide "originel" (A, U, G ou C)

A	52429
A2M	1417
6MZ	641

U	46244
PSU	6325
5MU	1950
OMU	1144
H2U	739
UR3	713

G	60098
OMG	2089
2MG	1599
7MG	840

C	49120
5MC	2902
OMC	1474
4OC	665

Table 6 : Nucléotides (1re colonne) ainsi que leurs effectifs (2d colonne) sélectionnés dans le jeu de données hybride séparés en fonction du nucléotide "originel" (A, U, G ou C)

Hyperparamètres et architecture des modèles de prédiction

- Modèle de machine learning

SVM	
regularization strength	10
kernel type	‘rbf’
kernel coefficient	0.002

KNearestNeighborsClassifier	
number of neighbors	10
algorithm	‘kd_tree’
leaf size	360
metric	‘minkowski’
weights	'distance’

RandomForestClassifier	
criterion	‘gini’
minimum split size	2
minimum life size	1
bootstrap	True

Table 7 : Hyperparamètres des modèles mis en place