Université de Montréal

**Prédiction et visualisation de la réactivité chimique des ARN**

**Proposition d’un modèle basé sur la cohérence des motifs cycliques nucléotidiques avec les données de réactivité chimique.**

par Philippe Malric

Département de biochimie

Faculté de Médecine

Mémoire présenté  
en vue de l’obtention du grade de Maître

en Bio-informatique

Décembre 2017

© Philippe Malric, 2017

# Résumé

Mesurer la réactivité chimique d’un nucléotide d’une séquence d’ARN permet d’inférer sa conformation locale. L’interprétation la plus simple de la conformation locale d’un nucléotide est de le qualifier par le terme pairé ou non pairé. Bien qu’ils soient reliés à la réactivité chimique des nucléotides, ces deux états ne l’expliquent pas entièrement.

Actuellement, la RNA mapping database (RMDB) contient 131,271 séquences sondées dans différentes conditions, ce qui donne un bon jeu de données pour étudier la structure des ARN en utilisant des algorithmes d’apprentissages machines. Dans mon projet de recherche, j’ai tenté de prédire et de comprendre la réactivité ou la non-réactivité des nucléotides à partir de plusieurs ensembles de données obtenues de la RMDB. Pour cela, j’ai utilisé deux algorithmes de repliement des ARN en structures secondaires : MCFlashFold du laboratoire de Dr François Major et RNAsubopt de la collection Vienna package. L’originalité de mon projet de recherche est qu’il se base sur une abstraction des composants de la structure secondaire nommée motif cyclique nucléotidique (MCN) pour caractériser les nucléotides.

Dans ce mémoire, je fais la démonstration que la majorité des valeurs de réactivités discrètes (hautes/basses) des nucléotides peuvent être prédites. L’apport majeur de ce mémoire est la mise sur pied d’une interface permettant d’évaluer la cohérence de la réactivité des nucléotides en les comparant à l’ensemble des nucléotides d’un même MCN, leurs homologues.

Vous constaterez aussi que le nombre de caractéristiques mesurables d’un nucléotide croît de façon exponentielle lorsqu’on considère les MCN adjacents. Ceci nous mène donc à nous poser la question suivante : peut-on expliquer la réactivité des nucléotides par un ensemble fini de règles formelles?

**Mots-clés** : ARN, MCN, structure secondaire, SHAPE, Eterna, RMDB

# Abstract

Measuring the chemical reactivity of a nucleotide within a sequence can infer the local conformation of a nucleotide most of the time. The simplest interpretation of the local conformation of a nucleotide is by the term paired or non-paired. However, these two states do not fully explain the reactivity of nucleotides in chemical probing experiments.

Currently, the RNA mapping database (RMDB) contains 131,271 sequences probed under different conditions, which gives us a relatively good data set to learn using machine learning algorithms. In my research project, I attempted to predict and understand the reactivity or non-reactivity of nucleotides of thousand of RNA obtained from RMDB. For this, I used two secondary structure RNA folding algorithms: MCFlashFold from Dr. François Major's laboratory and RNAsubopt from the Vienna package collection. The originality of my project is that it is based on an abstraction of the secondary structure called nucleotide cyclic motif.

In this thesis, I demonstrate that the majority of the discretized (high / low) reactivities values of nucleotides can be predicted, but not all. The proposed interface gives us the opportunity to quickly and easily compare homolog nucleotides between RNA.

You will also find that the number of measurable features of a nucleotide exponentially increases when nucleotide neighbors are considered. So, the question is what characteristics should be extracted to accurately predict the reactivity of a nucleotide?

**Keywords** : RNA, NCM, secondary structure, SHAPE, Eterna, RMDB

# Table des matières

Résumé 3

Abstract 4

Table des matières 5

Liste des tableaux 7

Liste des figures 8

Liste des sigles 14

Liste des abréviations 14

Remerciements 16

1 Introduction 17

1.1 Domaine de recherche 17

1.2 Mise en contexte 18

1.3 Historique 21

1.4 Pourquoi étudier la structure de l’ARN 22

1.5 Définition des termes et des concepts 23

Chapitre 1  RNASS : intégration de deux logiciels de prédiction des structures secondaire 45

De l’obtention des données à la prédiction discrète 45

Information sur les données collectées 51

Prédictions basées sur la cohérence 56

Conclusion du chapitre 1 57

Chapitre 2 : RDV : un logiciel de visualisation de la structure secondaire des ARN 58

Compilation des MCN 58

Visualisation du graphe des transitions 68

Visualisation de la cohérence 68

Visualisation de la SS 69

Obtenir des détails et rechercher des ARN semblables 69

Conclusion du chapitre 2 74

Chapitre 3 : Évaluation de la cohérence des MCN 75

Comparaison des prédictions de RNASS avec celles faites à l’aide de l’état pairé ou non d’un nucléotide. 77

Conclusion du chapitre 3 81

Conclusion 82

Qu’est-ce qui fait qu’un nucléotide réagit? 82

Champ d’étude à venir et algorithmes à considérer 83

Bibliographie i

Annexe i

Caractéristiques du serveur utilisé i

Identification des MCN i

API d’Eterna i

Évaluation du score de prédiction ii

Score de prédictions de la SS de de la *MFE* dans l’ensemble non filtré ii

Portrait d’un ARN i

# Liste des tableaux

Tableau I. Description des champs des fichiers JSON standardisés des ARN 48

Tableau II. Description du champs *nts* des fichiers JSON standardisés 49

Tableau III. MCN les plus présents et leurs valeurs de réactivités dans l’ensemble non-filtré 60

Tableau IV. Fréquence des nucléotides des MCN de RNAsubopt et MCFlashfold ordonnée selon la somme des nucléotides de hautes et basses réactivités de RNAsubopt dans l’ensemble de données filtré. 62

Tableau V. Cohérence des 10 MCN les plus fréquents et unique à MCFlashfold 65

Tableau VI. Signification des valeurs de prédictions 1 et -1 dans RDV 69

Tableau VII. Signification possible de la valeur 0 lors des prédictions et couleur du contour du nucléotide dans RDV 70

Tableau VIII. Tableau de contingence de l’état pairé ou non des nucléotides en fonction de leur niveau de réactivité. 76

Tableau IX. Comparaison entre le modèle basé sur l’état pairé ou non des nucléotides et le modèle de basé sur le score de RNASS 79

# Liste des figures

[Figure 1. **Nombre de séquences d’ARN ajoutées à la RMDB entre septembre 2014 et mars 2017**. Figure modifiée de l’article : *Updates to the RNA mapping database (RMDB), version 2.* Nucleic Acids Research, 2017. 22](#_Toc502671236)

[Figure 2. **A) Représentation d’un nucléotide, l’adénosine monophosphate (AMP)**. Le cercle rouge, le pentagone bleu et le carré vert représentent le groupement phosphate, le ribose et la base azotée, respectivement. Cette image a été modifié de l’article : *Structural analysis of nucleic acid aptamers* *.* Current opinion in chemical biology, 1997. **1**: p. 32-46, écrit par Patel, D.J. B) image 3D de A, avec en plus un groupe phosphate sur le carbone en 3’ du ribose. Source : PubChem3D. 26](#_Toc502671237)

[Figure 3. **Un groupement phosphate en 3D**. L’atome de phosphore, en vert, est au centre des atomes d’oxygène rouges. Les hydrogènes sont en blancs. Source : PubChem3D. 27](#_Toc502671238)

[Figure 4. **Représentation en balles et bâtons (*balls and sticks*) du beta-D-ribose**. Dans le polymère qu’est l’ARN, les groupes phosphate s’attachent aux carbones 3 et 5 du ribose. Le carbone 1, en beta, est lié à la base azotée. Les oxygènes sont rouge, les carbones gris et les hydrogènes blancs. Cette image est disponible sur https://commons.wikimedia.org/, elle a été produite par, Marina Vladivostok. 27](#_Toc502671239)

[Figure 5. A) L’adénine est une base azotée de la famille des purines. Son symbole est le A. B) La guanine est aussi une base azotée de la famille des purines. Son symbole est le G. C) La cytosine est une base azotée de la famille des pyrimidines. Son symbole est le C. D) Tout comme la cytosine, l’uracile est une base azotée de la famille des pyrimidines. Son symbole est le U. 28](#_Toc502671240)

[Figure 6. **Les MCN sont des structures cycliques dans un réseau formé par les liens sucre-phosphate et les paires de bases de l’ARN.** Dans la figure de gauche, trois MCN sont présents, les carrés pointillés bleu et vert englobent un MCN 2\_2. Celui qui est pointillé rouge entour une boucle formée par 5 nucléotides. À droite, se trouve les expressions régulières des MCN. La figure de gauche est tirée du manuel d’instruction de MCFlashfold (2015, Dallaire, P. and F. Major, *MCFlashfold (mcff version 34) user manual).* 30](#_Toc502671241)

[Figure 7. **Réaction chimique du 1M7 avec le 2’OH du sucre d’un nucléotide.** Cette image provient du wiki d’Eterna, un jeu vidéo consacré à la structure des ARN : http://eternawiki.org/wiki/index.php5/1M7. Elle a été créée par Omei. 36](#_Toc502671242)

[Figure 8. Représentation en deux dimensions de deux ARN provenant de la « *protein data bank* » (PDB) par le logiciel FORNA. Les structures se nomment « *RNA LOOP-LOOP COMPLEX* ». Leur identifiant est : « 1BJ2 ». 40](#_Toc502671243)

[Figure 9. Représentation en trois dimensions des ARN de la figure 10. Les pointillés mauves sont des ponts hydrogènes. La couleur des nucléotides est reliée à leur position allant du bleu (5’) au rouge (3’) 40](#_Toc502671244)

[Figure 10. **Prédiction de la SS faite par MCFlashfold (A) et de RNAsubopt (B) sur un ARN ayant une queue polyadénylée illustrée par RDV**. La couleur et la grosseur des nucléotides sont reliées à leur réactivité observée et la couleur de leur contour est reliée à la prédiction faite en compilant la valeur de plusieurs millions de nucléotides. La prédiction est une valeur de cohérence. Le bleu et le rouge représentent une réactivité élevée et basse respectivement. Le gris signifie que nous ne disposons pas de données de réactivité pour ces nucléotides. La SS de MCFlashfold est celle de gauche (A). Le symbole des nucléotides les identifie au centre de leur cercle. La forme n’est pas représentative de la structure 3D. 50](#_Toc502671245)

[Figure 11. **Pour chaque ARN la moyenne des scores des valeurs de réactivité de ses nucléotides est calculée et ordonnée.** On remarque que peu d’ARN ont un score moyen entre 2 et 6. Ces ARN proviennent de l’ensemble non filtré. 55](#_Toc502671246)

[Figure 12. **La différence entre cette figure et la figure 10 est qu’ici les données proviennent de l’ensemble filtré.** Le filtre écarte les ARN ayant une réactivité moyenne au-dessus de 1.5, ceux ayant un contenu en adénine au-dessus de 50% et ceux ayant un rapport signal sur bruit sous 2. En plus petit à droite, se trouve la même distribution, mais représenté sous forme d’histogramme. 55](#_Toc502671247)

[Figure 13. **Distribution de la longueur des séquences de l’ensemble de données filtré.** La moyenne de la longueur est de 103.98. 57](#_Toc502671248)

[Figure 14. **L’ARN le plus long (131 nt) à gauche (A) avec le plus petit (51 nt) à droite (B).** Le mode couleur selon la réactivité est activé. Un nucléotide bleu est réactif et un nucléotide rouge l’est peu. 57](#_Toc502671249)

[Figure 15. **SS provenant de FORNA.** Les doubles hélices sont vertes, les renflements sont rouges et les loupes terminales et intérieurs bleus et jaunes respectivement. Les MCN sont placés dans des boites. La figure originale est en bas à droite. 58](#_Toc502671250)

[Figure 16. **Distribution des MCN en fonction de leur fréquence pour l’ensemble non filtré.** Une transformation logarithmique de base 10 a été appliquée pour mieux voir les MCN les plus fréquents. Chaque logiciel est indépendant par rapport à l’ordre de ses MCN. Les MCN du logiciel de RNAsubopt sont représentés par des points bleus et les MCN de MCFlashfold sont représentés par des points verts. Certains MCN caractérisent plusieurs millions nucléotides dans la base de données complète de RMDB. Pour être considéré, le nucléotide doit être dans la conformation du MCN dans plus de 20% des SS sous-optimales. On trouve dans les boite des MCN les plus fréquents l’identificateur du MCN. 62](#_Toc502671251)

[Figure 17. **Distribution des fréquences des nucléotides par MCN dans l’ensemble filtré.** Les MCN sont ordonnés selon la fréquence des MCN de MCFlashfold. 65](#_Toc502671252)

[Figure 18. **Figure 16 inversée : les MCN sont ordonnés selon la fréquence des MCN de RNAsubopt.** 65](#_Toc502671253)

[Figure 19. **Les 32 MCN les plus fréquents selon l’ordre de MCFlashfold dans l’ensemble filtré.** Ces MCN ont des valeurs semblables pour les deux logiciels. La fréquence est le nombre de nucléotides dans la conformation d’un MCN donné. 67](#_Toc502671254)

[Figure 20. **Distribution de la différence entre le nombre de nucléotides réactifs et ceux non réactifs divisé par la somme de ces deux valeurs pour les MCN communs aux deux logiciels.** À gauche, l’ordre est donné par le ratio des MCN de MCFlashfold et à droite il l’est par le ratio de RNAsubopt. 68](#_Toc502671255)

[Figure 21. **Valeur du ratio des nucléotides réactifs moins les peu réactifs, divisés par leur total pour les MCN unique à MCFlashfold.** Autrement dit, la valeur de chacun des points composant cette distribution est calculée en soustrayant le nombre de nucléotides ayant une valeur de réactivité sous 0,5 de ceux ayant une valeur au-dessus de 1 et en divisant ce nombre par la somme des nucléotides de ces deux classes pour chaque MCN. À gauche complètement, le premier nucléotide de la boucle « UUCG », lorsqu’elle est fermée par une paire UA a une valeur élevée à 73 occasions et n’a jamais une valeur basse dans toute la base de données filtrée. De l’autre côté, le deuxième nucléotide de la boucle « GCCU » lorsqu’elle est fermée par une paire CG à une valeur basse pour 41 nucléotides distincts et n’a jamais de valeur haute. Ces deux MCN se retrouvent aux deux extrémités de ce graphique. En tout, 2 500 MCN sont représentés. 70](#_Toc502671256)

[Figure 22. Vue principale de RNA Dynamic Viewer (RDV). 75](#_Toc502671257)

[Figure 23. **Nombre d’occurrences de deux MCN classé par niveau de réactivité**. Juste en dessous, en retrouve deux listes (classée par niveau de réactivité) des adresses de 10 nucléotides homologues par leur MCN. Elles sont obtenues en cliquant sur le nucléotide d’intérêt dans la simulation de la figure 22. Ces listes apparaissent après avoir cliqué sur un des deux MCN. Lorsque c’est possible, 10 exemples de nucléotides ayant une haute et basse réactivité sont écrites. Ces adresses mènent vers la représentation des ARN pour lesquels un nucléotide est dans la conformation du MCN dans au moins 20% des SS du même logiciel de prédiction que celui d’où provient le MCN d’origine. Le nucléotide est mis en évidence dans l’ARN de l’adresse par un cercle jaune. Le rectangle noir sur fond gris donne une idée de la « pureté » du MCN. Sa longueur est proportionnelle à 4 fois nombre de « Hi » divisé par le nombre de « *Low*» plus le nombre de « *Hi* ». 76](#_Toc502671258)

[Figure 24. Protocole d’apprentissage graphique, partant des données brutes jusqu’à son analyse. La plateforme Azure permet d’entrainer un modèle en effectuant de la programmation graphique. Si on fait abstraction du bloc « *summarise data* », deux chemins partent des données brutes. À gauche, le modèle est entrainé sur les prédictions de RNASS et à droite sur l’état pairé ou non des nucléotides. Dans les deux cas une machine à vecteur de support a été utilisée, bien que pour la caractéristique : état pairé ou non, les valeurs de vrai « Hi », faux « Hi », de vrai « Low » et faux « Low » peuvent être directement obtenues. Le dernier bloc rassemble les données pour créer des graphiques, tels que la courbe ROC et la précision en fonction du rappel. 79](#_Toc502671259)

[Figure 25. **Cette figure met en perspective l’état (pairé ou non) des nucléotides avec leur réactivité chimique.** Le nombre de nucléotides réactifs non pairés est semblable au nombre de nucléotides peu réactifs non pairés. Cependant, il y a beaucoup moins de nucléotides réactifs pairés que de nucléotides peu réactifs pairés. Cette figure a été faite avec 10 000 nucléotides. 80](#_Toc502671260)

[Figure 26. **Courbe de ROC du modèle de RNASS et du modèle pairé / non pairé.** Le modèle de RNASS en plus d’être paramétrable au niveau du risque de ses prédictions performe mieux que le modèle basé sur l’état pairé ou non des nucléotides. 82](#_Toc502671261)

[Figure 27. **Précision en fonction du rappel du modèle de RNASS (en bleu) et de celui basé sur l’état pairé non pairé des nucléotides (en rouge).** Plus de 1 million de nucléotides ont servi à faire ce test. Chacune des 10 SS de RNAsubopt est prise séparément. Le modèle de RNASS performe mieux que celui basé sur l’état pairé ou non des nucléotides. 82](#_Toc502671262)

[Figure 28. **Distribution des nucléotides en fonction de leur état (axe des ordonnés), de leur score de prédiction de RNASS (axe des abscisses) et de leur réactivité chimique (couleur).** Les points rouges représentent les nucléotides réactifs. Ils sont beaucoup plus présents lorsque le nucléotide est non pairé et lorsque son score de prédiction est élevé. Cette visualisation a été faite avec Orange, un logiciel d’apprentissage machine. 83](#_Toc502671263)

[Figure 29. Diagramme de composition des identificateurs des MCN fait avec regexper i](#_Toc502671264)

[Figure 30. Le coefficient de corrélation de Pearson entre les scores de prédiction des deux logiciels est de 0.763083. Ceci indique une bonne corrélation entre les deux logiciels. Le score de MCFlashfold est en vert et celui de RNAsubopt en en bleu. iii](#_Toc502671265)

[Figure 31. Cette figure est la figure 8 inversée. Ici, ce sont les données de MCFlashfold qui sont ordonnées. La corrélation est visible par le nombre presqu’inexistant d’ARN de RNAsubopt (bleu) en bas à droite iii](#_Toc502671266)

[Figure 32. Après filtration des données, la moyenne des différences passe à 10.5 soit un gain de 2 points en faveur de MCFlashfold. L’étendue des différences s’est amoindrie aussi. À la défense de MCFlashfold, les données ont été calibrées à partir d’une structure en épingle prédite par un logiciel de la même suite que RNAsubopt, ViennaPackage[53]. iv](#_Toc502671267)

[Figure 33. Correspondance entre les valeurs de sondage chimique et les nucléotides d’un ARN repliés par RNAsubopt à gauche et MCFlashfold à droite. Le score de prédiction est de 46 contre 8 pour RNAsubopt. Lorsqu’on utilise le score pondéré, RNAsubopt obtient 66.62 et MCFlashfold 28.21. iv](#_Toc502671268)

[Figure 34. ARN ayant la différence de score la plus avantageuse pour MCflashfold. La différence entre les scores de prédiction pondérés est de 6.35. Les scores sont de 22.52 pour MCFlashfold et de 16.17 pour RNAsubopt. L’ARN du haut a été replié par MCFlashfold et celui du bas par RNAsubopt. v](#_Toc502671269)

# Liste des sigles

|  |  |
| --- | --- |
| ARN | : Acide ribonucléique |
| MCN | : Motif cyclique nucléotidique |
| RMDB | : RNA mapping data base |
| SS | : Structure secondaire |
| API | : Application programming interface |
| MFE | : Énergie libre minimum (*Minimum free energy*) |
| AUC | : Aire sous la courbe (*Area under the curve*) |
| ROC | : fonction d’efficacité du récepteur (*Receiver operating characteristic*) |
| URL | : Localisateur uniforme de ressource (*Uniform Resource Locator*) |
| RMN | : Résonance magnétique nucléaire |
| Nt : | : Nucléotide |

# Liste des abréviations

|  |  |
| --- | --- |
| 3D | : Trois dimensions |
| A | : Adénosine |
| C | : Cytosine |
| U | : Uracile |
| G | : Guanine |
| ed | : diversité d’ensemble (*ensemble diversity*) |
| STN | : rapport signal sur bruit (*signal to noise ratio*) |

*Je nous dédie ce mémoire!*

# Remerciements

Mes remerciements vont à me parents, mes frères, mon directeur de recherche, Dr François Major et aux 4 professeurs qui ont accepté d'être membre de mon jury : Gertraud Burger, Pascale Legault, Sylvie Hamel et Sergei Chteinberg. De plus, je tiens à remercier mes collègues de laboratoires, en particulier, Mathieu, Gabriel, Olivier, Zohra, Jordan, Roqaya, Blandine, Albert, Thomas, Nathanael, Nicolas, Marc Frédéric, Paul, Maria, Mosen et Frank. De plus, je veux souligner l’aide que Patrick Gendron m’a apportée avec le serveur. Élaine Meunier, la coordonnatrice des étudiants de bio-informatique, alias notre 2e maman, m’a aussi beaucoup aidé. À vous tous, merci beaucoup!

# 1 Introduction

Comprendre la structure des ARN est essentiel à l’élucidation de son rôle dans la cellule. Les ARN agissent à plusieurs niveaux pour réguler l’expression génique. Ces molécules portent le message des gènes; elles peuvent réguler d’autres ARN par leurs représentants les plus petits, les micro-ARN et elles ont aussi la capacité de sentir leur environnement et ainsi moduler leurs fonctions de façon autonomes par des motifs nommés *riboswitch*.

Pour mieux saisir les subtilités des structures des ARN, une méthode a été développée : le sondage chimique des ARN. Parmi plusieurs méthodes de sondage chimique, j’ai choisi d’étudier la méthode nommée SHAPE. Au début des années 2010, des chercheurs ont combiné cette technique avec le séquençage haut débit. Le résultat fut nommé : SHAPE-seq, une méthode qui génère beaucoup de données. Depuis 2014, le laboratoire du Dr Das à Stanford publie régulièrement des milliers d’ARN avec une valeur de réactivité et d’erreur pour presque tous les nucléotides des séquences sondées.

Dans mon projet de recherche, j’ai analysé ces données dans le but de pouvoir prédire la réactivité des nucléotides et pour mieux étudier les structures secondaires (SS) produites par des logiciels de prédiction de SS. La prédiction de la réactivité est une valeur de cohérence entre la réactivité du motif cyclique nucléotidiques (MCN) observée et celles d’une base de données.

Pour ce faire, il m’a fallu rassembler les connaissances sur les ARN sondés, visualiser et vérifier les prédictions. J’ai comparé mon approche à celle basée sur l’état des nucléotides (pairé ou non pairé).

## 1.1 Domaine de recherche

Les méthodes utilisées dans mon projet de recherches touchent plusieurs champs d’expertise. Des connaissances en physique, chimie et biochimie sont essentielles à la compréhension du repliement des ARN. Pour faire l’analyse, il m’a fallu créer des algorithmes et bâtir des bases de données performantes. Une solide base en informatique a donc été nécessaire. Finalement, j’ai dû faire preuve de curiosité et de persévérance tout au long de l’écriture de ce mémoire.

L’univers de la biologie moléculaire est très vaste et c’est pourquoi pendant toute ma maîtrise, j’ai tenté de garder les choses simples. J’ai voulu combiner plusieurs approches en un tout cohérent et facile d’utilisation.

Les méthodes à haut débit comme SHAPE-seq permettent de tester beaucoup d’hypothèses et les logiciels libres comme MCFlashfold et RNAsubopt peuvent être amélioré et utilisé de plusieurs façons. Ces fonctionnalités en font des outils scientifiques vraiment précieux.

## 1.2 Mise en contexte

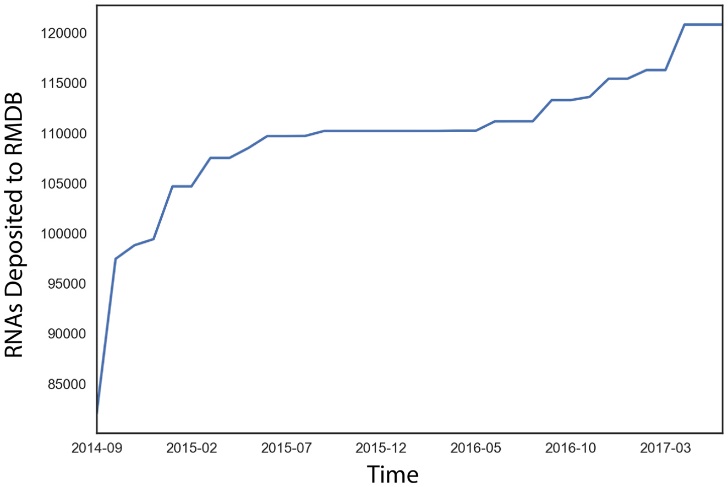
Ce travail a été effectué dans le laboratoire du Dr François Major à l’institut de recherche en immunologie et cancérologie. La bio-informatique moléculaire est une science récente et beaucoup de choses restent à faire. Actuellement, nous avons des outils informatiques pour représenter les ARN à plusieurs niveaux d’abstraction. Le niveau tout-atome est l’un des plus demandant en calcul, mais l’un des plus précis. Une simulation par dynamique moléculaire de quelques microsecondes ne peut considérer que des structures de la grosseur d’une boucle de 4 nucléotides en ce moment, mais cela risque de changer rapidement. Une autre limite de cette approche est que nul ne sait modéliser les interactions entre les ions magnésium et l’ARN. Il est bien connu que ces interactions sont essentielles au repliement de l’ARN [1]. La demande en calcul qu’entraine cette simulation empêche les modélisations d’être faites en temps réel, un changement de paramètre entraine un recalcul complet occupant un ordinateur pour plusieurs heures, voire des jours.

Le modèle que je propose dans ce mémoire permet de visualiser plusieurs structures secondaires (SS) d’un ARN dans la même vue et donne une idée de la vraisemblance de chacune d'entre elles. Il offre un compromis entre la résolution, le temps de calcul et l’espace mémoire nécessaire. Son utilité principale est de permettre la création d’hypothèses menant à améliorer les prédictions des SS faites à partir d'une séquence sondée par le 1M7.

Dans la littérature, l’évaluation d’une SS par le 1M7 est basée sur la flexibilité des nucléotides, dérivée de leur état pairé ou non pairé [2, 3].

Avant 2011, la date de la sortie du protocole de SHAPE-seq [4], vu le manque de données normalisées de sondage chimique, il était difficile d’utiliser des algorithmes d’apprentissage machines pour obtenir des informations sur la réactivité des nucléotides. En 2014, le nombre de séquences sondées rendues publiques a explosé, grâce au laboratoire du Dr Das [5].

Nombre d’ARN dans la RMDB en fonction du temps



ARN déposés dans la RMDB

Temps

1. **Nombre de séquences d’ARN ajoutées à la RMDB entre septembre 2014 et mars 2017**. Figure modifiée de l’article : *Updates to the RNA mapping database (RMDB), version 2.* Nucleic Acids Research, 2017.

Mon approche est basée sur des sous-structures de la SS. Elle considère la majorité des prédictions comme étant bonnes, ce qui pose un défi de sélection des séquences. J’ai récolté les niveaux de réactivités chimiques d’un grand ensemble de données, c’est-à-dire, plusieurs millions de nucléotides provenant de plusieurs milliers d’ARN, je les ai compilées, pour ensuite les utiliser pour prédire la réactivité des nucléotides provenant d’ARN non utilisés lors de l’apprentissage.

En plus de donner un ordre d’exactitude des prédictions des SS, les deux logiciels que j’ai créés permettent d’étudier l’ARN dans le contexte de la réactivité chimique de façon générale. Ils sont disponibles librement sur *github* dans mon répertoire et celui du laboratoire du Dr Major.

## 1.3 Historique

### 1.3.1 Début de l’étude des acides nucléiques

En 1869, Friedrich Miescher, découvre l’ADN[6]. En 1939, Caspersson et Schultz utilisent un spectromètre pour établir qu’il y a des acides nucléiques dans le cytoplasme[7].

### 1.3.2 Détermination de la structure de l’ADN et de l’ARN

En 1953, Francis Crick et James Watson élucident, avec l’aide des données de Rosalind Franklin, un des mystères les plus grands du milieu du 20e siècle, la structure de l’acide désoxyribonucléique (ADN)[8]. L’acide ribonucléique (ARN) est un acide nucléique tout comme l’ADN et sa structure lui est apparentée. En 1961, Monod et Jacob proposent que l’ARN soit un intermédiaire entre l’ADN et les protéines, ils auront raison[9]. Les mécanismes par lesquels les ARN jouent leurs rôles s’imposent alors comme questions de recherches. En 1964, Robert W. Holley se sert d’une enzyme pour digérer un ARN de transfert et c’est ainsi qu’il découvre sa structure secondaire [10]. C’est le début du sondage enzymatique de l’ARN.

### 1.3.3 Prédiction de la structure secondaire des ARN

L’idée de Tinocco, en 1971, est de minimiser l’énergie de ces molécules pour trouver la structure la plus stable. Des règles simples sont appliquées à une structure pour évaluer son énergie. Elles comprennent : une énergie d’initiation des hélices, une énergie de propagation et une énergie pour les boucles[11]. En 1978, Nussinov propose un algorithme rapide pour déterminer la structure secondaire d’un ARN. Son nombre d’étapes grossit proportionnellement au cube de la longueur d’une séquence de nucléotides, autrement dit, il est dans l’ordre de complexité de O(n3), « n » étant le nombre de nucléotides de la séquence [12, 13]. Ensuite, il y a eu les algorithmes donnant les SS sous-optimales en 1984 et 1989 [14, 15]. En 1990, McCaskill imagine un algorithme pour obtenir la fonction de partition en O(n3) [16]. Cette fonction est utile pour connaître la probabilité de formation d’une paire de bases. En 1991, MC-SYM, un logiciel qui prédit la structure 3D d’une molécule est publié par le Dr François Major, mon directeur de recherche[17]. En 1994, ViennaRNA package voit le jour [18]. Par la suite, l’intérêt pour l’ARN n’a fait qu’augmenter. En 1995, Turner publie un ensemble de données thermodynamiques sur les paires de bases et les boucles de l’ARN [19]. Ces données seront utiles pour des algorithmes tels que les proches voisins. Une autre date marquante pour ce domaine est l’année 2008, MC-Fold une version lente de MCFlashfold est publié par le laboratoire du Dr Major [20].

## 1.4 Pourquoi étudier la structure de l’ARN

Le repliement de l’ARN est intrinsèquement lié à ses nombreuses fonctions. On peut penser aux ARN de transfert, dont la forme tridimensionnelle, en « L », est nécessaire pour qu'ils soient reconnus par le ribosome. Dans une cellule vivante, chaque ARN naît de la transcription d’un brin d’ADN. Ensuite, chaque molécule se replie pour adopter une ou plusieurs formes plus ou moins stables pouvant se transformer une en l'autre sous certaines conditions. Ces molécules sont impliquées dans la majorité des processus de rétroaction cellulaires. Que ce soit au niveau de leur traduction en protéine ou directement. Les *riboswitchs* sont un bon exemple de régulation directe [21]. De façon indirecte, les micro-ARN (miARN) peuvent contrôle l’abondance de certaines protéines, en réprimant des ARN messagers (ARNm) [22]. L'effet éponge de certains ARN régule le niveau des miARN, le laboratoire du Dr Major est un pionnier dans le calcul des interactions des miARN in vivo [23].

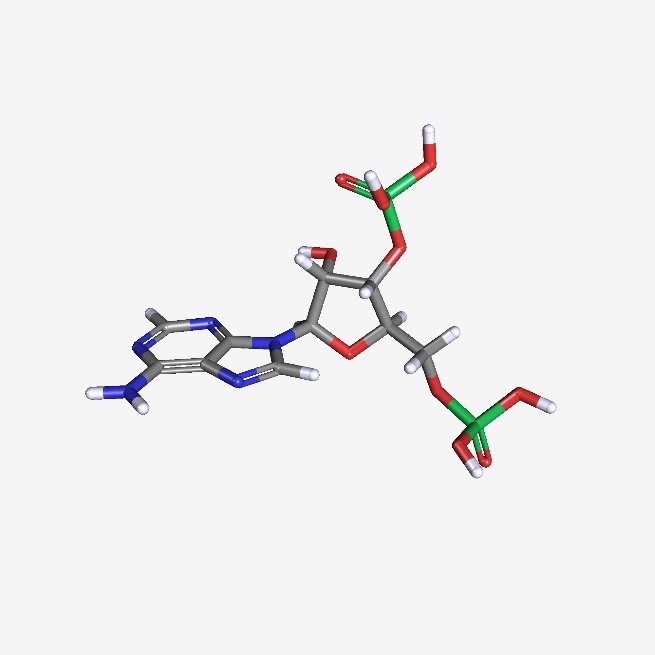
De plus, un problème complexe comme celui abordé dans ce mémoire offre l’occasion d’innover en matière d’algorithme d’apprentissage machine. La possibilité d’analyser des données publiques permet à la communauté scientifique de se rassembler et d’échanger sur les meilleures pratiques.

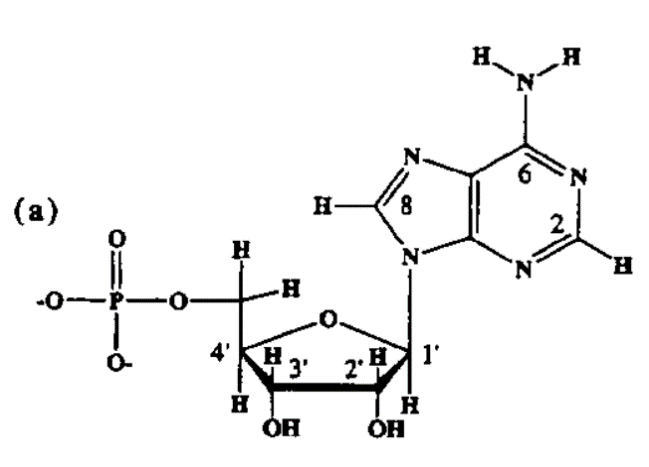
## 1.5 Définition des termes et des concepts utilisé pour prédire la réactivité des nucléotides d’un ARN.

### 1.5.1 Les nucléotides

D’un point de vue chimique, l’ARN est un polymère de ribonucléotides. Il est composé dans la nature de 4 ribonucléotides communs et de plusieurs ribonucléotides modifiés. Les 4 bases communes de l’ARN sont l’adénine, la cytosine, l’uracile et la guanine. L’image en A de la figure 2 provient de [24] et l’image en B de la même figure provient de [25].

Un nucléotide est toujours composé de 3 parties, voir figure 2 :

1. Le groupement phosphate
2. Le ribose
3. La base azotée



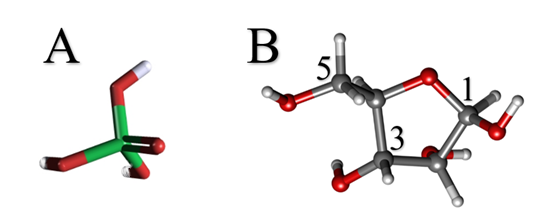
B

A

1. **A) Représentation d’un nucléotide, l’adénosine monophosphate (AMP)**. Le cercle rouge, le pentagone bleu et le carré vert représentent le groupement phosphate, le ribose et la base azotée, respectivement. Cette image a été modifié de l’article : *Structural analysis of nucleic acid aptamers* *.* Current opinion in chemical biology, 1997. **1**: p. 32-46, écrit par Patel, D.J. B) image 3D de A, avec en plus un groupe phosphate sur le carbone en 3’ du ribose. Source : PubChem3D.

#### 1.5.1.1 Le groupement phosphate

En ne comptant pas les atomes d’hydrogène, un groupement phosphate est composé de 5 atomes. 4 atomes d’oxygène et un de phosphore. Chaque groupement phosphate donne une charge négative à l’ARN.



1. **Représentation en balles et bâtons (*balls and sticks*) d’un groupement phosphate** **et** **d’un beta-D-ribose**. Le phosphore est vert, les oxygènes sont rouges, les carbones gris et les hydrogènes blancs **A)** L’atome de phosphore est au centre des atomes d’oxygène. Les hydrogènes en périférie. Source : PubChem3D. **B)** Dans le polymère qu’est l’ARN, les groupes phosphates s’attachent aux carbones 3 et 5 du ribose. Le carbone 1, en beta, est lié à la base azotée. Cette image est disponible sur <https://commons.wikimedia.org/>, elle a été produite par, Marina Vladivostok, les positions (1, 3, 5) des atomes ont été ajoutés.

#### 1.5.1.2 Le ribose

Le ribose est un pentose, un sucre composé de 20 atomes dans sa forme libre. Il a toujours 5 atomes de carbone. À l’état libre (pas dans une chaine), il a 5 atomes d’oxygène et lorsqu’il est dans l’ARN, avec sa base et ses deux phosphates, il a 2 atomes d’oxygène complètement à lui, 2 oxygènes qu’il partage avec les deux groupes phosphate liés au carbone 3’ et 5’ et un atome d’oxygène qu’il partage avec sa base. Dans la figure ci-dessous, on observe un beta-D-ribose. Son image miroir, la forme « L » n’est pas présente dans la nature.

#### 1.5.1.3 Les bases azotées

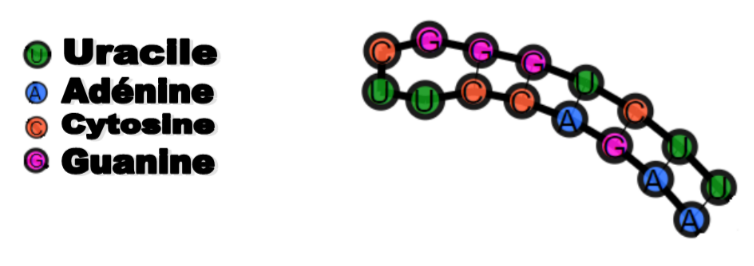
Les images suivantes viennent de Wikipédia, les atomes bleu, rouge, gris et blanc sont des azotes, des oxygènes, des carbones et des hydrogènes respectivement. La flèche rouge est le point d’attache du ribose. Comme mentionné plus haut, les ribonucléotides ont 4 représentants communs :

|  |  |
| --- | --- |
| 1. L’adénine | 1. La guanine |
| C:\Users\p\Pictures\Adenine-3D-balls.png | C:\Users\p\Pictures\Guanine-3D-balls.png |
| 1. La cytosine | 1. L’uracile |
| C:\Users\p\Pictures\Cytosine-3D-balls.png | C:\Users\p\Pictures\800px-Uracil-3D-balls.png |

1. **A)** L’adénine est une base azotée de la famille des purines. Son symbole est le A. **B)** La guanine est aussi une base azotée de la famille des purines. Son symbole est le G. **C)** La cytosine est une base azotée de la famille des pyrimidines. Son symbole est le C. **D)** Tout comme la cytosine, l’uracile est une base azotée de la famille des pyrimidines. Son symbole est le U.

### 1.5.2 La structure secondaire

La structure secondaire (SS) d’un ARN est formée par l’ensemble des paires de bases d’un ARN. Ce concept est utile pour comprendre et expliquer, la forme des ARN. Un des composants majeurs des ARN est la double hélice. La différence des doubles hélices des ARN par rapport à celles de l’ADN de nos chromosomes est que la plupart du temps les SS des ARN sont formées par deux brins de la même molécule.



1. Exemple de visualisation d’une structure secondaire générée par le logiciel *RNA dynamique viewer*. (Voir chapitre 2)

En général, on utilise un sous-ensemble des paires de bases répondant à deux critères, soit : l’absence de liens entre plus de 2 bases (sans compter les nucléotides adjacents liés de façon covalente) et l’absence de pseudo-nœuds. Les pseudo-nœuds sont des structures de la forme : ux1ux2uy1uy2u, où les « x » et « y » sont des suites de nucléotides pairés : les 1 (en exposant) avec les 2 (en exposant) et les « u » sont des nucléotides non pairés ou rien du tout [26]). Avec ces contraintes, le problème de la détermination de la structure secondaire peut être résolu avec la programmation dynamique en O(n3), sans ces contraintes le problème peut croite de façon exponentielle en fonction de la longueur de l’ARN. Il faut garder en tête que la majorité des liens entre les nucléotides sont compatibles avec ces critères, mais que les SS ainsi créés ne peuvent pas expliquer l’entièreté des phénomènes de réactivité chimique. La figure 6 est un exemple de tige suivi d’une boucle. De façon générale, la structure secondaire est l’ajout des paires de base à une séquence de nucléotides.

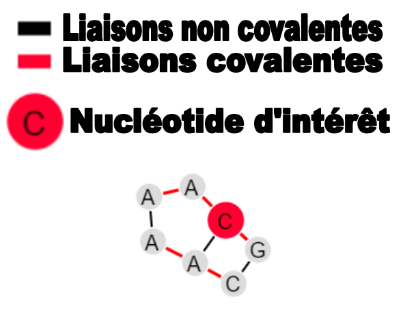
### 1.5.3 Les cycles minimale nucléotidique

Pour étudier la réactivité des nucléotides, il est utile de sous diviser l’ARN. Un cycles minimale nucléotidique (CMN) est une sous-structures de la SS des ARN. Un ARN, dans sa forme finale, est composé de boucles, de double hélice, de jonctions et de motifs tertiaires. La définition des CMN s’appuie sur les paires de bases. Par exemple, dans ce mémoire, on désigne par « 2\_2 », le CMN le plus fréquent. Il a deux paires de bases adjacentes et forment un CMN. Le tableau I présente 2 CMN, le « 2\_2 » est le composant principal des doubles hélices et le « 4\_4 » est un renflement interne à une double hélice.

1. Représentation de deux CMN

|  |  |  |
| --- | --- | --- |
| MCN | 2\_2 | 4\_4 |
| Figure |  |  |
| Identifiant | 2\_2\_GG\_CC | 4\_4\_GAAC-GAAC |

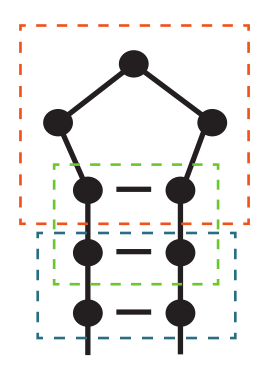
Un CMN « 3\_2 » est composé d’un nucléotide non pairé au centre de deux nucléotides pairés à deux nucléotides adjacents sur l’autre brin d’ARN, on appelle aussi cette structure : un renflement d’un nucléotide.

****

1. Sous-structure formée de deux MCN collés. L’identifiant de cette sous-structure est : « 3\_2-AAC-AA\_pos\_2&2\_2-CG-CA\_pos\_0 ». Le nucléotide d’intérêt est le premier nucléotide du NCM « 3\_2 », sont identifiant permet de comparer les nucléotides se trouvant dans la même sous-structure dans d’autres ARN.

La figure 7 illustre le chevauchement des nucléotides. Dans ce mémoire, pour caractériser un nucléotides deux types de sous structures ont été utilisé :

1. L’appartenance à un MCN peut importe l’état pairé ou non d’un nucléotide. C’est-à-dire, lorsque le nucléotide est pairé, deux entrées sont ajoutées dans la base de données (une pour le MCN en 5’ et l’autre pour le MCN en 3’).
2. Une sous-structure englobant le nucléotide. C’est-à-dire qu’il y a invariablement une entrée peut importe l’état pairé ou non du nucléotide.



1. Représentation de trois CMN. Ici, on voit que les CMN se chevauchent. La boite bleue et la boite verte entoure des CMN « 2\_2 ». La boite rouge entoure un CMN à 5 nucléotides, un « 5 ». Lorsqu’on s’intéresse à un nucléotide en particulier, les nucléotides pairés se trouve à la jonction de deux MCN tandis que les nucléotides non pairés font partie d’un seul MCN. Cette figure est tirée du manuel d’instruction de MCFlashfold (2015, Dallaire, P. and F. Major, *MCFlashfold (mcff version 34) user manual).* Les boites ont été ajoutées.

Dans ce mémoire, une chaine de caractère identifie les sous-structure que sont les CMN. Un « diagramme à rail » des expressions régulières représentant les CMN est placé en annexe. Une description textuelle est faite maintenant.

### 1.5.4 L’identifiant unique d’un nucléotide à l’intérieur d’une sous-structure

L’identifiant d’un nucléotide à l’intérieur d’une sous-structure est composé d’un ou deux « CMN augmenté », lorsqu’il y en a deux, un « & » les séparent. Chaque « CMN augmenté » est formé d’un type, d’une ou deux séquences et de la position du nucléotide d’intérêt.

#### 1.5.3.1 Le type

Dans tous les cas, la chaine de caractère débute toujours par le type du CMN (« 2\_2 », « 3\_2 », « 4\_4 », etc.).

Lorsque le CMN a 2 brin, il est délimité par 2 paires de bases (deux liaisons non covalentes), le type est composé de deux chiffres représentant le nombre de nucléotides entre les deux paires de bases plus les nucléotides des paires de bases. Il débute par le brin du nucléotide d'intérêt.

Les boucles sont traitées de la même manière, mais elles ont un seul brin. Précisément, pour les boucles, le type du CMN a seulement un chiffre. Ce chiffre est le nombre de nucléotides libres de la boucle plus les deux derniers nucléotides de l’hélice.

Seul les CMN ayant moins de sept nucléotides sur chaque brin ont un chiffre, les autres ont un « L ». Deux raisons expliquent ce choix d’implémentation.

1. Le nombre de représentant de chaque CMN de plus de 7 nucléotides par brin est bas.
2. L’espace de la base de données est précieuse, ses performances dépendent du nombre d’entrées différentes.

#### 1.5.3.2 La séquence

Pour les CMN à deux brins, le type est suivi par les séquences, en commençant, aussi, par le brin d'où provient le nucléotide d'intérêt. Les séquences sont de la longueur des chiffres du type ou si elles sont égales ou plus grandes à sept, elles sont remplacées par la lettre « L ».

Pour les boucles, la séquence débute par le nucléotide pairé en 5’ et se termine par le nucléotide pairé en 3’. Une fois encore, lorsque le nombre de nucléotides est égal ou supérieur à sept, la lettre « L » remplace la séquence.

#### 1.5.3.3 La position du nucléotide

Finalement, on ajoute à l’identifiant la position du nucléotide. Pour les CMN composés de deux brins le nt. se trouve sur le premier brin. Ici il faut noter qu’on ne prend pas en compte le coté du brin sur lequel se trouve le nt. (5’ ou 3’) lors de leur extraction de l’ARN. Pour les boucles, il n’y a pas d’ambiguïté à ce niveau.

Dans les deux cas (deux brins ou boucles), le premier nucléotide en 5’ à comme numéro de position le 0 et les autres suivent. Le dernier nt. a donc la position « n » - 1, « n » étant le nombre de nucléotide du brin.

#### 1.5.3.4 Appartenance d’un nucléotide à un MCN

Les nucléotides pairés font partie de 2 CMN, tandis que les nucléotides non pairés font partie d’un seul CMN par SS. Les nucléotides de début ou de fin de séquences pour lesquels aucun voisin n’est pairé d’un des deux côtés, ne font pas partie d’un CMN à proprement parler, mais il est utile d’identifier ces caractéristiques pour prédire la réactivité des nucléotides. L’algorithme qui permet d’extraire les CMN d’une SS est expliqué dans le chapitre 1.

### 1.5.4 La prédiction des structures secondaire

Pour comprendre et prédire la réactivité des nt. à partir d’une séquence de nucléotide, il faut avoir une idée de sa structure. Plusieurs méthodes ont été développées pour prédire la ou les structures secondaires d’un ARN à partir de sa séquence. Dans ce mémoire, j’utilise deux d’entre elles : la méthode des proches voisins et celle de l’échantillonnage des structures connues.

#### 1.5.4.1 Les proches voisins (RNAsubopt)

Une des deux méthodes de repliement des ARN en SS utilisé dans ce projet de recherche est la méthode des proches voisins.

Cette méthode est nommée « nearest neighbour » en anglais et a été développée par Nussinov et coll. en 1978 [15]. Zucker et Stiegler l’ont implémenté et placée dans un programme nommé mfold par la suite [27].

Dans cet algorithme, les duos de paires de bases adjacentes, les boucles internes, les boucles de fin d’hélice et les renflements obtiennent une énergie qui contribue à l’énergie totale de la molécule. L’idée, tout comme dans beaucoup d’autres algorithmes, est de former la structure secondaire ayant la plus petite énergie possible. Cette méthode a été optimisée et est très utilisée pour déterminer la température de fusion (ou de fonte) de deux acides nucléiques (ADN ou ARN) lors d’une PCR. J’utilise dans ce mémoire un logiciel qui produit en sortie un ensemble de SS qualifiées de sous-optimales, il se nomme : *RNAsubopt* [30].

#### 1.5.4.2 Échantillonnage à partir des structure 3D connues (MCFlashfold)

La deuxième méthode de repliement des ARN utilisés est une méthode développée dans le laboratoire du Dr. Major.

Elle consiste à échantillonner les MCN d’un ensemble d’ARN dont les structures sont connues. L’ARN est transformé en un réseau, où les nœuds sont des nucléotides et les liens sont soit des paires de bases ou des liens covalents médiés par un groupement phosphate entre les sucres des nucléotides.

Ce réseau est transformé en plus petits motifs. La fréquence d’observation d’un motif donné est convertie en probabilité et ces probabilités sont par la suite utilisées pour prédire les SS. La formule complète est offerte dans le matériel supplémentaire de l’article MC-Fold / MC-SYM pipeline [20].

MCFlashfold, la version utilisée, est une amélioration de MC-fold, créé par un ancien étudiant du Dr Major, Dr Paul Dallaire. MC-fold a été créé par Dr Marc Parisien.

### 1.5.6 Le sondage chimique de l’ARN

Une des méthodes fructueuses dans l’étude des structures des ARN est le sondage chimique des ARN. Les protocoles de sondages chimiques existent depuis longtemps. Ils ont été optimisés et ils produisent des résultats abondants grâce aux technologies de séquençages de nouvelle génération.

Le principe est le suivant. L’ARN est synthétisé en laboratoire. Il est placé en contact avec un agent chimique qui le modifie de façon spécifique. La flexibilité de l’ARN au niveau du nucléotide et l’accessibilité de l’agent chimique aux nucléotides est lié à sa réactivité [31]. Dans ce mémoire, une base placée dans une conformations précise décrites par sa sous-structures permet de prédire la réactivité du nucléotide.

RNAstructure, un logiciel de détermination de la SS des ARN, incorpore les données de réactivité chimiques en appliquant des contraintes à la SS. Selon cette méthode, si le nucléotide réagit, il ne peut être pairé à moins d’être dans une paire G-U ou de suivre une paire G-U [32]. Pour Turner, Zucker et Mathews, en plus des règles déjà décrites, les nucléotides peuvent être dans une paire A-U et G-C à la fin d’une hélice [33]. D’autres règles existent : Stadler ajoute des bonus d’énergie aux nucléotides réactifs, au lieu de déterminer complètement leur état (pairé ou non). Ces contraintes sont appelées des contraintes souples, par opposition aux contraintes dures citées précédemment [34].

À l’inverse, les nucléotides peu réactifs sont inaccessibles à l’agent modificateur. Cette inaccessibilité est due soit au pairage du nucléotide soit à la structure tertiaire de l’ARN, soit à une conformation locale précise. (Ces conformations sont décrites dans ce mémoire par les « CMN augmenté » ayant une réactivité basse.)

Le réactif et les conditions des expériences sont déterminants dans les résultats finaux. Dans ce mémoire, seul le réactif nommé « 1m7 » a été étudié. Toutes les données proviennent de la RMDB et sont associées à l’ensemble de donnée « ETERNA ».

Le nombre de modifications par ARN est contrôlé par la concentration du réactif et par son temps de contact avec l’ARN. Les protocoles les plus récents prennent en compte le décrochage de l’ARN en soustrayant les valeurs obtenues par un contrôle négatif. Dans le but de prendre en compte l’erreur due à un nucléotide très réactif, plusieurs sondages à différentes concentrations du modificateur sont nécessaires, ce phénomène se nomme en anglais : « *over modification* » [35]. Ensuite une enzyme, la rétrotranscriptase transcrit l’ARN en ADN. Lors de cette étape, la rétrotranscriptase s’arrête aux bases modifiées. À l’aide des technologies de séquençages, on obtient un décompte de toutes les sous-séquences de la séquence. Le logiciel « Mapseeker » établit ensuite une correspondance entre le nombre des molécules d'une certaine séquence et le lieu où l’ajout s'est produit [36]. Beaucoup de détails sur la standardisation de ces expériences sont disponibles dans le matériel supplémentaire de l’article [35].

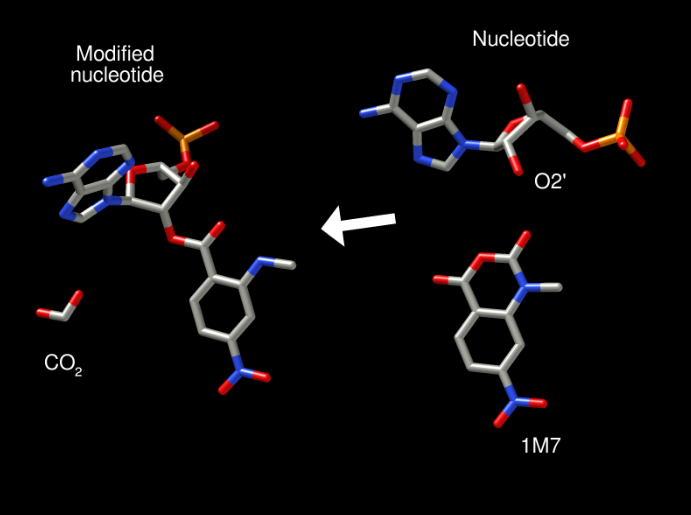
### 1.5.7 Agents modificateurs

#### 1.5.7.1 SHAPE-seq

Le laboratoire du Dr weeks à developpé un méthodes nommée SHAPE qui est l’acronyme de « *Selective 2′ Hydroxyl Acylation analyzed by Primer Extension* ». Cette méthode est composée de 3 réactifs : le « 1-methyl-7-nitroisatoic anhydride » (1m7), le « 1-ethyl-6-nitroisatoic anhydride » (1m6) et le « N-methyl isatoic anhydride » (NMIA). Par la suite, le laboratoire du Dr Lucks a développé une autre méthode basée sur SHAPE qui a permis d’augmenter le nombre d’ARN sondés (Doudna JA, Arkin AP.). Le Dr Das et ses collègues ont perfectionné cette méthode et ils ont créé une base de données pour obtenir les ARN et leurs valeurs de réactivité. Cette base de données se nomme : la RMDB, pour « RNA mapping database ». Comme il a été mentionné plus haut, les ARN analysés dans ce mémoire ont été sondés avec le 1M7.

#### 1.5.7.2 1M7

La figure 7 montre les réactifs et les produits de la réaction chimique entre le 1M7 et une adénine. Une molécule de dioxyde de carbone est créée lors de la réaction. C’est le groupement hydroxyle du carbone 2 qui réagit avec le 1M7. La sélectivité et le mécanisme de cette réaction ne sont pas complètement compris [37]. Récemment, une équipe à montrer à l’aide de la dynamique moléculaire que la conformation locale du nucléotide et spécialement de son sucre est importante dans ce mécanisme [38].



Nucléotide modifié

Nucléotide

1. **Réaction chimique du 1M7 avec le 2’OH du sucre d’un nucléotide.** Cette image provient du wiki d’Eterna, un jeu vidéo consacré à la structure des ARN : <http://eternawiki.org/wiki/index.php5/1M7>. Elle a été créée par Omei.

# Chapitre 1  RNASS\_v2 : Obtenir la réactivité des CMN

RNASS\_v2 est un script écrit en *python*. Il est spécialement conçu, pour faire le lien entre les données de sondages chimique de l’ARN et la structure secondaire produit par les logiciels *MCFlashfold* et *RNAsubopt*. Ce script prend en entrée une liste d’ARN. Il produit en sortie un fichier JSON. Ces fichiers sont ensuite utilisés pour peupler une base de données qui sert à prédire la réactivité des nucléotides. Dans ce chapitre, je décris les étapes de son fonctionnement. Il a deux modes de fonctionnement :

1. le mode apprentissage
2. le mode prédiction

## De l’obtention des données à la prédiction discrète

### Extraction des données

La procédure pour obtenir expérimentalement les données de la RMDB et les traitements numériques subséquents sont bien expliqués dans la documentation [35, 36] et sur le forum d’« ETERNA » [49]. Pour les données d’« ETERNA », une interface de programmation (API) a été mise en place pour que la communauté puisse les obtenir facilement. Il suffit de trouver la liste des « laboratoires » et de faire une requête à l’API (voir l’annexe). Les « laboratoires » sont des ensembles de données créés dans un but précis, ce but peut être l’étude d’une boucle, d’un renflement ou de toutes autres caractéristiques. Plusieurs informations sont disponibles sur les « laboratoires », nous retrouvons entre autres : leurs auteurs, leur identifiant numérique, la date de leur création et leur titre. J’ai choisi de me limiter aux données produites entre 2014 et 2018. Un script nommé : eternaTSV.py utilise cet API et produit une liste d’ARN avec leur séquence, les données de réactivité chimique et d’autres informations reliées à l’expérience qui les a créées.

Il est aussi possible de télécharger les fichiers RDAT de la base de données de RMDB. Ces fichiers viennent avec un script *python* qui se nomme *handler.py*. Chose surprenante le vecteur de réactivité n’est pas de la même longueur que la séquence elle-même. On doit phaser le vecteur de réactivité avec un attribut nommé : « offset ». Ceci s’explique par le fait que l’ARN possède une séquence qui permet sont identification lors du séquençage. Aucune donnée de réactivité ne peut être extraite de cette séquence.

Nous pouvons considérer ces données comme fiables, car elles ont été normalisées à l’aide d’un contrôle interne à la séquence et chaque séquence à un contrôle non modifié par le réactif. De plus, chaque ARN est sondé au minimum dans deux solutions, une 10 fois plus concentrée que l’autre [35].

### Retirer les ARN difficiles à prédire pour mieux apprendre

Après la première analyse de mon prédicteur, les ARN qui avait un score de prédiction moyen bas était systématiquement des ARN ayant un contenu en adénine ou un score de réactivité moyen élevé. J’ai donc décidé de les retirer de l’analyse finale.

Seul les ARN composés de moins de 50% d’adénine ont été gardé. Plusieurs scores de réactivité moyenne ont été testé. De plus, une valeur nommée « signal to noise » et la diversité d’ensemble des SS des ARN ont aussi servi à augmenter la fiabilité des prédictions.

### Repliement des ARN

Seuls les logiciels MCFlashfold et RNAsubopt ont été utilisés pour replier les ARN. RNAsubopt est un logiciel provenant de la suite Vienna package et MCFlashfold est un logiciel qui a été créé dans le laboratoire du Dr Major. Ces deux logiciels ont pour particularité de calculer un ensemble de SS auxquels ils attribuent une valeur d’énergie. Le nombre de SS calculée peut être modifié directement ou en fonction d’un écart d’énergie entre la SS de minimum d’énergie libre (minimum free energie (MFE)) et les autres SS. Vu la complexité qu’amène le calcul de chaque SS supplémentaire en raison du nombre de prédictions de réactivité chimique à faire (une par nucléotide de chaque SS), RNASS en considère relativement peu. Plusieurs valeurs ont été testées, toutes plus petites que cent. Dans la majorité des analyses qui suivent. Dix SS sous-optimales sont utilisées.

Un troisième logiciel, RNAfold, est utilisé pour obtenir la diversité d’ensemble et la fréquence de la SS d’énergie minimale (MFE) dans les SS prédites.

Toutes ces données sont placées dans un fichier JSON standardisé. La prochaine section décrit quelques-uns de ces champs.

### Champ des fichiers JSON standardisés

Les deux prochaines pages présentent la liste des champs des fichiers JSON de sorties. Chaque champ du tableau II est présent dans l’ensemble des séquences. Le champ est le nom de la variable. La structure du fichier est arborescente (un champ peut contenir plusieurs champs).

1. Description des champs des fichiers JSON standardisés des ARN

|  |  |  |
| --- | --- | --- |
| Champ | seq | frequenceNT |
| Explication | La séquence de l’ARN. | La fréquence relative des nucléotides. |
| Champ | ScoresMoy | graph\_transition |
| Explication | La moyenne des scores. | Les valeurs des énergies des SS et les liens de transitions entre les structures. |
| Champ | scoreTab | nts |
| Explication | Le vecteur de scores. | Les détails sur les nucléotides\*. |
| Champ | erreurTab | reactivityVector |
| Explication | Le vecteur d'erreurs. | Le vecteur de réactivité discret. |
| Champ | ed | sc\_mcff / sc\_so |
| Explication | La diversité d'ensemble. | Le score de prédiction pour chaque SS. |
| Champ | freqMfe | info |
| Explication | La fréquence de la mfe dans l'ensemble (RNAfold) | Information sur les paramètres de RNASS |

\*Voir le tableau de la prochaine page

Le tableau II présente les champs des nucléotides, un des champs des ARN.

1. Description du champs *nts* des fichiers JSON standardisés

|  |  |  |
| --- | --- | --- |
| Champ | score | position |
| Explication | La valeur de réactivité. | La position dans la séquence. |
| Champ | erreur | region |
| Explication | La valeur de l’erreur. | Le nucléotide est-il parmi les 5% premiers ou derniers nucléotides? |
| Champ | etat | rna\_id |
| Explication | La fréquence pairée par logiciel. | L’identifiant de l’ARN dans lequel, le nucléotide se trouve. |
| Champ | inHelice | seq\_motif3 / 5 |
| Explication | Le nucléotide est-il dans une hélice de trois paires de base et plus? | Le motif de séquence dans lequel le nucléotide se trouve. (3/5) |
| Champ | localNcmD | sc\_mcff / sc\_so |
| Explication | La fréquence du nucléotide dans chaque MCN. | Le score de prédiction du nucléotide dans toutes les SS. |
| Champ | mfp\_mcff / mfp\_so | sorte |
| Explication | Le partenaire le plus fréquent. | A, C, G ou U |

## Information sur les données collectées

Dans la prochaine section, j’ai fait ressortir les statistiques pertinentes de quelques champs des fichiers JSON présentés plus haut. J’ai bâti deux bases de données successivement.

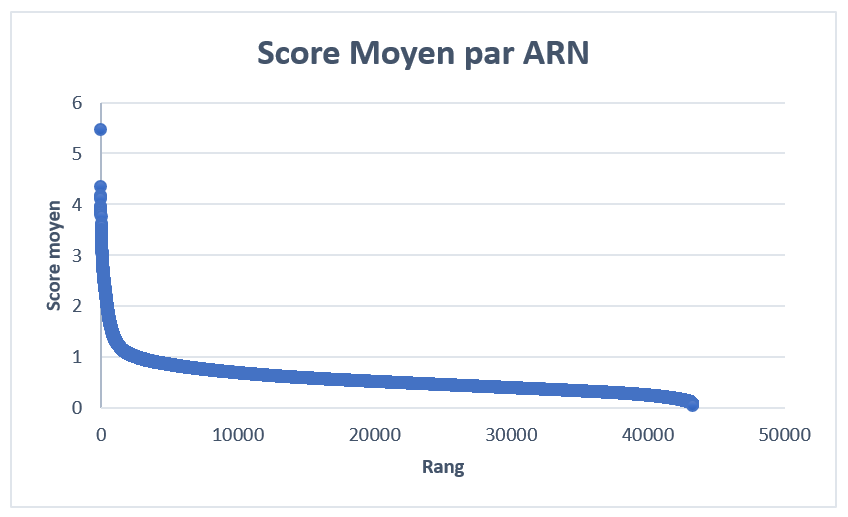
Dans la première que je nomme « non filtrée », plus de 43 000 ARN provenant de la RMDB ont servi pour ces statistiques, ce qui totalise un peu moins de 4 millions de nucléotides sondés. Pour cette base de données, la seule condition est que les ARN doivent avoir un rapport « signal sur bruit » supérieur à 1 pour être considérés.

La deuxième que je nomme filtrée contient plus de 16 000 ARN. Comme expliqué plus haut, le ratio d’adénine sur le nombre de nucléotides total doit être inférieur à 50% et la moyenne des scores plus petite que 1,5.

Les nucléotides analysés dans ce mémoire ont été sondés par l’agent chimique 1M7 à 24 °C dans 10 mM de MgCl et 50mM d’HEPES (pH8.0).

### Scores moyens des valeurs de réactivité

Comme vous pouvez le constater dans la figure 10, la majorité des moyennes des valeurs de réactivité se trouve sous la barre des 1. Cette analyse a été faite avant l’application du filtre sur la moyenne des scores de réactivité par ARN et le contenu relatif en adénosine.



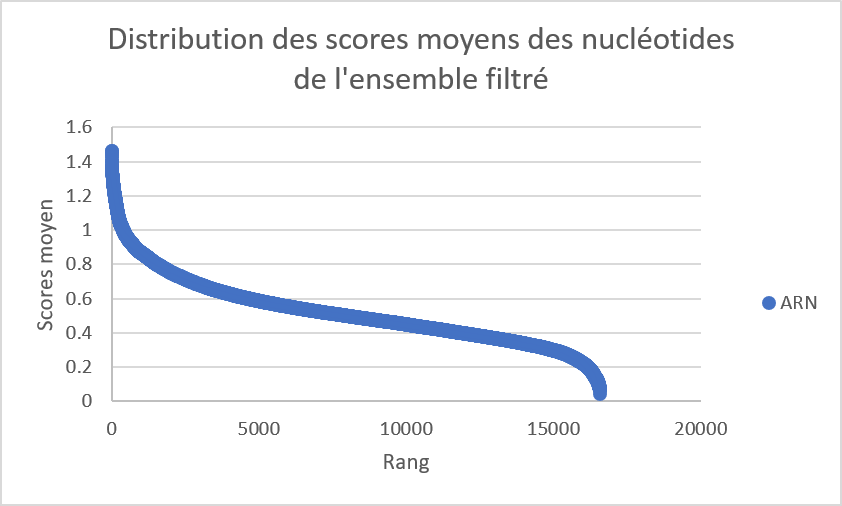
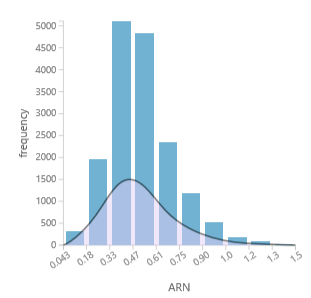
Distribution des scores de réactivité moyens par ARN

de l’ensemble non filtré

Score moyen

Rang des ARN

1. **Pour chaque ARN la moyenne des scores des valeurs de réactivité de ses nucléotides est calculée et ordonnée.** On remarque que peu d’ARN ont un score moyen entre 2 et 6. Ces ARN proviennent de l’ensemble non filtré.



Scores moyens

Fréquence des scores

Rang des ARN

1. **La différence entre cette figure et la figure 10 est qu’ici les données proviennent de l’ensemble filtré.** Le filtre écarte les ARN ayant une réactivité moyenne au-dessus de 1.5, ceux ayant un contenu en adénine au-dessus de 50% et ceux ayant un rapport signal sur bruit sous 2. En plus petit à droite, se trouve la même distribution, mais représenté sous forme d’histogramme.

La distribution des moyennes des scores de chaque ARN de l’ensemble filtré suit une courbe normale telle que déterminée par le test de Kolmogorov-Smirnov [51].

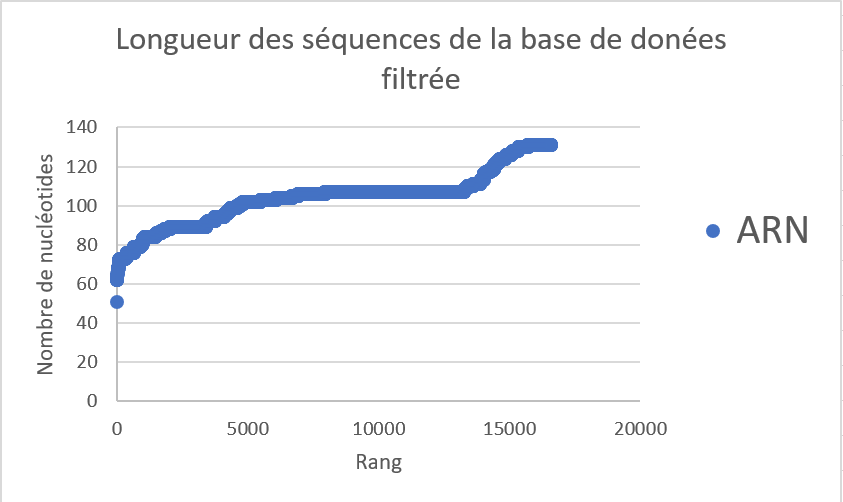
### Fréquence pairée normalisée

La fréquence pairée normalisée de chaque nucléotide pour chaque logiciel de prédiction de SS représente le nombre de fois que le nucléotide est pairé en moyenne dans l’ensemble de SS. 0 représente un nucléotide jamais pairé et 1 un nucléotide toujours pairé. En moyenne, les nucléotides des SS du logiciel RNAsubopt sont pairées à 53,9288% du temps et ceux de MCFlashfold sont pairés à 87,6139% du temps. Plus les nucléotides sont pairés, plus le nombre de MCN pour un ARN sera grand et plus le nombre de nucléotides qu’ils comprennent sera petit. Statistiquement, plus un MCN a de nucléotides, moins le nombre d’occurrences dans la base de données de ce MCN sera grand. Cette caractéristique est importante puisque pour déterminer la réactivité du nucléotide du MCN (lorsque la position du nucléotide lui est ajoutée), plus le nombre d’occurrences est grand plus sa puissance statistique augmente.

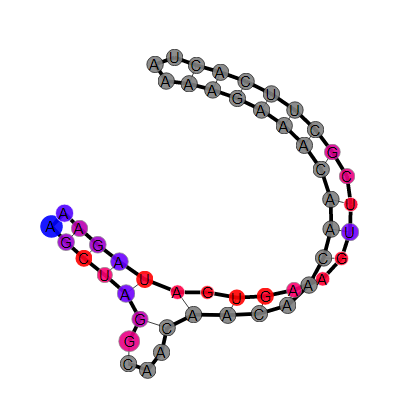
### Longueur des séquences

La longueur des séquences comprend la boucle qui sert de contrôle et la séquence qui sert à l’identification. La plus petite séquence comprend 51 nucléotides et la plus grande a 131 nucléotides.

Longueur des séquences de la base de données filtrée

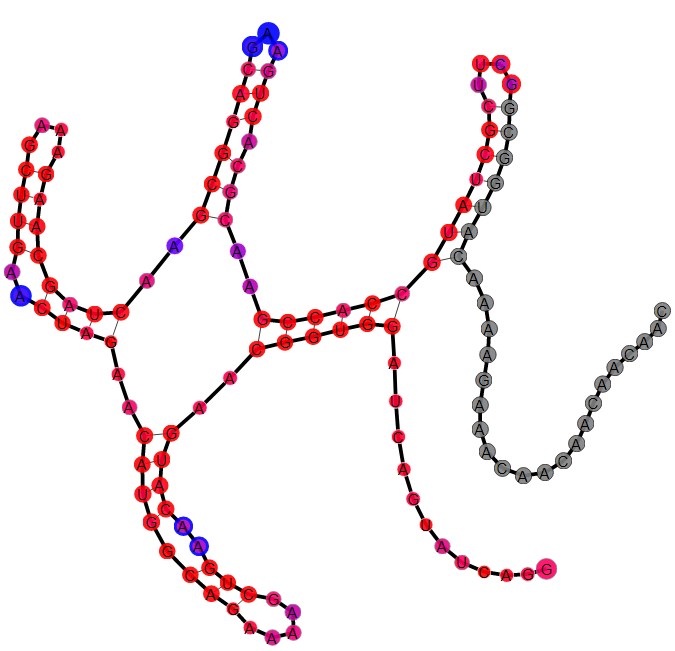


Rang des séquences

1. **Distribution de la longueur des séquences de l’ensemble de données filtré.** La moyenne de la longueur est de 103.98.

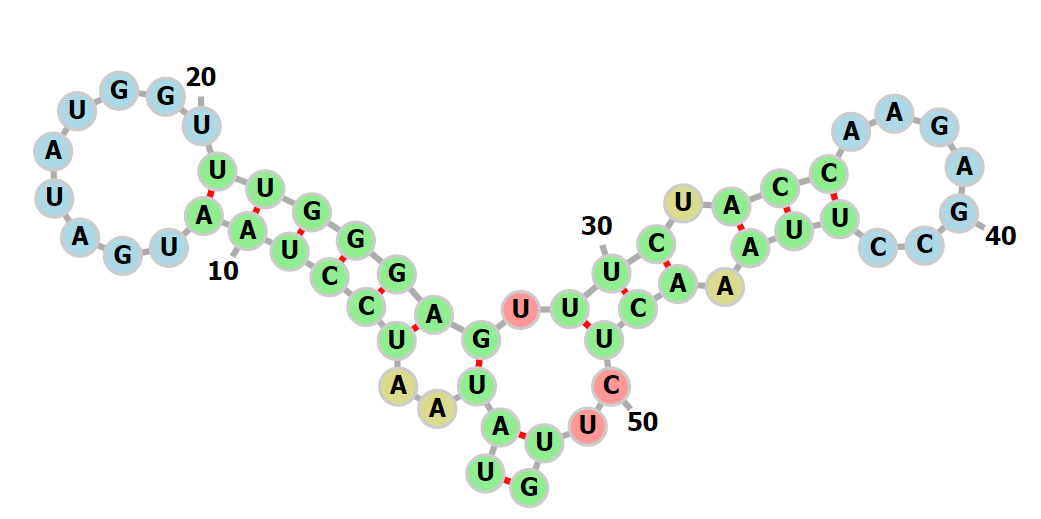
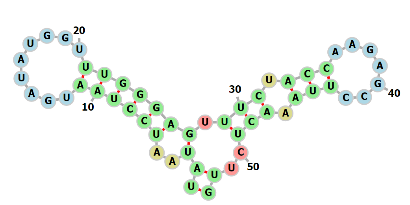
B

A



1. **L’ARN le plus long (131 nt) à gauche (A) avec le plus petit (51 nt) à droite (B).** Le mode couleur selon la réactivité est activé. Un nucléotide bleu est réactif et un nucléotide rouge l’est peu.

### Algorithme d’identification des MCN



1. **SS provenant de FORNA.** Les doubles hélices sont vertes, les renflements sont rouges et les loupes terminales et intérieurs bleus et jaunes respectivement. Les MCN sont placés dans des boites. La figure originale est en bas à droite.

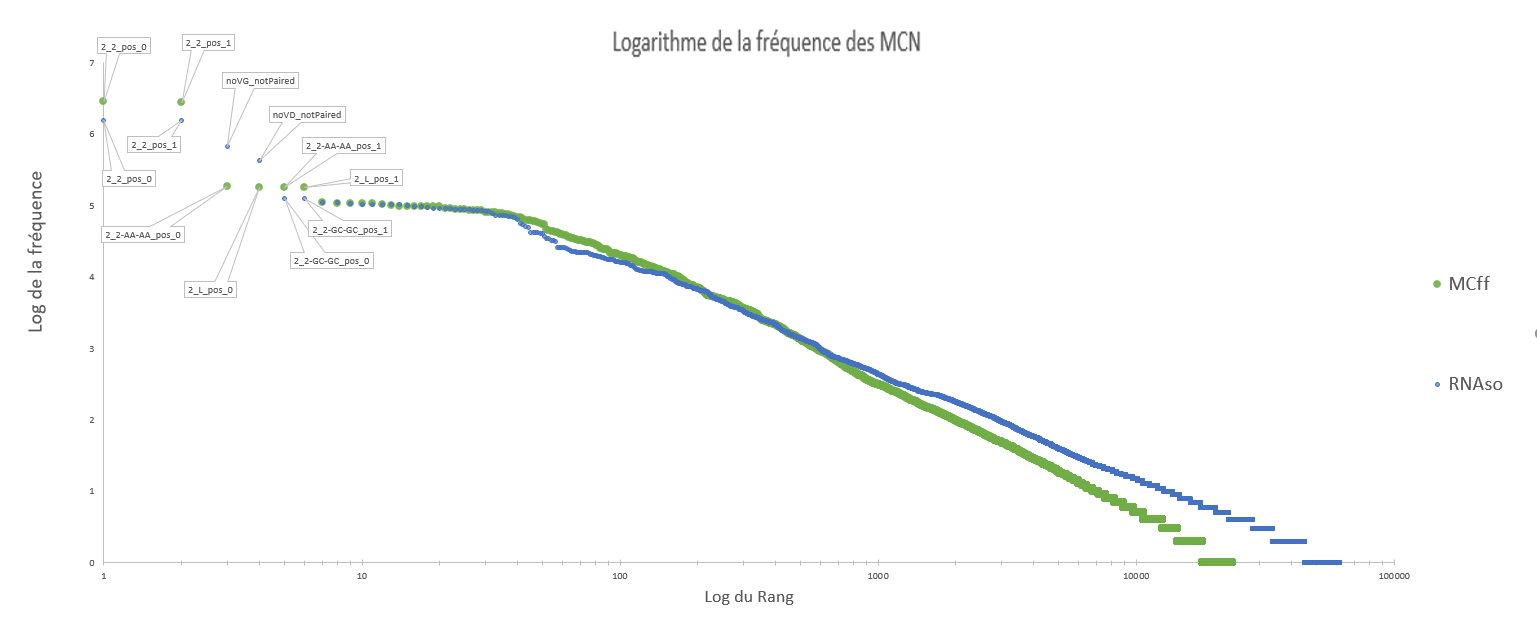
La première étape consiste à rechercher les voisins pairés du nucléotide, on les nommera : le voisin gauche (en 5’) et le voisin droit (en 3’). Ensuite, l’algorithme peut être divisé en deux branches selon que le nucléotide est pairé ou non pairé dans la SS analysée. Lorsqu’il est pairé, il fera partie d’un MCN comportant deux brins d’ARN, la longueur de chaque brin est déterminée et elles sont séparées par une barre de soulignement dans la chaine de caractères de sortie. Lorsque le nucléotide est non pairé, il y a deux cas possibles, soit qu’il se trouve dans une boucle ce qui est reconnu par le fait que le partenaire du premier voisin de gauche pairé est le premier voisin pairé de droite, soit qu’il est dans une boucle interne ou un renflement, les deux derniers cas sont gérés de la même façon. Dans tous les cas, lorsqu’il est non pairé, il appartient à un seul MCN, la figure 14 nous le démontre clairement. Dans le cas de la boucle, on comptabilise le nombre de nucléotides non pairés, plus les deux premiers et dans les deux autres cas, c’est la distance en nucléotide entre les deux voisins pairés et leurs partenaires qui est comptabilisé. Finalement, on identifie le MCN par la séquence des ou du brin(s) les composant et on ajoute la position du nucléotide. La valeur de la position débute toujours par 0. Les nucléotides de début et de fin de séquence ont un identifiant spécial qui leur est réservé.

Ceci est fait pour chaque SS de chaque logiciel. On peut vérifier l’exactitude de l’algorithme grâce au logiciel de visualisation. En effet, lorsqu’on passe la souris au-dessus d’un nucléotide, le ou les MCN s’illuminent. Une amélioration possible de cet algorithme serait de mieux définir les MCN des jonctions.

## Compilation des MCN

Pour prédire la réactivité des nucléotides à l’aide des MCN, il faut extraire les nucléotides en fonction du MCN auxquels ils appartiennent. Seuls les MCN présents dans plus de 20% des SS d’un logiciel sont considérés. Ceci force le nombre de MCN compilé, pour un nucléotide donné, à être inférieur à 5 puisque pour l’ensemble d’entrainement 10 SS sont généré par logiciel. Lors de cette étape, la fréquence des MCN pour un nucléotide est conservée dans la base de données pour des analyses ultérieures.

La figure 15 de la page suivante montre la distribution de la fréquence des MCN dans les données de la RMDB. Dans l’ensemble de données non filtré, le nombre total de MCN distincts est de 84 694. Il y en a 61 142 qui proviennent du logiciel RNAsubopt et MCFlashfold en compte 23 552. Cette différence s’explique par la grosseur moyenne d’un MCN. MCFlashfold prédit plus de paires de base ce qui diminue la grosseur moyenne des MCN. Le nombre de MCN ayant plus de 1000 nucléotides est de 575 pour MCFlashfold et de 596 pour RNAsubopt. De ces MCN, 213 sont communs aux deux logiciels.



1. **Distribution des MCN en fonction de leur fréquence pour l’ensemble non filtré.** Une transformation logarithmique de base 10 a été appliquée pour mieux voir les MCN les plus fréquents. Chaque logiciel est indépendant par rapport à l’ordre de ses MCN. Les MCN du logiciel de RNAsubopt sont représentés par des points bleus et les MCN de MCFlashfold sont représentés par des points verts. Certains MCN caractérisent plusieurs millions nucléotides dans la base de données complète de RMDB. Pour être considéré, le nucléotide doit être dans la conformation du MCN dans plus de 20% des SS sous-optimales. On trouve dans les boite des MCN les plus fréquents, l’identificateur du MCN.

Une transformation logarithmique en base 10 a été appliquée sur les deux axes de la figure 15 pour mieux distinguer les MCN ayant des valeurs élevées.

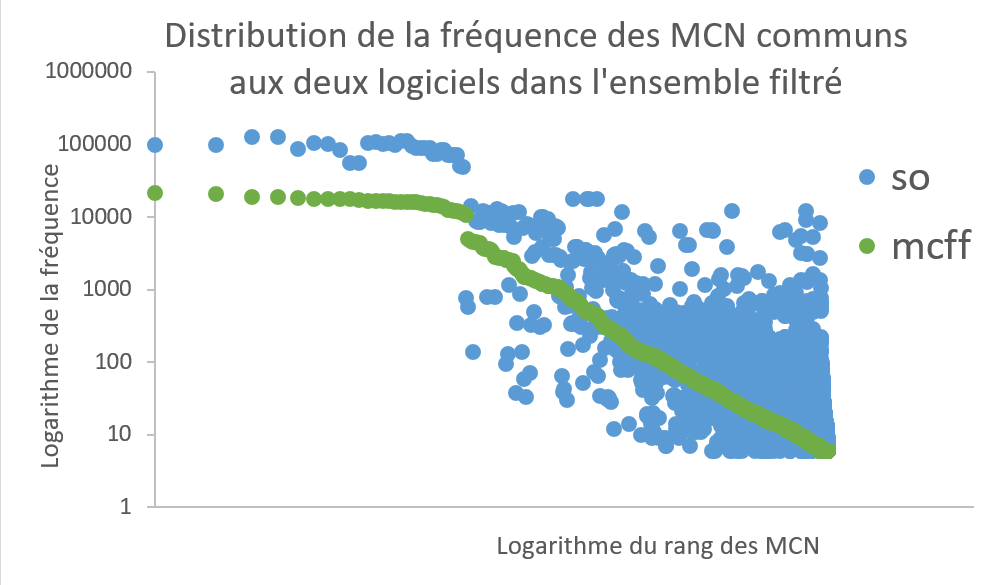
Leur valeur de réactivité est explicitée dans le tableau III, ci-dessous. Ce qui est remarquable dans ce tableau, c’est le nombre peu élevé de 0 des deux types de réactivités. Puisque dans chaque MCN on trouve des nucléotides des deux classes, on peut conclure que les MCN n’expliquent pas entièrement la réactivité, autrement dit, il y a d’autres facteurs à considérer ou les prédictions des SS demandent de l’amélioration. La réponse se trouve surement entre les deux.

Cependant, la plupart des MCN ont un biais marqué. Une autre chose remarquable de ce tableau est l’absence de contradiction entre les deux logiciels en ce qui concerne quelle valeur de réactivité à une fréquence majoritaire pour chaque MCN. Cette caractéristique est un argument en faveur de l’hypothèse qui veut qu’en général les logiciels de prédictions de SS font des bonnes prédictions puisque MCFlashfold et RNAsubopt sont deux méthodes indépendantes de prédictions des SS. Le terme « *noVG\_notPaired* » signifie que ce nucléotide n’est pas pairé et il n’a pas de voisin pairé en 5`. « *VG* » signifie voisin gauche. « *noVD\_notPaired* » identifie les nucléotides pour lesquels il n’y a pas de voisin droit pairé.

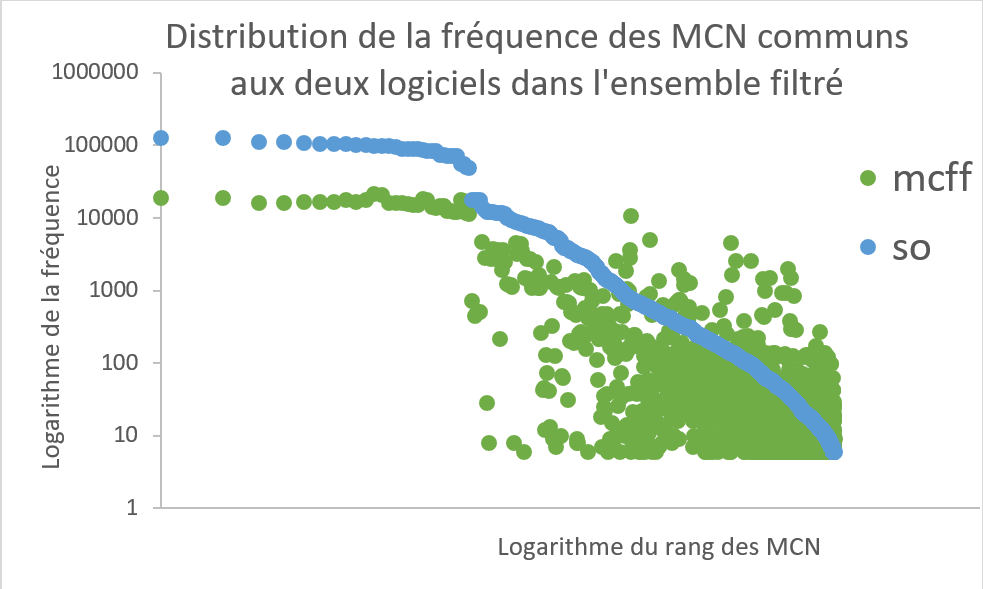
1. MCN les plus présents et leurs valeurs de réactivités   
   dans l’ensemble non-filtré

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MCN | Fréquence des MCN | | | |
| RNAsubopt | | MCFlashfold | |
| Basse | Élevé | Basse | Élevé |
| 2\_2\_pos\_0 | 1 357 111 | 64 727 | 1 961 587 | 381 301 |
| 2\_2\_pos\_1 | 1 324 986 | 74 831 | 1 934 676 | 385 707 |
| noVG\_notPaired | 203 500 | 229 648 | 464 | 642 |
| noVD\_notPaired | 428 746 | 956 | 2 728 | 0 |
| 2\_2-AA-AA\_pos\_0 | 0 | 0 | 72 742 | 70 214 |
| 2\_L\_pos\_0 | 27 986 | 1 963 | 81 088 | 64 071 |
| 2\_L\_pos\_1 | 28 736 | 1 912 | 81 287 | 60 744 |
| 2\_2-GC-GC\_pos\_0 | 114 177 | 2 061 | 99 099 | 1 819 |
| 2\_2-GC-GC\_pos\_1 | 114 247 | 1 551 | 99 280 | 1 645 |

Les figures 16 et 17 sont les fréquences des MCN de l’ensemble filtré ordonné par la fréquence des logiciels à tour de rôle. Les points alignés verticalement proviennent du même MCN. Seuls les MCN communs aux deux logiciels sont illustrés. Une transformation logarithmique en base 10 a été appliquée aux deux axes. On aperçoit un plateau à gauche des graphiques, il est formé des MCN les plus petits, les « 2\_2 ». Le MCN le plus fréquent est : « 2\_2-GU-AC\_pos\_0 ». Bien que le nombre total de 2\_2 soit plus élevé dans les SS de MCFlashfold (voir figure 15), RNAsubopt a constamment un nombre de nucléotides par MCN plus élevé. Ceci est explicable par le fait que MCFlashfold prédit des paires de base non-canoniques. Ce qui répartit les nucléotides dans plus de MCN.



1. **Distribution des fréquences des nucléotides par MCN dans l’ensemble filtré.** Les MCN sont ordonnés selon la fréquence des MCN de MCFlashfold.

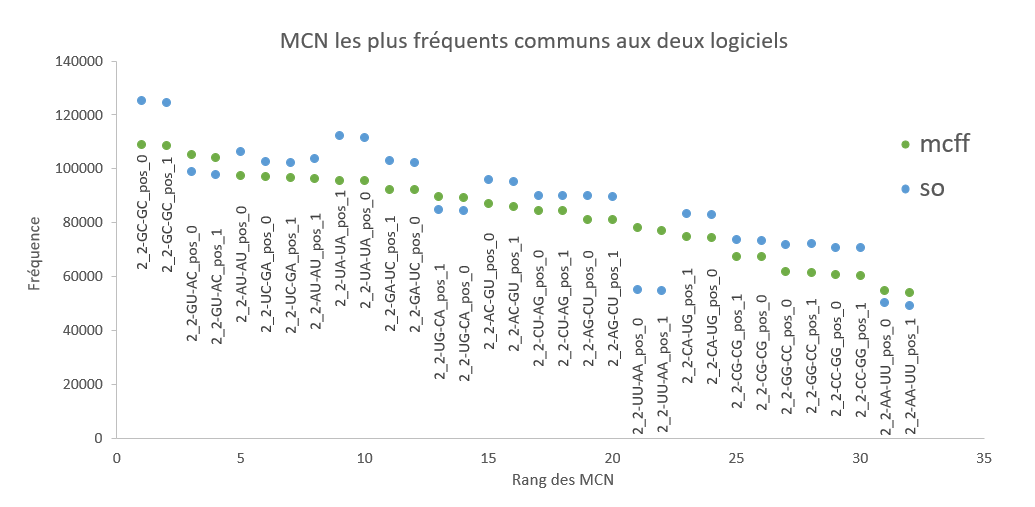


1. **Figure 16 inversée : les MCN sont ordonnés selon la fréquence des MCN de RNAsubopt.**

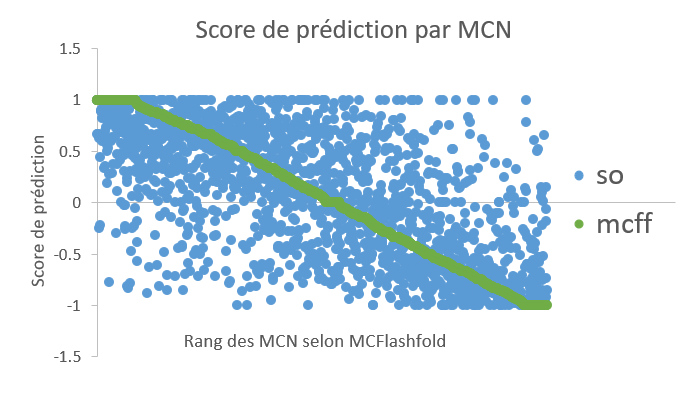
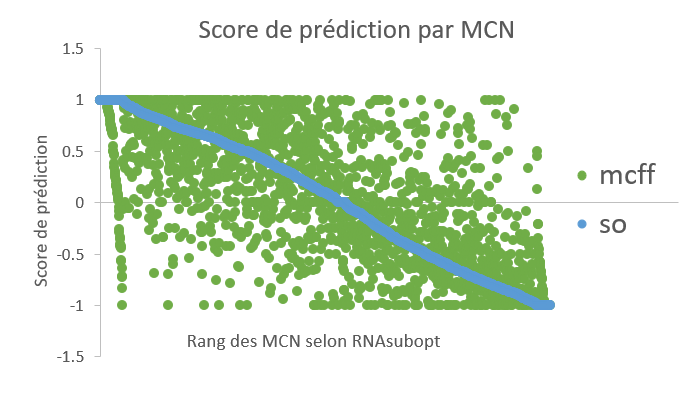
Le tableau IV est semblable au tableau III. On y voit le nombre de nucléotides faisant partie des MCN les plus fréquents de l’ensemble de données filtré. La colonne totale est le nombre de nucléotides de la classe « Hi » plus ceux de la classe « Low ». La colonne différence, comme son nom l’indique, est la différence entre le nombre de nucléotides ayant une réactivité élevée (« Hi ») et ceux ayant une réactivité basse (« Low ») pour le logiciel RNAsubopt et MCFlashfold. La colonne ratio est le résultat de la colonne différence divisée par le nombre total de nucléotides. Lorsque la valeur absolue de ce ratio tend vers 1, cela signifie que les nucléotides du MCN sont constamment réactifs ou non réactifs selon que le ratio est positif ou négatif respectivement.

1. Fréquence des nucléotides des MCN de RNAsubopt et MCFlashfold ordonnée selon la somme des nucléotides de hautes et basses réactivités de RNAsubopt dans l’ensemble de données filtré.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Nom du MCN** | **Nombre de nucléotides total (Hi + Low) so** | **Différence so** | **Nombre de nucléotides total (Hi + Low) mcff** | **Différence mcff** | **Ratio  So** | **Ratio MCff** |
| 2\_2-GC-GC\_pos\_0 | 125470 | -112116 | 109185 | -97280 | -0.893 | -0.890 |
| 2\_2-GC-GC\_pos\_1 | 124541 | -112696 | 108655 | -97635 | -0.904 | -0.898 |
| 2\_2-UA-UA\_pos\_1 | 112327 | -86408 | 95831 | -74733 | -0.769 | -0.779 |
| 2\_2-UA-UA\_pos\_0 | 111662 | -92867 | 95730 | -79500 | -0.831 | -0.830 |
| 2\_2-AU-AU\_pos\_0 | 106443 | -76420 | 97478 | -66774 | -0.717 | -0.685 |
| 2\_2-AU-AU\_pos\_1 | 103962 | -88838 | 96290 | -77878 | -0.854 | -0.808 |
| 2\_2-GA-UC\_pos\_1 | 102951 | -73121 | 92271 | -64809 | -0.710 | -0.702 |
| 2\_2-UC-GA\_pos\_0 | 102619 | -93206 | 97151 | -86758 | -0.908 | -0.893 |
| 2\_2-GA-UC\_pos\_0 | 102365 | -79364 | 92251 | -69446 | -0.7753 | -0.7527 |



1. **Les 32 MCN les plus fréquents selon l’ordre de MCFlashfold dans l’ensemble filtré.** Ces MCN ont des valeurs semblables pour les deux logiciels. La fréquence est le nombre de nucléotides dans la conformation d’un MCN donné.



B

A

1. **Distribution de la différence entre le nombre de nucléotides réactifs et ceux non réactifs divisé par la somme de ces deux valeurs pour les MCN communs aux deux logiciels.** À gauche, l’ordre est donné par le ratio des MCN de MCFlashfold et à droite il l’est par le ratio de RNAsubopt.

Les scores de prédiction de la figure 16 sont le ratio des deux colonnes de droite du tableau II pour tous les MCN ayant plus de 5 représentants pour les deux logiciels. Voici 3 observations tirés de ces graphiques :

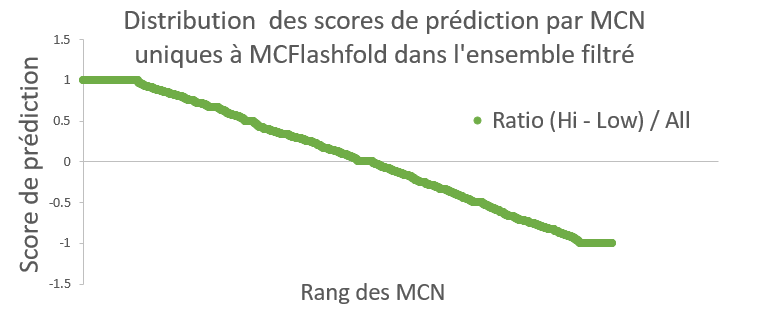
1. Il y a une corrélation de 0.68 entre les deux scores des logiciels, selon la corrélation de Pearson.
2. Une quantité non négligeable de MCN ont un ratio parfait, de 1 ou -1.
3. Le nombre de score parfait est plus grand pour le logiciel MCFlashFold (267) que chez RNAsubopt (155) pour les MCN communs aux deux logiciels.

Le tableau III montre les 10 MCN les plus présents lors des prédictions du logiciel MCFlashfold et absents de celles de RNAsubopt. Les prédictions ont été faites sur l’ensemble de données filtré. Une des caractéristiques distinguant MCFlashfold des autres logiciels de prédiction de SS est sa capacité à prédire les paires de bases non-canoniques. Le tableau III contient exclusivement des MCN composés d’au moins une paire de bases non-canoniques. On remarque que ces nucléotides n’ont pas une valeur prédictive très élevée. Par exemple, le MCN ayant le ratio le plus élevé (dans la boite du tableau V) en valeur absolue, le MCN « 2\_2‑AG‑CA\_pos\_1 » est associé à près de 14 000 nucléotides réactifs ou peu réactifs et la différence entre le nombre de nucléotides réactifs et les peu réactifs est de 6 821, ce qui représente environ 50% du total. Ceci signifie qu’il y a deux fois plus de nucléotides de basse réactivité que de nucléotides de haute réactivité.

Pour qu’un MCN soit prédictif, on aimerait que la différence soit supérieure. Ceci étant dit, une grande quantité de MCN peu prédictifs peut donner tout de même des informations utiles.

1. Cohérence des 10 MCN les plus fréquents et unique à MCFlashfold

|  |  |  |  |
| --- | --- | --- | --- |
| **Nom du MCN** | **Nombre de nucleotides total** | **Différence entre le nombre de  « Hi » et de « Low »** | **ratio** |
| 2\_2-AA-AA\_pos\_0 | 21 307 | 5 975 | 0.28042 |
| 2\_2-AA-AA\_pos\_1 | 19 863 | 2 712 | 0.13653 |
| 2\_2-GA-AC\_pos\_1 | 14 939 | 4 083 | 0.27331 |
| 2\_2-AG-CA\_pos\_0 | 14 378 | -341 | -0.02372 |
| 2\_2-GA-AC\_pos\_0 | 14 089 | -540 | -0.03833 |
| 2\_2-AG-CA\_pos\_1 | 13 750 | -6 821 | -0.49607 |
| 2\_2-CA-AG\_pos\_1 | 9 982 | -938 | -0.09397 |
| 2\_2-GG-AC\_pos\_1 | 9 783 | -1 168 | -0.11939 |
| 2\_2-GG-AC\_pos\_0 | 9 308 | -4 609 | -0.49517 |



1. **Valeur du ratio des nucléotides réactifs moins les peu réactifs, divisés par leur total pour les MCN unique à MCFlashfold.** Autrement dit, la valeur de chacun des points composant cette distribution est calculée en soustrayant le nombre de nucléotides ayant une valeur de réactivité sous 0,5 de ceux ayant une valeur au-dessus de 1 et en divisant ce nombre par la somme des nucléotides de ces deux classes pour chaque MCN. À gauche complètement, le premier nucléotide de la boucle « UUCG », lorsqu’elle est fermée par une paire UA a une valeur élevée à 73 occasions et n’a jamais une valeur basse dans toute la base de données filtrée. De l’autre côté, le deuxième nucléotide de la boucle « GCCU » lorsqu’elle est fermée par une paire CG à une valeur basse pour 41 nucléotides distincts et n’a jamais de valeur haute. Ces deux MCN se retrouvent aux deux extrémités de ce graphique. En tout, 2 500 MCN sont représentés.

## Prédictions basées sur la cohérence

Lors de la phase de prédiction, chaque nucléotide reçoit une valeur entre -1 et 1. Le signe qualifie la prédiction de vrai ou fausse et un poids entre 0 et 1 module la prédiction. Plus, un MCN est constant dans la valeur des étiquettes des nucléotides qui le compose, plus le score du poids se rapproche de 1. Contrairement, a un score fixe, ce poids permet de diminuer l’impact d’une prédiction pour laquelle les données ne sont pas en accords. De plus, il permet d’augmenter la pénalité d’une prédiction se trompant grandement. Ceci m’amène donc à penser que la somme des scores des nucléotides est une bonne mesure de la confiance qu’on peut avoir envers la prédiction de la SS d’un ARN. Le score de plus de 150 000 SS de plus de 15 000 ARN est disponible dans la base de données des MCN et par le logiciel de visualisation.

Les MCN des nucléotides qui n’ont pas assez de données et les nucléotides qui n’ont pas de valeur du tout n’ont pas de prédiction. Cela augmente la vitesse de l’algorithme de beaucoup en pratique. En effet, d’un point de vue algorithmique, cette étape est l’étape limitante. Ceci est dû au fait que le nombre de requêtes à la base de données est égal au nombre de SS sous-optimales multiplié par le nombre de nucléotides de la séquence pour chaque logiciel de prédiction pour un ARN. Le calcul de 10 SS de 1000 ARN composés de 100 nucléotides pour lesquels on a une valeur de réactivité, c’est 1 million de requêtes qui sont faites à la base de données, d’où l’importance d’optimiser les performances de la base de données en compilant les MCN par valeur de réactivité et en les indexant sur les MCN et les logiciels.

En ce moment, pour avoir les prédictions sur 100 SS de 20 ARN, cela prend 1 minute. Les détails du serveur sur lequel mes expériences sont exécutées sont en annexe.

## Conclusion du chapitre 1

RNASS est un ensemble de fonctions intégrant trois logiciels de prédiction des SS. Il crée des fichiers standardisés ayant une structure arborescente. Ces fichiers sont dans le format JSON, un format très populaire. Vu leur structure, il est facile d’ajouter des informations sans compromettre leur compatibilité avec les logiciels qui les utilisent. Toutes les données qu’ils contiennent pourraient être calculées au besoin, mais leur « précalcule » augmente la fluidité des outils d’analyses qui leur sont dédiés. Les limites du « précalcule » sont surtout l’espace disponible et notre imagination. L’espace mémoire est de moins en moins couteux cela rend le « précalcule » de plus en plus avantageux.

RNASS\_v2 a été amélioré tout au long de mon parcours et j’espère qu’il le sera encore dans le futur.

# Chapitre 2 : RDV : un logiciel de visualisation de la structure secondaire des ARN

RDV est l’abréviation de RNA Dynamic Viewer. Ce logiciel permet de visualiser les données provenant de RNASS. Il a été écrit en *Javascript* et utilise Node.js une plateforme qui permet de créer des sites *web* modernes. RDV superpose les données de réactivités chimiques sur les SS d’un ARN. Tout comme le logiciel de visualisation de structure secondaire FORNA, RDV utilise une librairie codée en *Javascript* nommé d3.js pour représenter la structure secondaire d’un ARN. Cette librairie implémente un champ de force dans l’objectif de dessiner des réseaux formés de nœuds et de liens.

## Visualisation du graphe des transitions

Parmi les fonctions innovantes de RDV se trouve la possibilité de voir un ensemble de SS les unes après les autres en continuité. Il est ainsi plus facile de voir la différence entre deux SS.

La transition fluide entre les SS permet aussi de mieux comprendre la dynamique des ARN. Grâce à ce logiciel, il m’a été possible d’avoir une vue d’ensemble sur les ARN sondés. L’avantage de la flexibilité des composantes est la possibilité de placer l’ARN de façon à faire ressortir une information importante. Les réseaux de transition des premières SS de chaque logiciel de prédiction des SS sont dessinées à la droite de la SS actuelle. Chaque cercle représente une SS et leur position sur l’axe vertical est proportionnelle à leur énergie, plus l’énergie des nucléotides est basse plus le cercle sera bas. Un lien est dessiné entre deux SS partageant 90 % de leurs paires de bases. Un « click » sur un SS, la sélectionne.

## Visualisation de la cohérence

Une autre fonction innovante de RDV est la possibilité de valider une structure secondaire prédite en se basant sur la réactivité moyenne d’un nucléotide à l’intérieur d’un MCN.

La couleur des SS dans le graphe des transitions est reliée à la vraisemblance de la SS, plus elle est pâle meilleur elle est.

Comme je l’ai expliqué dans le chapitre 1, le contour des nucléotides est relié à la réactivité moyenne des autres nucléotides à la même position dans les MCN de la base de données. Un contour rouge signifie une réactivité moyenne basse et un contour bleu signifie une réactivité moyenne élevée. Voir tableau I et II. Une cohérence globale entre la couleur de contour et la couleur de l’intérieur des nucléotides signifie que la structure secondaire est plausible. De plus, l’épaisseur du contour est reliée à la confiance de la prédiction, ce qui fait ressortir les mauvaises prédictions ayant une confiance élevée. La valeur exacte de la prédiction globale est disponible en plaçant la souris au-dessus de la conformation ou du nucléotide d’intérêt.

## Visualisation de la SS

Du côté gauche, la SS sélectionnée est dessinée avec le champ de force décris plus haut. Dans cette visualisation, chaque cercle correspond à un nucléotide, trois modes de couleur sont disponibles,

* le mode : couleur selon la sorte du nucléotide
* le mode : couleur selon la réactivité du nucléotide.
* le mode : couleur selon l’erreur reliée à chaque nucléotide.

## Obtenir des détails et rechercher des ARN semblables

La nécessité d’avoir un outil de visualisation pour ce type de prédiction vient de la présence de nucléotides se comportant de façon inattendue. Ces nucléotides sont extrêmement intéressants lorsque vient le temps de comprendre les interactions tertiaires des nucléotides au sein d’un ARN, mais sont difficilement explicables par les MCN. Heureusement, la majorité des interactions sont locales.

Lorsqu’un MCN a un nombre significatif de nucléotides appartenant à la classe de réactivité basse ou haute, la prédiction est considérée comme fiable et l’algorithme peut se prononcer. Un système de pointage donnant 1 point à une bonne prédiction et enlevant un point à une mauvaise prédiction a été mis sur pied. La somme des prédictions pour chaque conformation donne une idée du niveau de confiance que nous avons pour cette conformation. La couleur des nœuds du réseau des conformations est reliée à cette somme. Le gradient de couleur passe du vert pâle au vert foncé pour le logiciel MCFlashfold et du rouge pâle au rouge foncé pour le logiciel RNAsubopt. Dans les deux cas, une couleur pâle signifie une moins bonne prédiction qu’une couleur foncée. L’étendue des valeurs est calculée en prenant les conformations des deux logiciels en considération. Lorsqu’on passe la souris par-dessus un nœud, on obtient la somme de façon numérique pour une meilleure précision.

Une autre caractéristique intéressante de ce logiciel est la possibilité d’obtenir les détails pour chaque nucléotide prédit. En cliquant sur un nucléotide, on obtient les statistiques sur le ou les MCN du nucléotide en question pour la SS dans laquelle il se trouve lors du « click ». Ce profil comprend 3 décomptes : les nucléotides de basses, moyennes et hautes réactivités de tous les ARN de la base de données filtrée. Il comprend aussi une jauge visuelle donnant le ratio du nombre de nucléotides hautement réactifs par rapport au nombre de nucléotides non réactifs. La jauge est complètement noire pour un MCN ayant que des nucléotides hautement réactifs et grise pour les MCN ayant que des nucléotides peu réactifs. De plus, on peut obtenir une URL pour chaque nucléotide ayant un comportement inattendu cela dans le but de tester des hypothèses rapidement sans avoir besoin de les coder. Les types d’hypothèses pouvant être testé pour le moment sont : y a-t-il un motif nucléotidique dans la SS qui est récurent pouvant se lier au nucléotide et ainsi le rendre inaccessible au réactif? Est-ce que le nucléotide apparait dans un sous-ensemble de SS distinctes d’un autre sous-ensemble, rendant le premier ensemble moins probable?

La couleur du contour des nucléotides représente la prédiction ou son absence. Tel que mentionné dans le tableau, une couleur rouge signifie que l’on s’attend à ce que le nucléotide ait une basse réactivité, une couleur bleue signifie une haute réactivité, le mauve représente une incohérence entre les deux MCN du nucléotide et le gris de l’incertitude. Les nucléotides qui n’ont pas de valeur de réactivité n’ont pas de prédiction non plus, ce qui est représenté par un contour noir. Lorsqu’il n’y a pas de donnée sur le MCN, le contour est blanc.

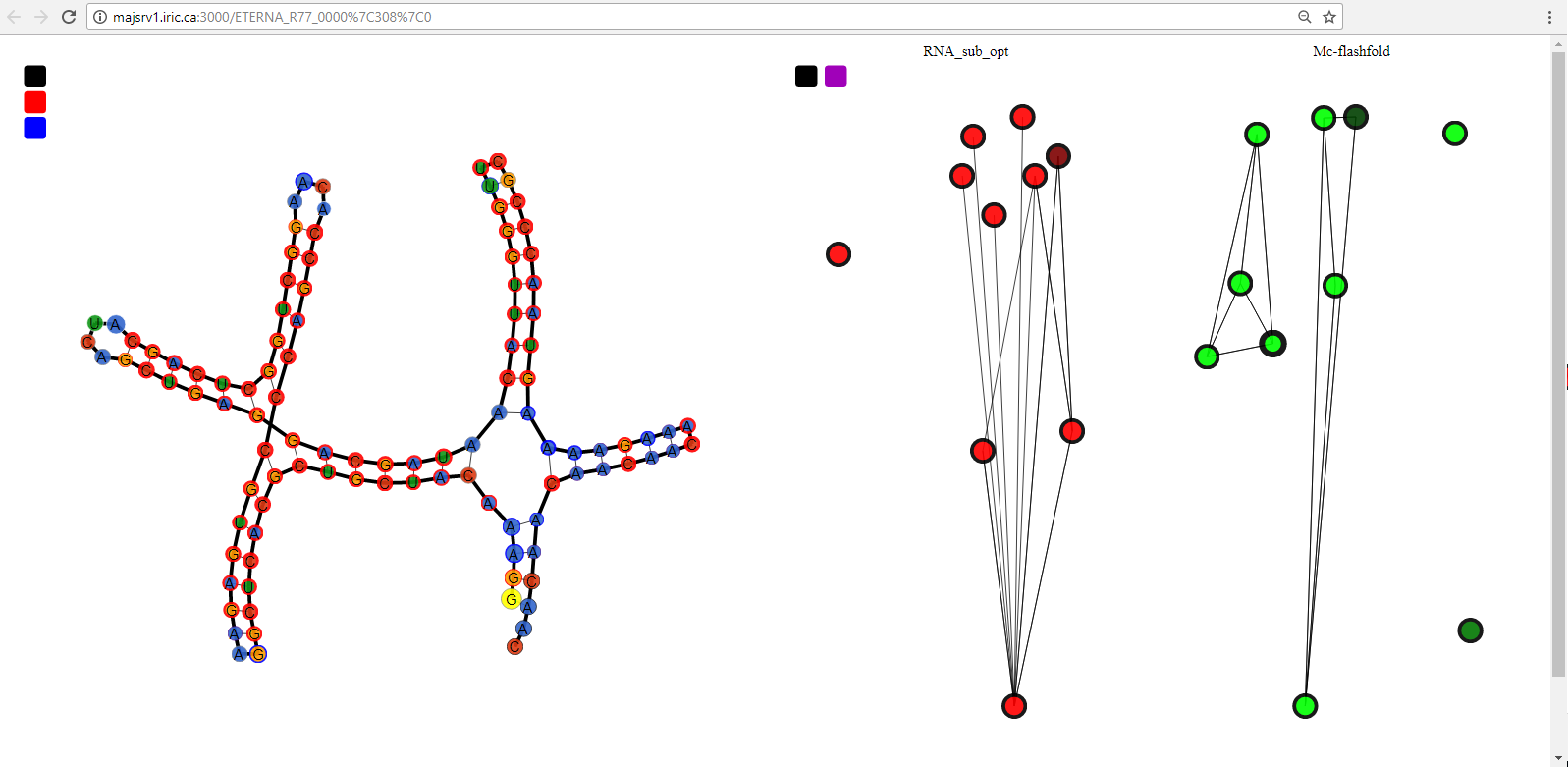
1. Signification des valeurs de prédictions 1 et -1 dans RDV

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Valeur de prédiction | 1 | | -1 | |
| Signification | Bonne prédiction | | Mauvaise prédiction | |
| Réactivité observé | Haute | Basse | Haute | Basse |
| Couleur | Bleu | Rouge | Rouge | Bleu |

1. Signification possible de la valeur 0 lors des prédictions  
    et couleur du contour du nucléotide dans RDV

|  |  |  |
| --- | --- | --- |
| Valeur de prédiction | Signification | Couleur |
| 0 | La réactivité observée est dans la classe *Bg*. | Gris |
| Incohérence entre les deux MCN prédits | Violet |
| Pas de valeur pour ce nucléotide. | Noir |
| Pas assez de données pour faire une prédiction | Blanc |

.



Traverser le graphe des transitions

Mouvement aléatoire de la SS (gardez enfoncé pour activer)

Mouvement aléatoire

du graphe des transitions

Changer le mode de couleur des nucléotides

Arrêt prématuré du

mouvement aléatoire

du graphe des transitions

SS de

minimum d’énergie

(MFE)

SS sélectionnée

Information sur la SS

pour laquelle la souris

est au-dessus

Prédiction de RNASS

Information sur le MCN

pour lequel la souris est au-dessus

Nucléotide mis en évidence

(dernier champ de l’URL)



1. Vue principale de RNA Dynamic Viewer (RDV).

La « simulation » commence en hélice et se déploie automatiquement. Pour démêler les bras à l’aide du mouvement aléatoire on peut appuyer sur « z », pour activer une minuterie de quelques secondes. Pour agiter les nucléotides, on « click » sur le rectangle noir en haut à gauche. Celui qui est rouge lance et arrête la traversée des graphes de transitions. On change le mode de couleur en appuyant sur le rectangle bleu. En cliquant sur le cercle lui correspondant dans le graphe des transitions, on change la SS. Les performances de la simulation dépendent des performances de votre ordinateur, il a été testé dans le navigateur « chrome ». L’objet « jsonData » nous renseigne sur l’ARN dans la console JavaScript (f12).

Détails des MCN du nucléotide sélectionné



1. **Nombre d’occurrences de deux MCN classé par niveau de réactivité**. Juste en dessous, en retrouve deux listes (classée par niveau de réactivité) des adresses de 10 nucléotides homologues par leur MCN. Elles sont obtenues en cliquant sur le nucléotide d’intérêt dans la simulation de la figure 22. Ces listes apparaissent après avoir cliqué sur un des deux MCN. Lorsque c’est possible, 10 exemples de nucléotides ayant une haute et basse réactivité sont écrites. Ces adresses mènent vers la représentation des ARN pour lesquels un nucléotide est dans la conformation du MCN dans au moins 20% des SS du même logiciel de prédiction que celui d’où provient le MCN d’origine. Le nucléotide est mis en évidence dans l’ARN de l’adresse par un cercle jaune. Le rectangle noir sur fond gris donne une idée de la « pureté » du MCN. Sa longueur est proportionnelle à 4 fois nombre de « Hi » divisé par le nombre de « *Low*» plus le nombre de « *Hi* ».

## Conclusion du chapitre 2

L’outil de visualisation RDV se démarque des autres logiciels de visualisation de SS par la possibilité de voir le graphe des transitions des ARN et la valeur de cohérence attribuée à chacune de ses SS. De plus, il est possible de vérifier des hypothèses rapidement. Il aussi possible de naviguer entre les nucléotides ayant des MCN communs.

Je crois que cet outil à un avenir intéressant puisqu’il est écrit en *JavaScript* avec la librairie d3.js, une librairie libre. Le code de RDV est libre lui aussi.

Les améliorations possibles sont entre autres :

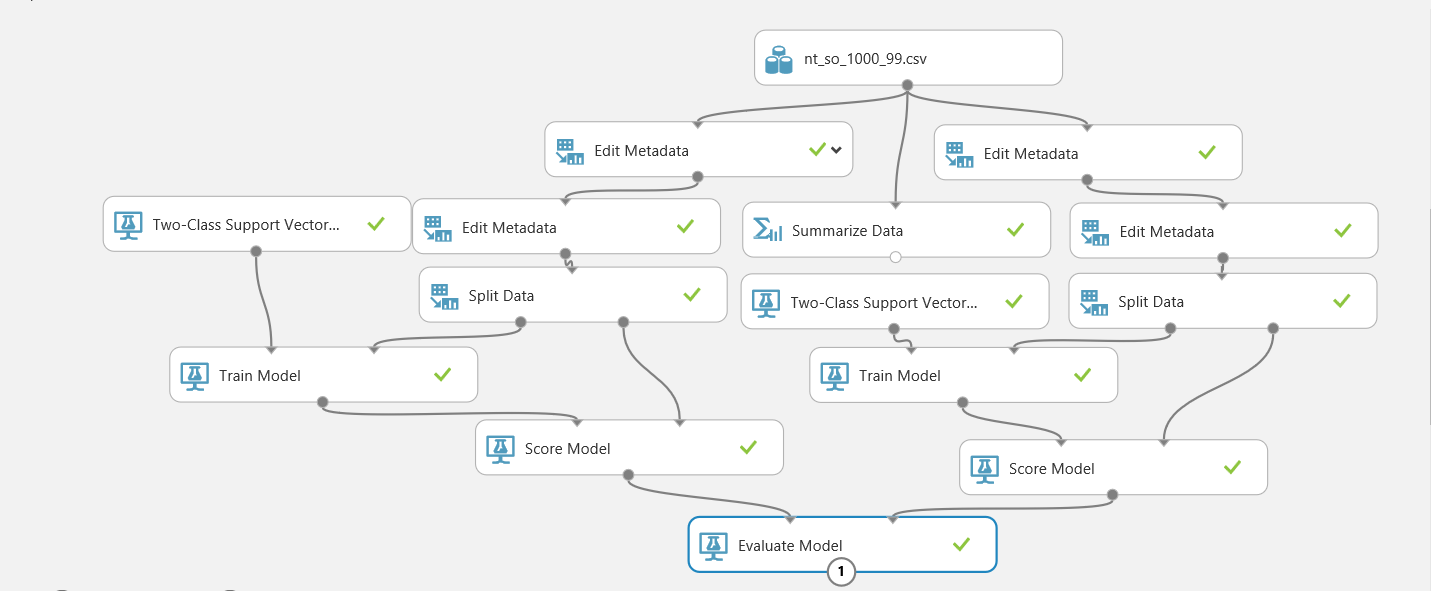
* Navigation automatique entre les SS par marche aléatoire.
* Formation des liens entre les SS dynamique en fonction d’un seuil de distance.
* Visualisation en 3D de la SS.

# Chapitre 3 : Évaluation de la cohérence des MCN

Dans le but de mesurer les performances des prédictions effectuées par RNASS, je les ai évalués dans ce chapitre. J’ai utilisé une machine à vecteur de support pour entrainer un modèle avec deux ensembles de données différents. Le premier ensemble contient les valeurs de pairage des nucléotides et le deuxième contient les valeurs de prédiction de RNASS. Ces nucléotides ont été extraits de l’ensemble de données filtré (voir chapitre 1).

Le graphique de la figure 23 montre les étapes de l’entrainement. La plateforme Azure permet de créer un protocole graphique. Du haut vers le bas, voici les étapes :

1. Les données sont dans cet élément. 2 116 180 nucléotides sont considérés. Ils ont une valeur de réactivité haute (>1) ou basse (<0,5).
2. On configure les métadonnées pour que l’algorithme d’apprentissage sache sur quelles caractéristiques apprendre et quelle caractéristique prédire.
3. On choisit le modèle d’apprentissage et on lui donne les paramètres voulus. Dans le présent cas, les paramètres sont ceux d’origines.
4. On sépare les données pour avoir un ensemble d’entrainement et un ensemble « test ». Ici, on a séparé en deux l’ensemble de départ.
5. On entraine le modèle.
6. On attribue un score avec l’ensemble « test ».
7. On l’évalue. Dans cette étape on calcul les performances avec les métriques expliquées dans l’introduction.



1. Protocole d’apprentissage graphique, partant des données brutes jusqu’à son analyse. La plateforme Azure permet d’entrainer un modèle en effectuant de la programmation graphique. Si on fait abstraction du bloc « *summarise data* », deux chemins partent des données brutes. À gauche, le modèle est entrainé sur les prédictions de RNASS et à droite sur l’état pairé ou non des nucléotides. Dans les deux cas une machine à vecteur de support a été utilisée, bien que pour la caractéristique : état pairé ou non, les valeurs de vrai « Hi », faux « Hi », de vrai « Low » et faux « Low » peuvent être directement obtenues. Le dernier bloc rassemble les données pour créer des graphiques, tels que la courbe ROC et la précision en fonction du rappel.

## Comparaison des prédictions de RNASS avec celles faites à l’aide de l’état pairé ou non d’un nucléotide.

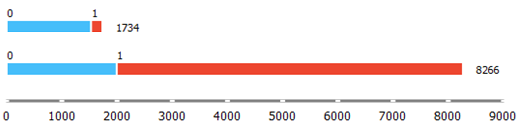
Les modèles ont été entrainés et évalués sur deux ensembles distincts de plus d’un million de nucléotides (1 058 090) provenant des prédictions de la SS de RNAsubopt. Pour les deux modèles, deux ensembles ont été créés en séparant de façon aléatoire un grand ensemble de 2 116 180 nucléotides, l’ensemble d’entrainement et l’ensemble « test ».

1. Tableau de contingence   
   de l’état pairé ou non des nucléotides en fonction de leur niveau de réactivité.

Total :

1 058 090 nucléotides

|  |  |  |  |
| --- | --- | --- | --- |
|  | | État des nucléotides | |
| Non pairés | Pairés |
| Réactivité | Haute | **155 247** | **22 179** |
| Basse | **210 901** | **669 763** |



Nucléotides peu réactifs

Non pairé

Pairé

Nucléotides réactifs

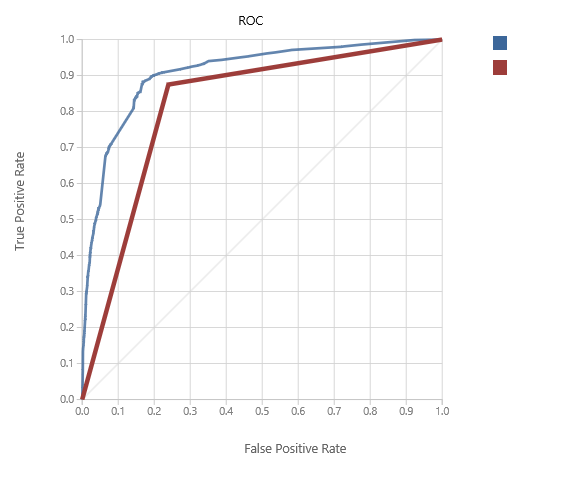
1. **Cette figure met en perspective l’état (pairé ou non) des nucléotides avec leur réactivité chimique.** Le nombre de nucléotides réactifs non pairés est semblable au nombre de nucléotides peu réactifs non pairés. Cependant, il y a beaucoup moins de nucléotides réactifs pairés que de nucléotides peu réactifs pairés. Cette figure a été faite avec 10 000 nucléotides.

Le pourcentage de nucléotides pairés dans ces ensembles est d’environ 65 % et le pourcentage de nucléotides peu réactifs est de 83 %. Ces deux pourcentages supposent déjà que certains nucléotides peu réactifs seront non pairés. En réalité, c’est 20% des nucléotides totaux qui sont peu réactifs et non pairés. 24 % des nucléotides peu réactifs ne sont pas pairés. En revanche, 12.5 % des nucléotides avec une réactivité élevée sont pairés.

La figure 24 est une courbe ROC des deux modèles superposés. Concentrons-nous sur la signification de la valeur de l’axe des « x ». De gauche à droite, le taux de faux « Hi » (nucléotides peu réactifs mal prédit divisé par les nucléotides peu réactifs totaux) augmente linéairement. Plus le taux de faux « Hi » est bas pendant que le taux de vrai « Hi » (nt réactifs bien prédit divisé par les nt réactifs totaux) est élevé, meilleur est l’algorithme.

RNASS donne une probabilité avec ses prédictions. Dans le cas du modèle basé sur l’état pairé ou non des nucléotides, le nombre de prédictions « Hi » est fixe, c’est la raison de la forme « carrée » de la « courbe » rouge du graphique de la figure 24. En fait, un seul point est calculé. On constate que le modèle basé sur l’état des nucléotides à un taux de faux « Hi » d’environ 24 %, ce qui correspond à un taux de vrai « Hi » de 87.4 %. Pour un même taux de faux positifs, RNASS à un taux de vrai « Hi » au-dessus de 90 %, un gain de plus de 2.4 %

Courbe ROC



Taux de vrais « Hi »

Taux de faux « Hi »

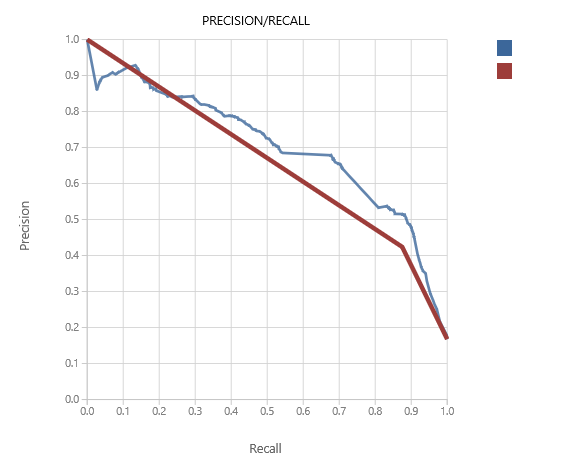
Score de RNASS

Pairé ou pas

1. **Courbe de ROC du modèle de RNASS et du modèle pairé / non pairé.** Le modèle de RNASS en plus d’être paramétrable au niveau du risque de ses prédictions performe mieux que le modèle basé sur l’état pairé ou non des nucléotides.

Lorsqu’on considère les nucléotides réactifs comme des positifs, la précision de ce modèle est de 42.4 % (la précision lorsqu’on considère les nucléotides peu réactifs comme positifs est de près de 97 %). Cela veut dire que 42.4 % les nucléotides non pairés sont des réactifs. Pour le même nombre de vrai « Hi », RNASS obtient une précision de 51 %. Cependant, le modèle de RNASS peut être ajusté. Par exemple, si l’on désire une précision de 85 % sur nos prédictions « Hi », on le peut, mais le rappel est de seulement 2 %. Voir le graphique précision en fonction du rappel de la figure 25.

Précision en fonction du rappel (« Hi »)



Rappel

Précision

Score de RNASS

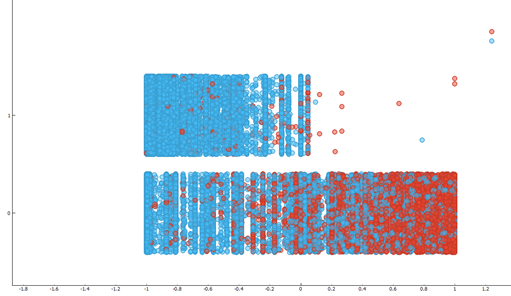
Pairé ou pas

1. **Précision en fonction du rappel du modèle de RNASS (en bleu) et de celui basé sur l’état pairé non pairé des nucléotides (en rouge).** Plus de 1 million de nucléotides ont servi à faire ce test. Chacune des 10 SS de RNAsubopt est prise séparément. Le modèle de RNASS performe mieux que celui basé sur l’état pairé ou non des nucléotides.

Chose surprenante, du point de vue du nombre des bonnes prédictions total divisé par le nombre de prédictions total (*accuracy*) du modèle prenant seulement en compte l’état pairé ou non d’un nucléotide, il est plus avantageux de prédire tous les nucléotides comme étant de basse réactivité plutôt que de se fier sur le pairage de ceux-ci. Autrement dit, le nombre de nucléotides non pairés et peu réactifs est tellement grand par rapport au nombre de nucléotides non pairés et réactifs que l’état non pairé aide peu à trouver les nucléotides réactifs (moins de 1 nucléotide non pairé sur 2 est réactif).

1. Comparaison entre le modèle basé sur l’état pairé ou non des nucléotides et le modèle de basé sur le score de RNASS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Métrique | État pairé ou non | RNASS | | |
| AUC | **0.818** | **0.908** | | |
| Accuracy | **0.780** | **0.818** | **0.890** | **0.843** |
| Recall | **0.875** | **0.900** | **0.683** | **0.070** |
| Precision | **0.424** | **0.477** | **0.668** | **0.909** |
| F1 Score | **0.571** | **0.624** | **0.676** | **0.130** |



1

-1

0

Score de prédiction de RNASS

Nucléotides pairés

Nucléotides non pairés

Nucléotides réactifs

Nucléotides peu réactifs

1. **Distribution des nucléotides en fonction de leur état (axe des ordonnés), de leur score de prédiction de RNASS (axe des abscisses) et de leur réactivité chimique (couleur).** Les points rouges représentent les nucléotides réactifs. Ils sont beaucoup plus présents lorsque le nucléotide est non pairé et lorsque son score de prédiction est élevé. Cette visualisation a été faite avec Orange, un logiciel d’apprentissage machine.

Le tableau IV montre quelques valeurs des métriques de performance du modèle basé sur le score de RNASS. Pour obtenir un taux de bonne prédiction de plus de 90% sur les nucléotides de haute réactivité, il faut diminuer le rappel en dessous du 1 %. Cela veut dire que moins de 1% des prédictions faites avec les données sur les MCN ont un taux de bonne prédiction sur le nombre de prédictions total au-dessus de 90 %. Les nucléotides de la figure 27 ont été divisés en deux groupes, les nucléotides pairés en haut et ceux non pairés en bas. Les nucléotides sont ordonnés selon leur score de prédiction, -1 à gauche jusqu’à 1 à droite. Finalement, les nucléotides rouges sont ceux qui ont un score de réactivité au-dessus de 1 et ceux en bleu ont un score de réactivité en dessous de 0.5.

## Conclusion du chapitre 3

Le score de prédiction de RNASS performe mieux que les prédictions faites avec l’état pairé ou non des nucléotides. Il se démarque par le fait qu’il peut prendre plus ou moins de risque dépendamment des besoins.

Au fur et à mesure que d’autres données sur les ARN seront recueillies, ce score pourra être amélioré. En effet, pour certains MCN le nombre de nucléotides est trop bas pour emmètre de bon jugement.

La qualité des prédictions des SS utilisés lors de l’entrainement de ce score est un élément crucial à son fonctionnement, c’est pourquoi un fichier *csv* comprenant les ARN avec leur niveau de cohérence accompagne ce mémoire.

# Conclusion

## Qu’est-ce qui fait qu’un nucléotide réagit?

Les règles permettant de prédire la réactivité d’un nucléotide sont beaucoup plus complexes que le simple fait qu’un nucléotide soit pairé ou non. Lorsqu’on considère seulement l’état pairé ou non des nucléotides, les nucléotides ayant une basse réactivité sont plus faciles à prédire que ceux ayant une réactivité élevée parce qu’il y en a beaucoup plus et que leur MCN. Le score de prédiction de RNASS n’est pas vraiment une prédiction dans la mesure qu’il est basé sur l’observation de cas identique à ceux prédits. Cependant, un nombre minimal d’observations assure sa puissance statistique.

Ce qui est admis en ce moment c’est que les nucléotides pairés ne réagissent pas sauf à de rares occasions et ceux non pairés réagissent. Pour les nucléotides pairés, c’est le cas la majorité du temps, mais pour les nucléotides non pairés, d’autres règles doivent être établies. Les règles proposées dans ce mémoire sont basées sur la structure locale du nucléotide, le MCN.

Le taux élevé de faux « Hi » nous amènent à faire de fausses prédictions lorsqu’on veut prédire la ou les SS d’un ARN en particulier. En prenant en compte les MCN cette erreur diminue. Il faut noter que l’information du pairage d’un nucléotide est inscrite dans le MCN ce qui lui assure de mieux performer sauf dans les cas où les données sur le MCN sont trop rares.

Statistiquement, plus un MCN est petit, plus il y aura d’occurrences de celui-ci dans la base de données et moins le nombre total de MCN sera grand. Ceci favorise les logiciels prédisant plusieurs petits MCN au lieu de quelques grands. MCFlashfold répond à ces exigences.

Avec mon approche, les ARN ont un score de cohérence basé sur la correspondance entre le niveau de réactivité de ses MCN et le niveau de réactivité observé

Mon algorithme s’appuie sur le fait que les logiciels de prédiction de SS font de bonnes prédictions en moyenne. Elle ne fait que vérifier la cohérence des logiciels avec eux même et lève des drapeaux lorsque ce n’est pas le cas.

## Champ d’étude à venir et algorithmes à considérer

Les liens entre les nucléotides distants sont difficiles à prédire, mais nous savons qu’ils existent, il est donc essentiel de laisser une marge de manœuvre aux modèles. Cette liberté laisse la place à des hypothèses qui pourront être réfutées ou confirmées. Dans un article d’Alain Denise[52], les liens tertiaires sont calculés en minimisant une fonction de coût qui oblige la majorité des boucles à faire une interaction avec un autre élément de l’ARN. Ce groupe a obtenu de bons résultats et leur méthode pourrait être appliquée à ce modèle ou vice et versa. Précisément, la non-réactivité anormale d’une boucle pourrait se voir attribuer une importance plus grande dans la fonction de coût, ce qui l’obligerait à être pairé.

Pour le moment aucun logiciel offert librement ne montre la dynamique d’un ARN à partir d’une SS. RDV permet d’imaginer cette fonction. La modification du champ de force en y ajoutant des contraintes physiques par d’autres moyens tels que la RMN combinée à une visualisation de la SS en 3D, nous donnera une meilleure idée des mouvements des molécules. Il est permis de croire que des hypothèses naîtront de cette visualisation. Le mécanisme des ARN catalysant des réactions chimiques serait un bon exemple à étudier.

Mon analyse s’inspire du processus scientifique. Les ARN et les SS prédites entrant dans la construction de la base de données des MCN sont fiables en raison des contrôles effectués et de leur nature même. Dans un prochain projet, il sera possible d’élargir l’espace de recherche, dans notre cas, apprendre des séquences plus difficiles à prédire. On pourrait par exemple abaisser le rapport signal sur bruit ou augmenter le taux d’adénine et le score moyen de réactivité des ARN.

Pour mon algorithme, ce qui détermine si une séquence est difficile à prédire ou cohérente, c'est la combinaison de sa composition en MCN et les valeurs de réactivité associées. Plus le nombre de MCN ayant une réactivité variable (autant de « Hi » que de « Low ») est grand, plus l'ARN est considéré comme difficile et obtient un score global bas. Ces ARN ont des mystères à élucider et beaucoup de questions restent sans réponse encore à ce jour. Par exemple, un ARN peut-il alterner entre deux SS rapidement comme on le voit dans la simulation de RDV?

Dans le futur, les expériences de réactivité chimique seront surement guidées par des algorithmes d’apprentissage machine. Un bon algorithme d’apprentissage par rétroaction devrait garder le nombre de MCN variable constamment petit, mais non nul. Ainsi, l’espace total des incertitudes se rétrécirait. Ce processus itératif tirerait avantage des « mauvaises prédictions », il les transformerait en occasions d’apprentissage. Ultimement, cet algorithme convergera vers un ensemble de règles qui expliquent tous les phénomènes de réactivités chimiques des ARN, mais pour atteindre ce but beaucoup de sondages chimiques des ARN restent à faire. Orienter ces expériences a été l'un des objectifs de mon approche. Les MCN ayant une faible occurrence et ceux variables doivent être plus étudiés.

Finalement, les interactions tertiaires doivent être prises en compte, un algorithme qui listerait les MCN distants entrainant une réactivité anormale pour chaque MCN variable aurait un bel avenir selon moi.

La liste des MCN avec le nombre de nucléotides contenues dans chacune des classes de réactivité chimique (basse, moyenne et élevé) est offerte avec ce mémoire en format csv.

De plus, une liste des ARN de l’ensemble filtré est placée dans un fichier csv en pièce jointe. Le fichier inclus la séquence, le vecteur de réactivité chimique, le score de cohérence pour les deux logiciels et les informations nécessaires pour le retrouver dans la RMDB.

# Bibliographie

1. Vangaveti, S., S.V. Ranganathan, and A.A. Chen, *Advances in RNA molecular dynamics: a simulator's guide to RNA force fields.* Wiley Interdisciplinary Reviews: RNA, 2017. **8**.

2. Deigan, K.E., et al., *Accurate SHAPE-directed RNA structure determination.* Proceedings of the National Academy of Sciences, 2009. **106**: p. 97-102.

3. Hajdin, C.E., et al., *Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots.* Proceedings of the National Academy of Sciences, 2013. **110**: p. 5498-5503.

4. Lucks, J.B., et al., *Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq).* Proceedings of the National Academy of Sciences, 2011. **108**: p. 11063-11068.

5. Yesselman, J.D., et al., *Updates to the RNA mapping database (RMDB), version 2.* Nucleic Acids Research, 2017: p. 1-5.

6. Dahm, R., *Friedrich Miescher and the discovery of DNA.* Developmental Biology, 2005. **278**: p. 274-288.

7. Caspersson, T. and J. Schultz, *Pentose Nucleotides in the Cytoplasm of Growing Tissues.* Nature, 1939. **143**: p. 602.

8. Watson, J.D. and F.H.C. Crick, *Molecular structure of nucleic acids*, in *Nature*. 1953. p. 737-738.

9. Jacob, F. and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins.* Journal of Molecular Biology, 1961. **3**(3): p. 318-356.

10. Holley Rw Fau - Apgar, J., et al., *STRUCTURE OF A RIBONUCLEIC ACID.* (0036-8075 (Print)).

11. Tinoco I Jr Fau - Uhlenbeck, O.C., M.D. Uhlenbeck Oc Fau - Levine, and M.D. Levine, *Estimation of secondary structure in ribonucleic acids.* (0028-0836 (Print)).

12. Nussinov, R., et al., *Algorithms for Loop Matchings.* SIAM Journal on Applied Mathematics, 1978. **35**(1): p. 68-82.

13. Brassard, G. and P. Bratley, *Fundamentals of Algorithmics*. 1995. p. 524.

14. Zuker, M., *On finding all suboptimal foldings of an RNA molecule.* 1989(0036-8075 (Print)).

15. Zuker, M. and D. Sankoff, *RNA secondary structures and their prediction.* Bulletin of Mathematical Biology, 1984. **46**(4): p. 591-621.

16. McCaskill, J.S., *The equilibrium partition function and base pair binding probabilities for RNA secondary structure.* (0006-3525 (Print)).

17. Major, F., et al., *The combination of symbolic and numerical computation for three-dimensional modeling of RNA.* Science (New York, N.Y.), 1991. **253**: p. 1255-60.

18. Hofacker, I.L., et al., *Fast folding and comparison of RNA secondary structures.* Monatshefte f??r Chemie Chemical Monthly, 1989. **125**: p. 167-188.

19. Serra, M.J. and D.H. Turner, *Predicting thermodynamic properties of RNA.* (0076-6879 (Print)).

20. Parisien, M. and F. Major, *The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data.* Nature, 2008. **452**: p. 51-55.

21. Mandal, M. and R.R. Breaker, *Gene regulation by riboswitches.* Nature reviews. Molecular cell biology, 2004. **5**: p. 451-463.

22. Almeida, M.I., et al., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2015. **33**: p. 4663-4670.

23. Weill, N., et al., *MiRBooking simulates the stoichiometric mode of action of microRNAs.* Nucleic Acids Res, 2015. **43**(14): p. 6730-8.

24. Patel, D.J., *Structural analysis of nucleic acid aptamers.* Current opinion in chemical biology, 1997. **1**: p. 32-46.

25. Bolton, E.E., et al., *PubChem3D: A new resource for scientists.* Journal of Cheminformatics, 2011. **3**: p. 1-15.

26. Staple, D.W. and S.E. Butcher, *Pseudoknots: RNA structures with diverse functions.* PLoS Biology, 2005. **3**: p. 0956-0959.

27. Zuker, M. and P. Stiegler, *Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.* Nucleic Acids Research, 1981. **9**: p. 133-148.

28. Gorodkin, J. and J.M. Walker, *RNA Sequence , Structure , and Function : Computational and Bioinformatic Methods IN Series Editor.*

29. Dallaire, P., *Une signature du polymorphisme structural d ’ acides ribonucléiques non-codants permettant de comparer leurs niveaux d ’ activités biochimiques par Résumé*. 2014.

30. Flamm, C., et al., *Barrier Trees of Degenerate Landscapes.* Zeitschrift für Physikalische Chemie, 2002. **216**: p. 155.

31. Weeks, K.M. and D.M. Mauger, *Exploring RNA structural codes with SHAPE chemistry.* 2012. **44**: p. 1280-1291.

32. Mathews, D.H., *Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization.* Rna, 2004. **10**: p. 1178-1190.

33. Mathews, D.H., et al., *Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure.* Proceedings of the National Academy of Sciences, 2004. **101**: p. 7287-7292.

34. Lorenz, R., I.L. Hofacker, and P.F. Stadler, *RNA folding with hard and soft constraints.* Algorithms for Molecular Biology, 2016. **11**: p. 8.

35. Kladwang, W., et al., *Standardization of RNA chemical mapping experiments.* Biochemistry, 2014. **53**: p. 3063-3065.

36. Seetin, M.G., et al., *Massively Parallel RNA Chemical Mapping with a Reduced Bias MAP-Seq Protocol.* 2014. **1086**: p. 95-117.

37. Kasner, E., et al., *The Mechanisms of RNA SHAPE Chemistry.* 2015. **70**: p. 646-656.

38. Mlynsky, V. and G. Bussi, *Molecular Simulations Reveal an Interplay Between SHAPE Reagent Binding and RNA Flexibilty.* The Journal of Physical Chemistry Letters, 2017: p. acs.jpclett.7b02921.

39. Lee, J., et al., *RNA design rules from a massive open laboratory.* Proceedings of the National Academy of Sciences, 2014. **111**: p. 2122-2127.

40. Das, R., *Eterna Forum.* 2014.

41. Darty, K., A. Denise, and Y. Ponty, *VARNA: Interactive drawing and editing of the RNA secondary structure.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2010. **7**: p. 309-322.

42. Kerpedjiev, P., S. Hammer, and I.L. Hofacker, *Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams.* Bioinformatics, 2015. **31**: p. 3377-3379.

43. leinbaum, D.G., M. Klein, and E.R. Pryor, *Logistic regression: a self-learning text.* 2002.

44. Steinwart, I. and A. Christmann, *Support Vector Machines.* 2008: p. 618.

45. Barros, R.C., A.C.P.L.F. de Carvalho, and A.A. Freitas, *Automatic Design of Decision-Tree Induction Algorithms.* 2015.

46. Friedman, J.H., *Greedy function approximation: A gradient boosting machine.* Annals of Statistics, 2001. **29**: p. 1189-1232.

47. Herbrich RHERB, R., et al., *Bayes Point Machines.* Journal of Machine Learning Research, 2001. **1**: p. 245-279.

48. POWERS, D.M.W., *Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation.* Journal of Machine Learning Technologies, 2011. **2**: p. 37-63.

49. d'Eterna, J.

50. Watkins, A., et al.

51. Marsaglia, G., W.W. Tsang, and J. Wang, *Evaluating Kolmogorov's Distribution.* Journal of Statistical Software, 2003. **8**: p. 1-4.

52. Lamiable, A., et al., *An Algorithmic Game-Theory Approach for Coarse-Grain Prediction of RNA 3D Structure.* 2013. **10**: p. 193-199.

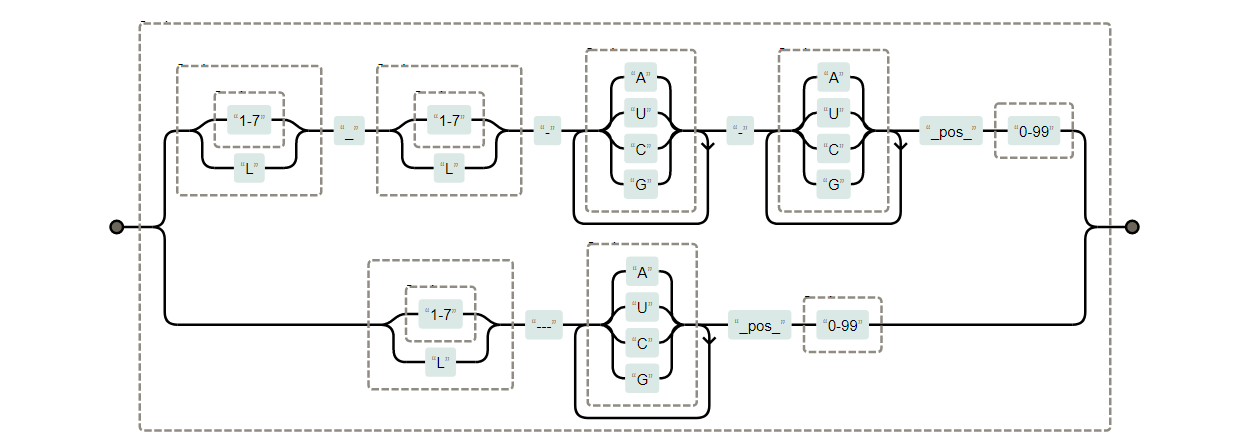
53. Lorenz, R., et al., *ViennaRNA Package 2.0.* Algorithms for Molecular Biology, 2011. **6**: p. 26.

# Annexe

## **Caractéristiques** du serveur utilisé

24 cores, 96GB : Intel(R) Xeon(R) CPU X5650  @ 2.67GHz

## Identification des MCN



1. Diagramme de composition des identificateurs des MCN fait avec regexper

## API d’Eterna

L’université Stanford rend disponible plusieurs ensembles de données par une interface de programmation (API) distincte de la RMDB, ils ont tous un identifiant sous forme de nombre. Pour obtenir les données brutes en format tsv (tab separated value), on remplace **« !id »** par l’identifiant dans cette URL :

* http://www.eternagame.org/tsv/synthesis!id.tsv

On peut aussi explorer les données avec l’interface d’Eterna en utilisant cette URL:

* http://www.eternagame.org/game/browse/!id

Pour retrouver notre ARN d’intérêt, on peut le chercher avec sa séquence.

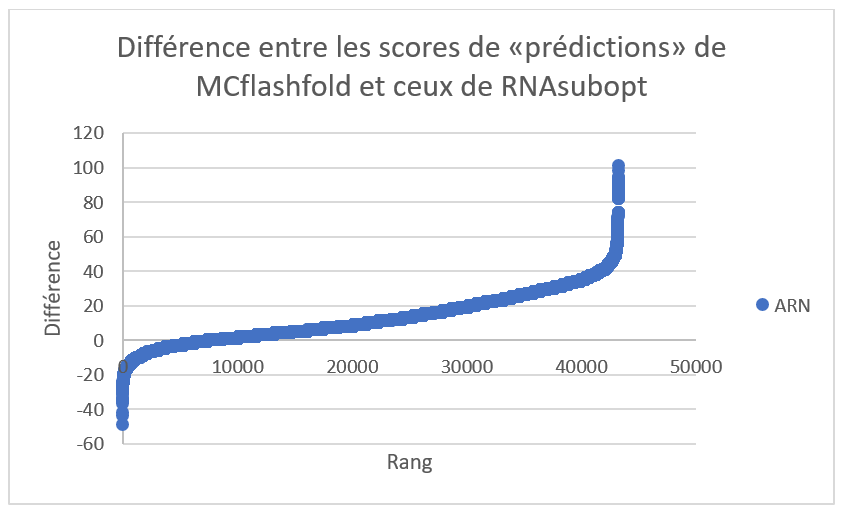
L’ensemble des identifiants des laboratoires est disponible à cette adresse :

* http://www.eternagame.org/get/?type=labs&size=400&skip=0

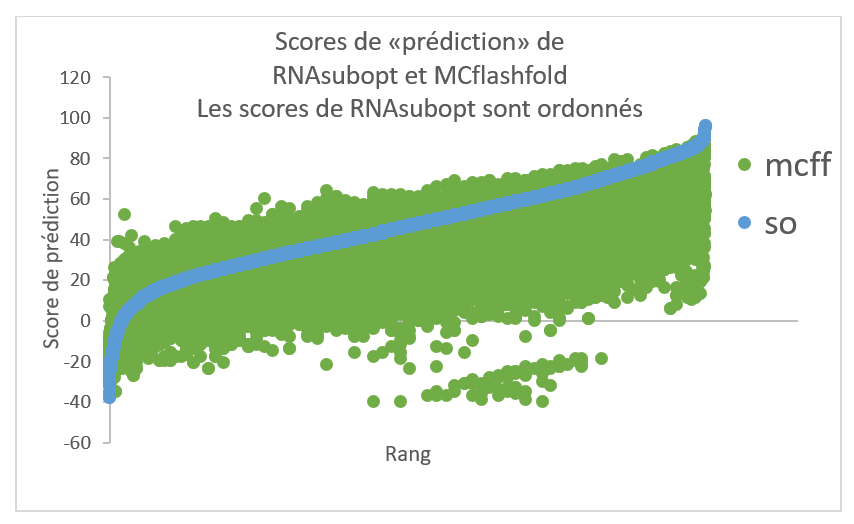
# Évaluation du score de prédiction

## Score de prédictions de la SS de de la *MFE* dans l’ensemble non filtré

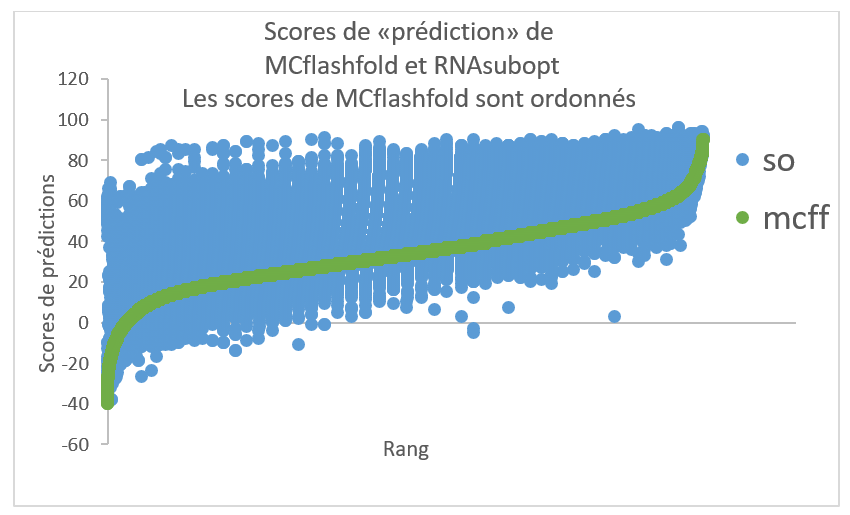
Les ARN de la ci-dessous sont classés par leur différence de score prédit entre les deux logiciels. On observe que les prédictions de MCFlashfold sont en moyenne moins précises que celle de RNAsubopt de 12.5 point. Cela signifie qu’en moyenne pour un ARN les prédictions de 12.5 nucléotides sont soit ratés par MCFlashfold pendant que RNasubopt ne se prononce pas, soit le contraire, soit un mélange des deux. Puisqu’avec ce nombre de données, il est rare que le logiciel ne se prononce pas, il est plus probable qu’en moyenne 6 prédictions de MCFlashfold soient ratées pendant que RNAsubopt en réussi 6 de plus. Les deux figures de la page suivante montrent les scores absolus. À tour de rôle, les données ont été ordonnées pour chaque logiciel. Le coefficient de corrélation de Pearson est relativement élevé, il est légèrement au-dessus de 0.76. Cela indique que certain ARN sont plus difficiles à prédire que d’autre peu importe la méthode.



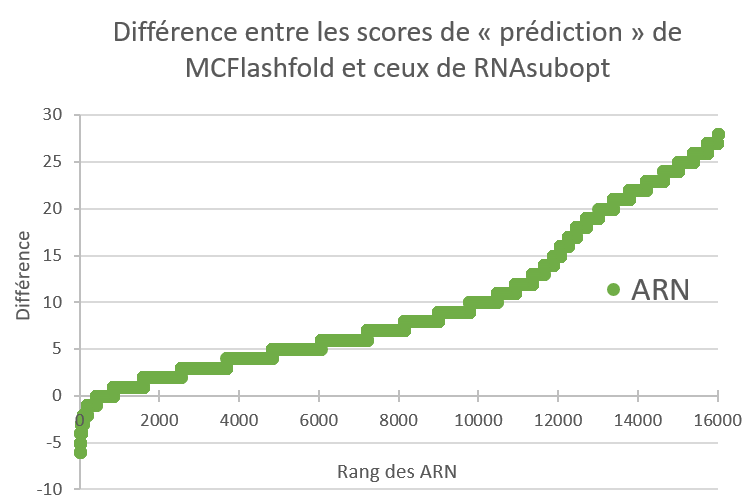
La différence de score de prédiction entre MCFlashfold et RNAsubopt est en moyenne de 12.5 point en faveur de RNAsubopt. Les pires ARN ont une forte concentration en adénine ou ont un score moyen élevé



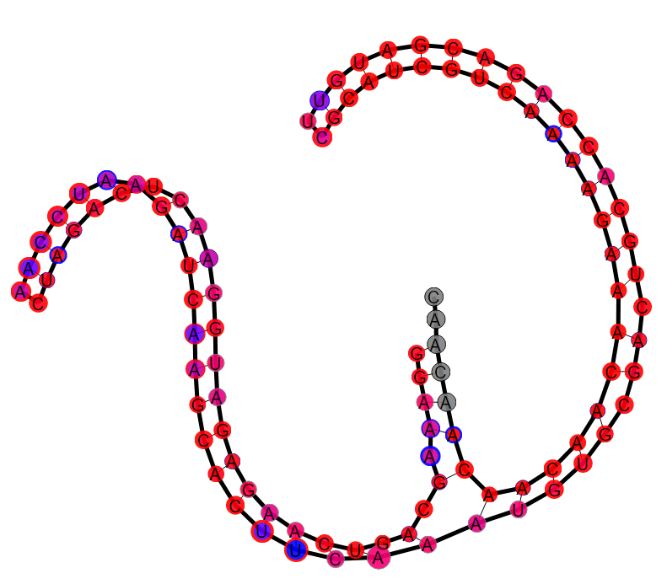
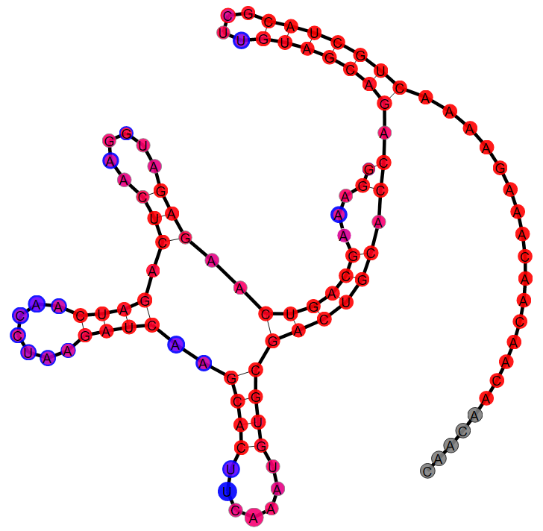
1. Le coefficient de corrélation de Pearson entre les scores de prédiction des deux logiciels est de 0.763083. Ceci indique une bonne corrélation entre les deux logiciels. Le score de MCFlashfold est en vert et celui de RNAsubopt en en bleu.



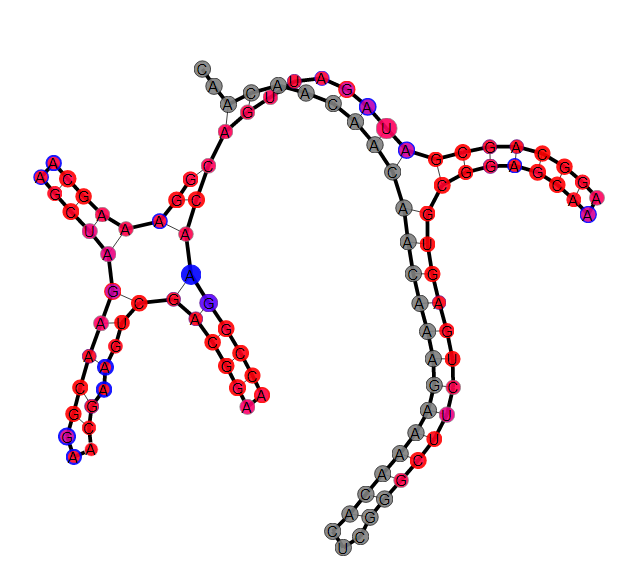
1. Cette figure est la figure 8 inversée. Ici, ce sont les données de MCFlashfold qui sont ordonnées. La corrélation est visible par le nombre presqu’inexistant d’ARN de RNAsubopt (bleu) en bas à droite

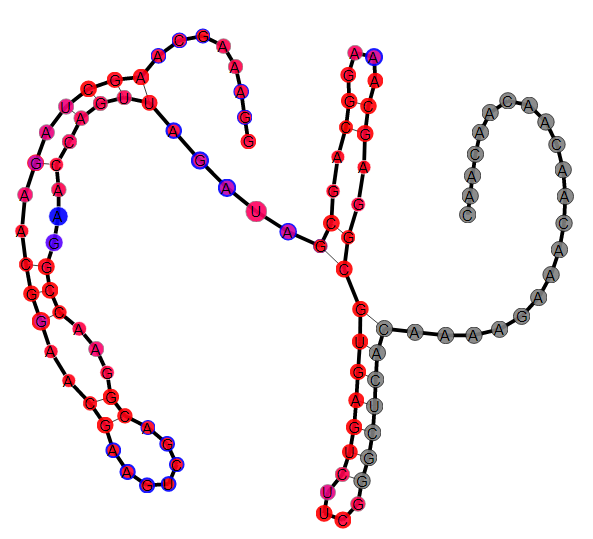


1. Après filtration des données, la moyenne des différences passe à 10.5 soit un gain de 2 points en faveur de MCFlashfold. L’étendue des différences s’est amoindrie aussi. À la défense de MCFlashfold, les données ont été calibrées à partir d’une structure en épingle prédite par un logiciel de la même suite que RNAsubopt, ViennaPackage[53].



1. Correspondance entre les valeurs de sondage chimique et les nucléotides d’un ARN repliés par RNAsubopt à gauche et MCFlashfold à droite. Le score de prédiction est de 46 contre 8 pour RNAsubopt. Lorsqu’on utilise le score pondéré, RNAsubopt obtient 66.62 et MCFlashfold 28.21.



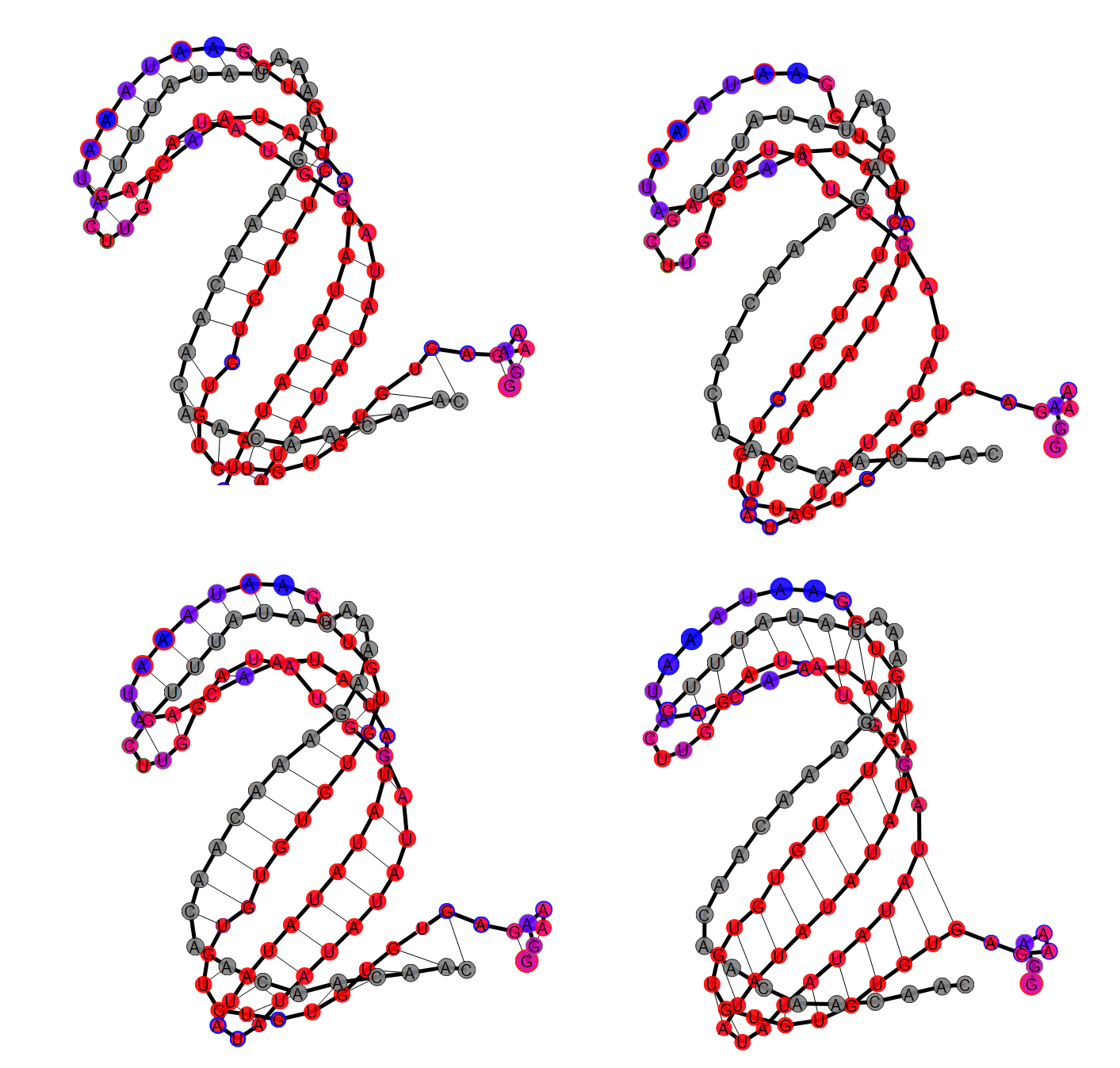


1. ARN ayant la différence de score la plus avantageuse pour MCflashfold. La différence entre les scores de prédiction pondérés est de 6.35. Les scores sont de 22.52 pour MCFlashfold et de 16.17 pour RNAsubopt. L’ARN du haut a été replié par MCFlashfold et celui du bas par RNAsubopt.

Ce tableau présente les MCN les plus souvent retrouvés dans la base de données. Le ratio donne une idée de la pureté du MCN. Un ratio de 1 signifie que tous les nucléotides du MCN sont réactifs, tandis qu’un ratio de -1 signifie que tous les nucléotides du MCN sont peu réactifs.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Nom du MCN** | **Nombre de nucléotides total so** | **Différence so** | **Nombre de nucléotides total mcff** | **Différence mcff** | **ratio So** | **ratio MCff** |
| 2\_2-GC-GC\_pos\_0 | 125470 | -112116 | 109185 | -97280 | -0.89357 | -0.89096 |
| 2\_2-GC-GC\_pos\_1 | 124541 | -112696 | 108655 | -97635 | -0.90489 | -0.89858 |
| 2\_2-UA-UA\_pos\_1 | 112327 | -86408 | 95831 | -74733 | -0.76925 | -0.77984 |
| 2\_2-UA-UA\_pos\_0 | 111662 | -92867 | 95730 | -79500 | -0.83168 | -0.83046 |
| 2\_2-AU-AU\_pos\_0 | 106443 | -76420 | 97478 | -66774 | -0.71794 | -0.68502 |
| 2\_2-AU-AU\_pos\_1 | 103962 | -88838 | 96290 | -77878 | -0.85452 | -0.80879 |
| 2\_2-GA-UC\_pos\_1 | 102951 | -73121 | 92271 | -64809 | -0.71025 | -0.70238 |
| 2\_2-UC-GA\_pos\_0 | 102619 | -93206 | 97151 | -86758 | -0.90827 | -0.89302 |
| 2\_2-GA-UC\_pos\_0 | 102365 | -79364 | 92251 | -69446 | -0.7753 | -0.75279 |
| contrôle visuel |  |  |  |  | -1 | 1 |

# Portrait d’un ARN



Séquence : GGAAAGAGUGUGUGUGUGUGUGUGUGUGUGGAAUAAAUAACAAUAUAUAUAUAUAUAUAUAUAUAUAGGUAUAGGUUCGUUUAUAUAAAAGAAACAACAACAACAAC

Score : [0.4927, 0.5641, 0.8659, 0.3409, 0.345, 0.3438, 0.1098, 0.0627, 0.0418, 0, 0.0209, 0, 0.0209, 0, 0, -0.0089, 0.0209, 0.0209, 0.0209, 0, 0.0626, 0.0417, 0, 0.0417, 0.0417, 0.0832, 0.0208, 0.0624, 0.0312, 0.0844, 0.3638, 1.1921, 1.1511, 0.8807, 0.8363, 1.2392, 1.0229, 0.7493, 0.8263, 0.2343, 0.1953, 0.919, 0.1515, 0.1045, 0.0421, -0.0196, 0.0748, 0.0374, 0.0374, 0.0747, 0.0187, 0.0932, 0.0076, 0.0373, 0.0372, 0.0558, 0.1116, 0.0557, 0.0186, 0.0743, 0.1298, 0.185, 0.0555, 0.1293, 0.1107, 0.1657, 0.2938, 0.4751, 0.2006, 0.3271, 0.308, 0.361, 0.0064, 0.0721, 0.2756, 0.6014, 0.113, 0.444, 0.3474]

### 1.5.5 Le graphe des transitions d’un ARN

Les deux algorithmes donnent en sortie des SS classées en ordre selon leur énergie libre, ordonnée de la plus basse vers la plus élevée. Les SS prédites sont plus ou moins différentes les unes des autres selon plusieurs aspects : le nombre d’hélices, la composition en bases des jonctions, le nombre et l’identité des paires de bases, etc. La plus grande différence est que *MCFlashfold* permet les paires de bases non canoniques tandis que *RNAsubopt* permet un type de paire de base non canonique : la paire GU. Une valeur de distance ou de similarité peut être calculée entre chacune des SS, ce qui permet par la suite une représentation sous forme de réseau. Ce réseau est nommé le graphe des transitions.

Dans sa forme la plus simple, le calcul de la distance entre deux structures se fait en dénombrant le nombre de paires de bases communes [28]. Plus ce nombre est élevé, plus les deux SS sont semblables. Une façon plus stricte en générale de relier les SS est de considérer le changement d’une seule paire de base, Dr Paul Dallaire en discute dans sa thèse [29] et fait référence à un article de Dr Christoph Flamm [30].

L’ensemble des structures (nœuds) et des liens entre les SS semblables se nomme : le graphe des transitions, puisque l'on fait l’hypothèse qu’une structure peut transiter en une autre lorsqu'elles sont reliées dans le graphe. À ma connaissance, les règles de ces transitions n’ont pas été testées de façon approfondie en laboratoire. Actuellement, ce réseau nous permet de comparer la diversité des SS prédites entre deux ARN différents et il nous donne une idée du paysage énergétique de la molécule.

En se repliant, un ARN emprunte des chemins dans ce réseau passant d’une forme de haute énergie à une forme de basse énergie. En clair, l’ARN tentera de minimiser son énergie en créant des structures stables telles que des paires de bases ou des boucles de basses énergies. L’ensemble des chemins menant vers un minimum local est appelé un puits d’énergie et l’une des causes du mauvais repliement d’un ARN est la présence de plusieurs puits d’énergie semblablent piégeant l’ARN dans un minimum local.

Une explication non validée de la difficulté à prédire la réactivité est l’absence de considération de la structure tertiaire, c’est-à-dire les interactions à longue distance.

### 1.5.8 Base de données RMDB

La « RNA mapping data base » est une base de données des expériences de sondages chimiques de l’ARN. Elle contient plusieurs types d’expériences. Certaines sont faites sur des ARN provenant de cellules vivantes et d’autres sont faites sur des ARN synthétisés en laboratoire. Elle fournit les séquences d’ARN avec une valeur associée aux nucléotides donnant la mesure de leurs réactivités. Les données utilisées dans ce mémoire sont normalisées sur une séquence formant une boucle ajoutée en 3`. Cette boucle a une réactivité connue et constante.

Lorsque la valeur de réactivité chimique est basse, cela signifie une absence de réactivité et une valeur haute signifie une réactivité élevée.

J’ai classé les nucléotides en deux groupes, les nucléotides qui réagissent peu, les « Low » et ceux réagissant fortement les « Hi ». Il y a une zone de réactivité qui n’est pas considérée, elle correspond aux nucléotides qui sont entre le « Low » et le « Hi ».

La RMDB contient 3 catégories d’expériences :

1. Des expériences publiées ou provenant du laboratoire du Dr Das.
2. Des expériences utiles aux puzzles sur l’ARN (RNA puzzles).
3. Des expériences reliées à Eterna.

Mes analyses ont été faites sur les données reliées à Eterna.

### 1.5.9 Eterna

Eterna est un jeu vidéo scientifique. Il propose aux joueurs des casse-têtes ayant pour thème l’ARN. Un des jeux consiste à trouver une séquence qui se replie en une SS donnée, c’est l’inverse du problème de repliement résolu par RNAsubopt et MCFlashfold.

### 1.5.10 Cloud Labs

Lorsqu’un joueur expérimenté a trouvé une séquence se repliant en la structure demandée, il peut la soumettre au laboratoire pour confirmer sa découverte. Le laboratoire synthétise l’ARN et établit son profil par sondage chimique.

### 1.5.11 Fonction d’évaluation des séquences soumises

Dans ce mémoire, pour évaluer les SS quant à leur degré de similarité avec les données de SHAPE, je me suis inspiré de la fonction de score d’Eterna. Dans ce jeu, les séquences soumises sont évaluées en fonction de la réactivité de leurs bases. Un seuil est déterminé et si le nucléotide est censé être pairé dans la SS cible, mais qu’il réagit à l’agent modificateur, un point est retiré du total. Dans le cas contraire, un nucléotide non pairé doit ne pas réagir du tout pour qu’un point soit retiré. Cette évaluation est très « gentille » pour le joueur, puisqu’un nucléotide tombant dans la zone grise n’entraine pas de pénalité et que les valeurs de seuil sont optimisées pour maximiser le score du joueur. Ceci est dû à l’incertitude de la vraie signification de la réactivité chimique d’un nucléotide. Cette valeur est nommée : « Structure mapping score » [39]. Je me suis inspiré de ce score pour distinguer les nucléotides réactis des nucléotides peu réactifs. La seule différence est que les seuils sont fixes.

Dans ce mémoire, vous constaterez que le concept pairé ou non pairé peut être précisé pour mieux évaluer la réactivité ou la non-réactivité d’un nucléotide

### 1.5.12 Mapseeker

*Mapseeker* est le logiciel faisant la correspondance entre les séquences déterminées lors de l’étape de séquençage et la valeur de réactivité pour chaque nucléotide de la séquence. Son fonctionnement est décrit dans [35]

#### 1.5.12.1 Prise en considération des décrochements naturels de la polymérase

Comme mentionné brièvement plus haut, un contrôle négatif est effectué pour chaque ARN sondé. Ceci permet de distinguer les arrêts de transcriptions non provoqués par l’agent modificateur. Par exemple une SS très stable peut empêcher la rétrotranscriptase de poursuivre sa transcription.

#### 1.5.12.2 Normalisation sur un segment connu

L’ajout d’une tige-boucle permet de mesurer chaque réaction de façon extrêmement précise. Dans le passé, comparer les expériences faites sur différents ARN était une tâche presque impossible. Il est maintenant beaucoup plus facile de le faire grâce à cet ajout.

#### 1.5.12.3 Le biais de ligation

Certaines séquences ont moins de chance de se faire transcrire par la polymérase, créant ainsi un biais. Lors de la calibration des méthodes de détermination de la réactivité chimique des ARN par séquençage, les auteurs de [35] ont remarqué que les données de réactivités normalisées adjacentes à la tige-boucle « GAGUA » sont plus élevées que ce qu’ils avaient découvert avec les expériences faites par électrophorèse capillaire. Ils ont déterminé un facteur pour corriger cette différence.

#### 1.5.12.4 Rapport « signal sur bruit » (*Signal to noise ratio*)

Cette valeur est calculée en divisant la moyenne des valeurs de réactivité par la moyenne des erreurs de tous les nucléotides d’un ARN [40]. Pour plus d’information sur l’erreur, on peut se référer à l’article de Matthew G. Seetin [36]. On y apprend entre autres que l’erreur est basée sur la loi de poisson.

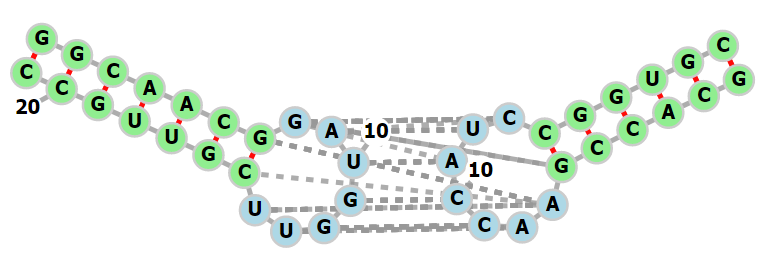
### 1.5.13 Outils de visualisation des structures secondaires

#### 1.5.13.1 VARNA

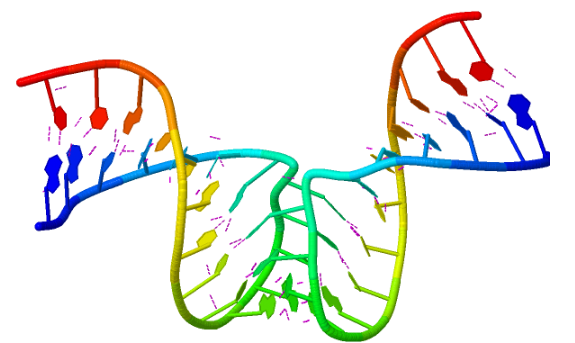
La base de données RMDB utilise un outil nommé VARNA pour représenter la structure secondaire et les données de sondage. Cet outil est écrit en java et le code source est offert gratuitement sous la licence GNU GLP. Trois limitations m’ont poussé à créer m’ont propre outil. Premièrement, Java contrairement à JavaScript n’est pas interprété directement dans le navigateur, il doit être compilé par une machine virtuelle non native et gérée à l’intérieur d’un « applet ». La deuxième raison est qu’il ne permet pas de visualiser le graphe des transitions des ARN. Finalement, il ne permet pas de mettre l’accent sur un nucléotide en particulier [41].

#### 1.5.13.2 FORNA

FORNA est un logiciel de représentation de la structure secondaire écrit en JavaScript et utilisant la même technologie que RNA Dynamic Viewer (RDV) (mon logiciel), soit d3js. FORNA est offert gratuitement sur *github* (un répertoire en ligne de code informatique basé sur git un logiciel de gestion des versions). La seule raison pour laquelle, je ne l’ai pas utilisé c’est qu’il n’existait pas lors du début de la conception de RDV, en 2014 [42].



1. Représentation en deux dimensions de deux ARN provenant de la « *protein data bank* » (PDB) par le logiciel FORNA. Les structures se nomment « *RNA LOOP-LOOP COMPLEX* ». Leur identifiant est : « 1BJ2 ».



1. Représentation en trois dimensions des ARN de la figure 10. Les pointillés mauves sont des ponts hydrogènes. La couleur des nucléotides est reliée à leur position allant du bleu (5’) au rouge (3’)

### 1.5.14 Azure

Azure est une plateforme développée par l’équipe de Microsoft. Elle est fournie gratuitement aux étudiants de l’Université de Montréal en version d’essais. Elle permet de faire de l’apprentissage machine de façon intuitive et graphique. Plusieurs algorithmes sont offerts et ils sont tous paramétrables. J’ai utilisé la plupart d’entre eux et j’ai comparé leur efficacité dans la tâche de prédire la réactivité d’un nucléotide sur différentes entrées. Un réglage automatique des hyper-paramètres est possible, ce qui permet une utilisation par un scientifique non expert en apprentissage automatisé. De plus, c’est une très bonne façon de s’instruire sur le sujet tout en testant les algorithmes.

### 1.5.15 Classifieurs

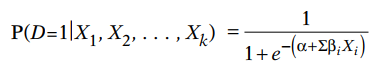
Les classifieurs sont des algorithmes qui prennent en entrée des éléments ayant tous les mêmes classes de caractéristiques et qui produit des catégories classant ces éléments. Les classifieurs suivants ont été utilisés pour prédire la classe de réactivité des nucléotides (basse ou haute).

#### 1.5.15.1 Réseau de neurones

Un réseau de neurones est une classe d’algorithme d’apprentissage machine complexe. En générale, les réseaux de neurones sont composés de nœuds et de liens entre les nœuds ayant chacun un poids. Ils sont faits de plusieurs couches de nœuds. Il y a la couche d’entrée où chaque nœud correspond à une dimension du problème à résoudre. Ensuite, il y a une ou plusieurs couches cachées. Dans un réseau de neurones classique, chaque nœud de la première couche cachée est relié à tous les nœuds d’entrées.

#### 1.5.15.2 Régression logistique

La régression logistique est un modèle d’apprentissage machine populaire dans lequel on tente de maximiser la vraisemblance d’un modèle logistique de la forme :



« D » est la variable à prédire et les « X » sont les caractéristiques données en entrée. Les paramètres à déterminer sont « α » et les « β ». Plus le « β » d’une variable est élevé plus l’importance de la variable est grande dans le modèle. Lorsque la variable est catégorique elle est transformée en variable binaire au début de l’algorithme [43].

#### 1.5.15.3 Machine à vecteurs de support.

Ce sont des algorithmes qui tentent de trouver un hyperplan qui sépare les deux classes à prédire. Pour avoir plus de détail sur les différents types de machine à vecteur de support ou « *support vector machine* » (SVM) en anglais, je vous conseille de lire [44] .

#### 1.5.15.4 Arbres de décisions

Les arbres de décisions sont des structures de prises de décisions extrêmement simples. Comme leur nom l’indique, leur structure est en forme d’arbre. À chaque nœud, une décision est prise, éliminant du même coup les ramifications non explorées. Lorsqu’il est binaire, seulement deux branches sont accessibles par nœud. Au bout de chaque chemin, constitué de nœuds et de branches, se trouve la réponse. Pour entrainer un arbre de décision, on génère des arbres aléatoirement et l'on compare leur succès par des métriques tel que ceux expliqués plus loin, dans la section : « Métriques de mesure de la performance des algorithmes d’apprentissage machine ». On peut aussi utiliser un algorithme génétique pour combiner les arbres qui performent bien. Pour plus de détail référez-vous à [45]

#### 1.5.15.5 Arbre de décision « *boosté* » ou « *Boosted Decision Tree* »

Cette méthode de classification s’appuie sur les arbres de décision. De façon itérative, on tente de diminuer l’erreur des arbres de décision modélisés au paravent. Pour plus de détails, veuillez-vous référer à [46].

#### 1.5.15.6 « *Bayes point machine* »

Cette technique de classification repose sur la probabilité d’observer une caractéristique étant donné une autre caractéristique. On ajoute à cela un biais pour maximiser l’algorithme. Pour plus de détails, veuillez-vous référer à [47].

### 1.5.16 Métriques couramment utilisées

Dans cette section plusieurs métriques sont expliquées. Grossièrement, une métrique permet de mesurer un phénomène. Les métriques de corrélation prennent en entrée deux vecteurs de la même taille, tandis que les métriques de mesure de performance des algorithmes d’apprentissage machine présentée ici prennent en entrée un classifieur binaire tels que ceux présentés plus haut.

#### 1.5.16.1 Corrélation de Pearson (corrélation linéaire)

Cette métrique permet d’établir une corrélation entre deux vecteurs de la même longueur. Chaque élément d’un vecteur doit être relié à l’élément de même indice de l’autre vecteur. Une valeur de 1 signifie une corrélation parfaite et une valeur de -1 signifie une corrélation inversée. Une valeur de 0 signifie pas de corrélation du tout. D’autre métriques de mesure de la corrélation existent, par exemple la corrélation de Spearman et la covariance.

#### 1.5.16.2 Métriques de mesure de la performance des algorithmes d’apprentissage machine

#### 1.5.16.2.1 Les vrais positifs

Lorsqu’on prédit des éléments pouvant se trouver dans deux classes distinctes, les positifs et les négatifs, les vrais positifs sont de vraies (bonnes) prédictions d’un élément positif.

#### 1.5.16.2.2 Les faux positifs

Tous comme les vrais positifs, les faux positifs sont un jugement sur la prédiction d’un élément pouvant être soit positif ou négatif. Les faux positifs sont donc de fausses (mauvaises) prédictions ayant été prédites dans la classe des positifs. Les prédictions sont faussement positives.

#### 1.5.16.2.3 Les vrais négatifs et les faux négatifs

Il suffit de remplacer le terme positif par le terme négatif dans les deux définitions précédentes pour avoir la définition des vrais et faux négatifs.

Dans le reste du mémoire, il sera question de vrais « Hi », faux « Hi », vrais « Low » et faux « Low », cela fait référence à la prédiction des classes discrètes de réactivité chimique des nucléotides. Les seuils sont de 0,5 et 1. ( « Hi » > 1 & « Low » < 0,5 )

#### 1.5.16.2.4 Le taux de vrais positifs (« *True positive rate* »)

Le taux de vrais positifs est le nombre de vrais positifs divisé par le nombre d’éléments positifs total. L’axe verticale de la courbe ROC est le taux de vrais positifs.

#### 1.5.16.2.5 Le taux de faux positifs (« *False positive rate* »)

Le taux de faux positifs est le nombre de faux positifs divisé par le nombre total d’éléments négatifs. Autrement dit, c’est le nombre d’éléments prédits positifs, mais qui sont en réalité négatifs divisé par le nombre total d’éléments négatif. L’axe des abscisses (horizontale) du plan de la courbe ROC est le taux de faux positifs. En général, moins la sensibilité ou le rappel est grand, moins le taux de faux positifs est grand, c’est ce que la courbe « ROC » illustre. Ceci est vrai pour les algorithmes tel que les classifieurs bayésiens naïfs puisqu’ils donnent une probabilité d’appartenance à une classe [41]. Plus on augmente le seuil de risque que prend l’algorithme, plus le nombre total de vrais positifs augmente, mais de moins en moins rapidement. L’augmentation du risque que prend l’algorithme entraine aussi le nombre de faux positifs à la hausse. Un algorithme risqué prédit plus d’éléments dans la classe des positifs qu’un algorithme peu risqué. Un bon algorithme à un rapport élevé entre le taux de vrais positifs et le taux de faux positifs.

#### 1.5.16.2.6 La précision

Cette valeur s’étend aussi de 0 à 1. La précision est le nombre d’éléments bien prédit divisé par l’ensemble des éléments prédits dans cette classe. Conceptuellement, moins l’algorithme prend de risque plus la précision devrait être élevée (moins de fausses prédictions). Cependant, le nombre total de vrais positifs est par le fait même diminué. Pour bien comprendre cette notion, on peut se placer dans le contexte d’un test diagnostic. Un jugement positif signifie que le patient a la maladie. Pour avoir un test précis, donc avec un taux de faux positifs bas, il faut diminuer le nombre de déclarations positives totales et se restreindre aux patients ayant des signes évidents. Dans le cas de la prédiction de la réactivité chimique des nucléotides, une précision élevée déclare comme réactifs une petite fraction des nucléotides qui le sont, mais se trompe rarement. Pour atteindre une plus grande précision, on doit donc simplement prendre moins de risques et prédire moins d’éléments de la classe d’intérêt.

#### 1.5.16.2.7 Le rappel ou sensitivité

C’est le nombre d’éléments bien prédit d’une classe divisée par le total des éléments de la classe. Un ratio de 1 signifie que tous les éléments de la classe ont été trouvés et bien prédits. Dans l’exemple d’un algorithme avec un seuil de risque au-dessus duquel il se prononce, il y a un compromis à faire entre le rappel et la précision. La solution la plus simple pour avoir un rappel de 1 est de prédire tous les éléments comme étant positifs. Cependant, la précision risque d’être moins grande.

#### 1.5.16.2.8 Score F1

Cette valeur considère le rappel et la précision dans son calcul. Plus précisément, le score F1 et la moyenne harmonique de la précision est du rappel. Plus la valeur est près de 1, meilleur est l’algorithme [48].

#### 1.5.16.2.9 Table de contingence

Cette table est une matrice n x n, « n » étant le nombre de classes possibles dans laquelle le nombre de vraies « classes i » et de fausses « classes i » est indiqué, i étant le nom des classes, voir tableau VIII. La table de contingence est une généralisation de la matrice de confusion.

#### 1.5.16.2.10 Courbe ROC

Lorsqu’un algorithme donne une probabilité d’appartenance à une classe donnée, un seuil est nécessaire pour départager les prédictions. Si le seuil est un nombre réel et qu’il peut être ajusté, on peut générer « une infinité » de table de contingence. La courbe ROC permet de voir un portrait d’ensemble de ces matrices. Chaque point de la courbe représente le taux de vrai positif en fonction du taux de faux positifs.

### 1.5.17 Autres logiciels utilisés

#### 1.5.17.1 Mongodb

MongoDb est une base de données non relationnelle (BDNR). À la différence des bases de données relationnelles (BDR), les BDNR peuvent contenir des structures de données en forme d’arbre. Chaque document ajouté à la base de données à un identifiant unique. Comme dans la plupart des BDR, quatre commandes de base permettent d’interagir avec la base de données : « insert, remove, update et find ».

On doit créer la base de données en lui donnant un nom. À l’intérieur de celle-ci se trouvent des collections, c’est dans ces contenants que sont placées les données. Chaque collection est indépendante. Pour accélérer la recherche de la valeur d’un des champs, on « index » ce champ avec la commande *createIndex*. MongoDB m’a permis de compiler les MCN comme vous le lirez dans le chapitre 1.

#### 1.5.17.2 Node.js

Node.js est un serveur web. Contrairement au serveur « Apache », il n’interprète pas le langage PHP, il interprète le JavaScript, ce qui est pratique dans la mesure où JavaScript est le langage interprété par les navigateurs. Dans mon cas, JavaScript a été essentiel pour la réalisation des interfaces interactive et dynamique. La combinaison de Node.js et de MongoDB est facile à mettre en place. RDV utilise ces deux technologies.

### 1.5.18 Formats de fichiers

RNASS produit en sortie des fichiers JSON qui peuvent être converti en fichier CSV pour faciliter leur étude ultérieure.

#### 1.5.18.1 JSON (JavaScript object notation)

Ce format est largement utilisé. Il permet de conserver des données dans une structure en forme d’arbre. Il peut être lu par des éditeurs de texte directement. Il est compatible avec la base de données Mongodb.

#### 1.5.18.2 CSV (comma separated value)

Ce format est très simple. Tout comme le format JSON, il peut être lu directement par un éditeur de texte. Il range les données sous forme de tables, c’est-à-dire une liste de rangées avec un nombre de champs définis formant des colonnes. Les rangées sont séparées par des sauts de ligne et les colonnes par des virgules. Il est compatible avec logiciel Excel de Microsoft. Les fichiers TSV (*tab separed value*) diffèrent des fichiers CSV par le remplacement des virgules par des tabulations.

### Le format RDAT

Les données de la RMDB sont disponibles en deux formats : le format ISATAB (non utilisé) et le format RDAT. Les spécifications de ces formats sont disponibles sur le site web de la RMDB[50]. Les fichiers RDAT sont générés par MapSeeker [36]. Un outil pour les interpréter est mis à notre disposition, il se nomme RDATKit. Cette librairie est écrite en *Python* et est particulièrement complexe. Un fichier RDAT correspond à une expérience distincte pouvant contenir des milliers de séquences sondées. Chaque fichier a une liste de « *constructs* », qui contient la majorité du temps un seul élément. Cet élément à un champ « data » contenant l’ensemble des données de réactivités, les données relatives aux erreurs entre les réplicas, les séquences et le « signal sur bruit ». Les séquences sont parfois plus longues que les vecteurs de réactivités et d’erreurs. Pour phaser les deux vecteurs, un champ nommé « offset » accompagne chaque séquence, les nucléotides avant cet indice n’ont pas de valeur de réactivité. Les conditions dans lesquels l’expérience a été menée, tel que la température et la concentration des espèces chimiques, sont disponibles dans la variable annotation de chaque séquence. De plus, c’est dans ce fichier qu’on retrouve le traitement numérique qui a été fait sur les données.