ALIFREEFOLDMULTI : UNE MÉTHODE SANS ALIGNEMENT POUR PRÉDIRE LES STRUCTURES SECONDAIRES D'ARN HOMOLOGUES

par

Marc-André Bossanyi

Mémoire présenté au Département d'informatique
en vue de l'obtention du grade de maître ès sciences (M.Sc.)

FACULTÉ DES SCIENCES

UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, 22 février 2021

Le 22 février 2021

Le jury a accepté le mémoire de Marc-André Bossanyi dans sa version finale

Membres du jury

Professeur Aïda Ouangraoua
Directrice
Département d'Informatique

Professeur Manuel Lafond
Membre interne
Département d'Informatique

Professeur Richard Egli
Président-rapporteur
Département d'Informatique

Sommaire

Le sujet de ce mémoire est la prédiction de structures secondaires de familles d'ARN homologues. La structure secondaire d'une séquence d'ARN non-codant définit généralement la fonction de cet ARN au sein de la cellule. Dans cette maîtrise, nous avons développé un nouvel algorithme sans alignement pour prédire la structure secondaire de chacune des séquences d'ARN d'une famille d'ARN non-codants. Cet outil, aliFreeFoldMulti, est une extension d'un outil qui a été précédemment développé au sein du laboratoire CoBIUS, soit aliFreeFold. Initialement, aliFreeFold permet de prédire une seule structure secondaire représentative pour une famille de séquences d'ARN homologues. Avec les algorithmes développés dans cette maîtrise, aliFreeFoldMulti a la capacité de retourner une structure secondaire prédite pour chacune des séquences d'ARN qui composent une famille. Quatre stratégies ont été développées afin d'explorer de nouvelles approches pour prédire des structures secondaires à partir d'aliFreeFold. En comparant l'outil aliFreeFoldMulti avec les différents outils existants permettant de faire de la prédiction de structures secondaires de plusieurs séquences d'ARN homologues, aliFreeFoldMulti est le plus rapide et retourne des scores du même ordre de grandeur que les autres méthodes et les scores maximaux les plus élevés. Une analyse approfondie des résultats d'aliFreeFoldMulti permet de mettre en évidence le potentiel des méthodes sans alignement pour la prédiction de structures secondaires d'ARN.

Mots-clés: ARN non-codant ; structure secondaire ; prédiction sans-alignement ; homologie

Remerciements

Je désire consacrer cette page spécialement à remercier les personnes importantes durant ma maîtrise.

Tout d'abord, je tiens à remercier énormément ma directrice de recherche Aïda Ouangraoua pour m'avoir permis de faire cette maîtrise dans son laboratoire de recherche CoBIUS, de m'avoir accordé son précieux temps, pour ses conseils et son soutien au travers de ma maîtrise.

Je voudrais remercier Yoann Anselmetti pour avoir contribué à la rédaction de mon premier article scientifique ainsi que de son temps et de sa patience lors de mes explications sur le fonctionnement d'aliFreeFoldMulti.

Je remercie Yanchun Qi pour m'avoir aidé durant la conception du serveur Web pour mon outil aliFreeFoldMulti.

Je remercie Manuel Lafond pour m'avoir invité à son cours d'algorithmique et de m'avoir donné le sourire tout au long de ma maîtrise.

Je remercie également les membres du laboratoire CoBIUS durant ma maîtrise, soit Davy, Esaie, Anaïs, Abigail, Ibrahim, Safa et Siham.

Je voudrais remercier spécialement Guillaume, mon petit frère qui m'a toujours considéré comme un modèle et qui a toujours eu confiance en moi et ce malgré la distance nous séparant.

Finalement, je tiens à remercier les membres du jury d'avoir accepté l'évaluation de mon mémoire.

Abréviations

A Adénine

ADN Acide désoxyribonucléique

ARN Acide ribonucléique

ARNnc ARN non-codant

ARNpi ARN PIWI

ARNr ARN ribosomal

ARNt ARN de transfert

C Cytosine

G Guanine

lARNnc Long ARN non-codant

MCC *Matthew Correlation Coefficient*

miARN Micro ARN

nt Nucléotide

pARNnc Petit ARN non-codant

pARNi Petit ARN interférant

pARNn Petit ARN nucléaire

pARNno Petit ARN nucléolaire

PPV *Positive Predictive Value*

SENS Sensitivity

T Thymine

Abréviations

U Uracile

Xist *X-inactive specific transcript*

Table des matières

Sommaire ii

Remerciements iii

Abréviations iv

Table des matières vi

Table des figures viii

1 Introduction générale 1

1.1 Le dogme central de la biologie moléculaire 1

1.1.1 La transcription 4

1.1.2 La traduction 6

1.2 Les ARN non-codants 8

1.2.1 Les ARN non-codants ménagers 10

1.2.2 Les ARN non-codants à potentiel de régulation 11

1.2.3 La structure d'un ARN non-codant 11

1.3 Prédiction de la structure des ARN non-codants 12

1.3.1 Prédiction pour une seule séquence d'ARN non-codant 14

1.3.2 Prédiction pour un ensemble de séquences d'ARN non-codants 16

1.3.3 Avantages et limites des différentes stratégies de prédiction . . 22

1.4 Hypothèse et objectif de la maîtrise 24

1.5 Structure du mémoire 25

2 Méthode «sans-alignement» de prédiction de structures secondaires 26
2.1 La méthode «sans-alignement» aliFreeFold 26
2.1.1 Le modèle des n-motifs 27
2.1.2 Fonctionnement d'aliFreeFold 28
2.2 aliFreeFoldMulti, une extension d'aliFreeFold 29
2.2.1 Stratégie «centroïde» 29
2.2.2 Stratégie «centroïde ajustée» 30
2.2.3 Stratégie «plongement des tiges» 30
2.2.4 Stratégie «plus proche du sous-optimal» 31
2.2.5 Résumé des résultats 31
2.3 Article «aliFreeFoldMulti : alignment-free method to predict secondary structures of multiple RNA homologs» 32

Conclusion 55

Table des figures

1.1 Cellule eucaryote et cellule procaryote 2
1.2 Dogme central de la biologie moléculaire 3
1.3 ADN et ARN 3
1.4 La transcription 5
1.5 La traduction 7
1.6 ARNm et ARNnc 8
1.7 Classification générale des ARNnc 9
1.8 Structure d'un ARNt 10
1.9 Structure d'un miARN 11
1.10 Structure secondaire d'un ARNnc 12
1.11 Algorithme de Fibonacci par récursivité et algorithme de Fibonacci par programmation dynamique 13
1.12 Exemple d'un alignment multiples de 5 séquences d'ARN ainsi que les différentes régions conservées 17
1.13 Exemple de graphe 20
1.14 Différentes stratégies de prédiction pour un ensemble de séquences d'ARNnc 21

2.1 Une structure secondaire d'ARN et ses différents éléments structuraux 27

Chapitre 1

Introduction générale

Dans ce chapitre, nous présentons les notions biologiques de base et les notions de prédiction de structures secondaires d'ARN utiles pour comprendre le problème traité dans ce mémoire. La première section porte sur l'ADN et l'ARN d'une cellule. La seconde décrit les ARN non-codants, leurs fonctions et leurs structures secondaires. La troisième section présente les différentes méthodes de prédiction de structures secondaires pour les ARN non-codants.

1.1 Le dogme central de la biologie moléculaire

Les organismes vivants sont classés en deux groupes distincts suivant si la cellule comporte un noyau ou non. Un procaryote est un organisme dont la cellule ne contient pas de noyau alors qu'un eucaryote est un organisme composé d'une ou de plusieurs cellules contenant un noyau. (*Voir Figure 1.1.*) Il est possible de séparer davantage le groupe des procaryotes en deux sous-groupes, soient les bactéries et les archées [48]. Les organismes vivants sont composés d'une ou de plusieurs cellules. Ces organismes sont classifiés en deux groupes distincts suivant le nombre de cellules qui les composent. Il y a les pluricellulaires comme les animaux qui sont composés de plusieurs cellules et les unicellulaires comme les bactéries qui sont composés d'une seule cellule [28]. La multicellularité est apparue plusieurs fois au cours du vivant,

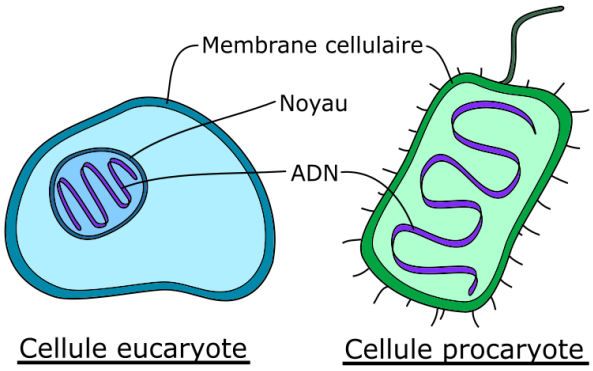

FIGURE 1.1 – Cellule eucaryote et cellule procaryote

mais uniquement chez les eucaryotes et pas chez les archées et les bactéries.

L'acide désoxyribonucléique (ADN) est le matériel génétique contenu dans chaque cellule de l'organisme. L'ADN d'une cellule procaryote se retrouve directement dans le cytoplasme de la cellule alors que l'ADN d'une cellule eucaryote se retrouve dans le noyau de la cellule. (*Voir Figure 1.1.*) Les gènes sont des segments de l'ADN permettant de produire des molécules nécessaires au fonctionnement de la cellule et de l'organisme. Le dogme central de la biologie moléculaire décrit le processus de production des protéines à partir des gènes en deux étapes, soient la transcription et la traduction. (*Voir Figure 1.2.*)

Dans la cellule, l'ADN se présente sous la forme stable d'une double hélice composée de deux chaînes de nucléotides s'appariant par complémentarité entre les bases azotées [46]. Les bases azotées composant l'ADN sont l'adénine (A), la cytosine (C), la guanine (G) et la thymine (T). L'étape de transcription permet de produire de l'acide ribonucléique (ARN) à partir d'un gène de l'ADN. L'ARN, comparativement à l'ADN, est composé d'un seul brin de nucléotides. Ce sont les mêmes bases azotées de l'ADN qui composent l'ARN à l'exception de la thymine (T) qui est remplacée par l'uracile (U) dans l'ARN. (*Voir Figure 1.3.*) Un nucléotide est l'association d'un groupe phosphate permettant de lier les nucléotides entre eux afin de former une chaîne nucléotidique, un sucre, soit le **d**ésoxyribose pour l'ADN et le **r**ibose pour l'ARN, et une base azotée permettant l'appariement de deux nucléotides par complémentarité.

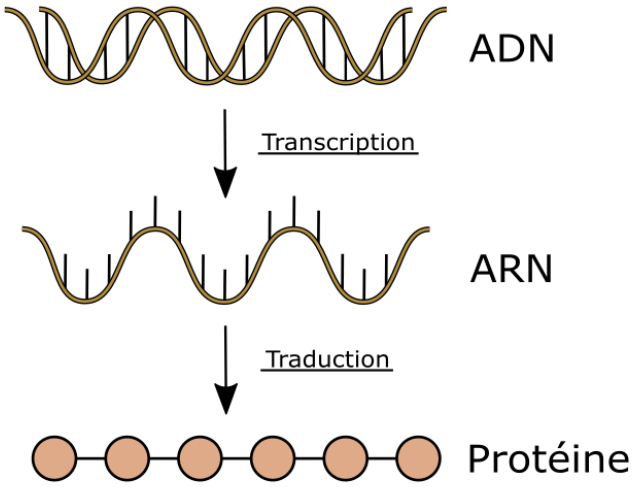

FIGURE 1.2 – Dogme central de la biologie moléculaire

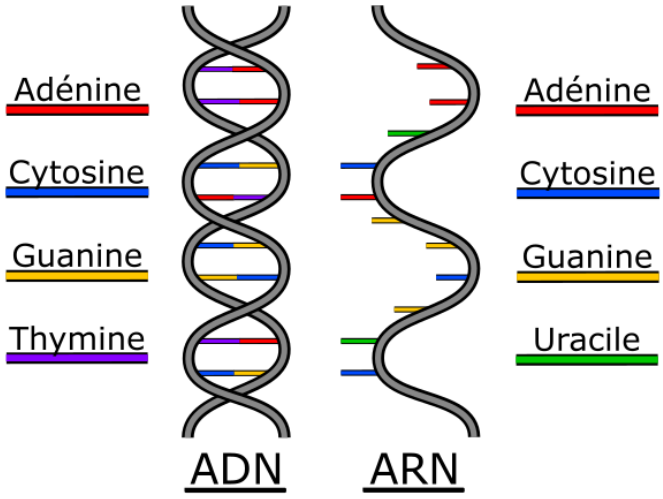

FIGURE 1.3 – ADN et ARN

1.1.1 La transcription

Lors de la transcription de l'ADN vers l'ARN, les facteurs de transcription et l'ARN polymérase jouent un rôle important. Les facteurs de transcription sont des protéines qui permettent de réguler la transcription d'un gène de l'ADN vers un ARN en se liant directement sur la séquence de l'ADN [23, 21]. Ils peuvent promouvoir, en tant qu'activateur, ou bloquer, en tant que répresseur, la liaison de l'ARN polymérase sur la séquence du gène [38]. L'ARN polymérase est un enzyme permettant de synthétiser de l'ARN à partir de l'ADN. Pour que la transcription se réalise, l'ADN doit se scinder en son milieu séparant ainsi les deux brins de l'ADN. Par la suite, un facteur de transcription se lie sur un site de liaison d'un des brins, appelé région promotrice, afin de permettre à l'ARN polymérase d'initier la transcription. L'ARN polymérase permet de produire soit des ARN codants pour des protéines, appelés ARN messager (ARNm), ou des ARN non-codants, tels que les ARN de transfert (ARNt), les ARN ribosomaux (ARNr) et les micros ARN (miARN). L'ARN produit par l'ARN polymérase n'est pas encore mature. Il s'agit d'un ARN primaire. Ce dernier va être modifié par des processus chimiques, par l'ajout d'une coiffe, d'une queue poly(A) et l'épissage chez les eucaryotes. Lors du processus de l'épissage, des segments appelés introns sont retirés de la séquence d'ARN et les segments restants de la séquence d'ARN appelés exons sont concaténés afin de produire un ARN mature. Cependant, il est important de préciser que certains exons sont éliminés lors de l'épissage lorsque ceux-ci sont situés entre 2 introns épissés, notamment lors de l'épissage alternatif. En résumé, l'étape de la transcription du dogme central de la biologie moléculaire permet de créer une séquence d'ARN codant ou d'un ARN non-codant à partir d'un gène. (*Voir Figure 1.4.*)

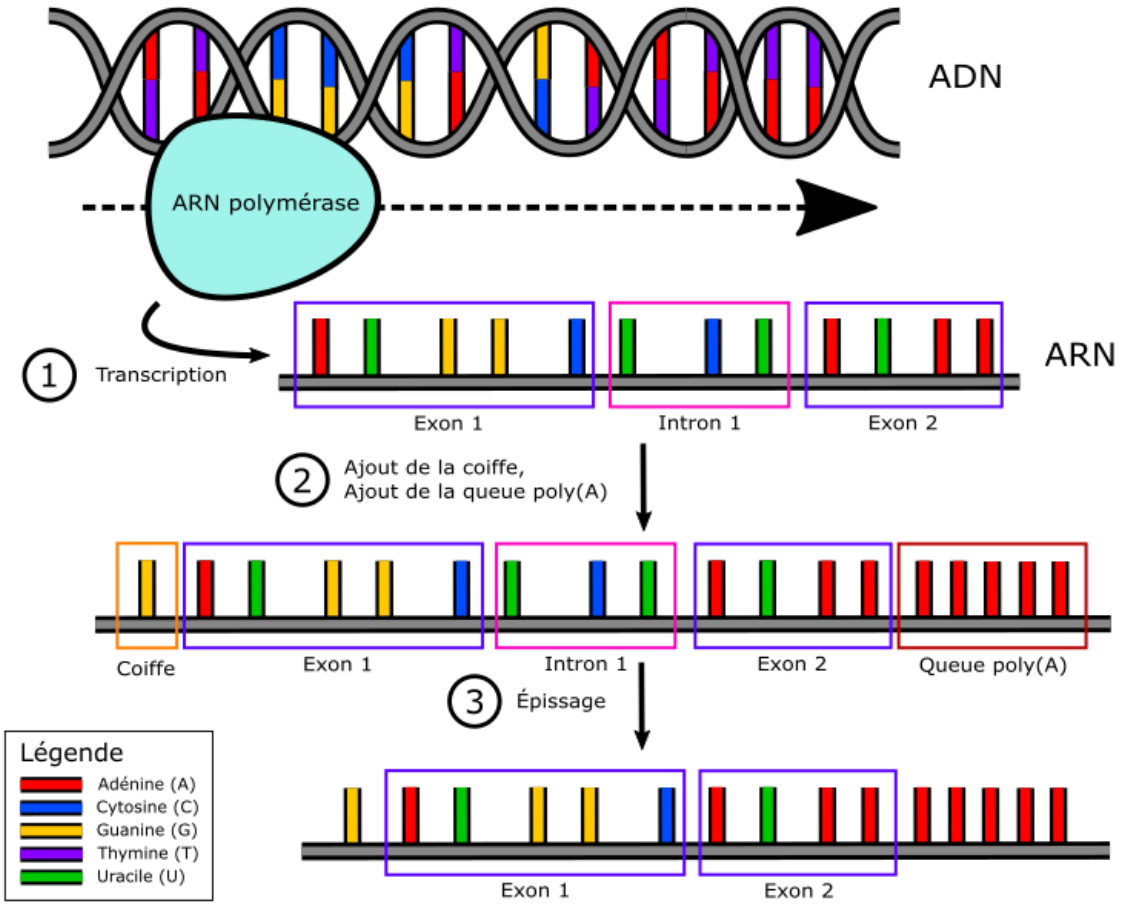

FIGURE 1.4 – L'ARN polymérase transcrit un gène de l'ADN pour produire un brin d'ARN (*étape 1*). Par la suite, L'ARN passe par l'épissage qui consiste à retirer les introns et concaténer les restants (*étape 2*). Finalement, dans le cas d'un pré-ARNm, on ajoute à celui-ci une coiffe et une queue poly(A) (*étape 3*).

1.1.2 La traduction

Dans le cas des ARN codants, la seconde étape du dogme central de la biologie moléculaire consiste à produire une protéine à partir d'un ARNm. Cette étape de traduction est réalisée par un ribosome. Le ribosome est un complexe ribonucléoprotéique composé de deux sous-unités, soient la grande sous-unité et la petite sous-unité. Ce complexe ribonucléoprotéique est constitué d'une combinaison d'ARN non-codants et de protéines. Le ribosome commence la traduction de l'ARNm par un codon de départ, AUG, soit un triplet de bases azotées. Les complexes de facteurs d'initiation et des facteurs d'élongation apportent un ARNt au ribosome en associant le codon de l'ARNm avec l'anticodon de l'ARNt. L'ARNt transporte un acide aminé spécifique à son anti-codon. L'acide aminé est lié sur la queue de l'ARNt grâce à l'enzyme aminoacyl-ARNt synthétase. Il y a un aminoacyl-ARNt synthétase pour chaque acide aminé et cet enzyme permet de lier un acide aminé à un ARNt selon l'anticodon de celui-ci [20]. Lorsque la traduction est initiée par le codon de départ, le ribosome continue la traduction sur l'ensemble de l'ARNm en apportant un nouvel ARNt pour chaque codon composant la séquence d'ARNm. Cet assemblage de plusieurs acides aminés s'appelle un polypeptide ou encore une protéine. Lorsque le ribosome rencontre un codon d'arrêt, soit UAA, UAG ou UGA, la traduction de l'ARNm est complétée et le ribosome relâche le polypeptide. En résumé, l'étape de la traduction du dogme central de biologie moléculaire permet de créer une protéine ou plusieurs protéines à partir d'un ARNm. (*Voir Figure 1.5.*)

Seulement 2% à 3% de l'ADN du génome humain est transcrit en ARNm codant pour des protéines [13]. Les études récentes d'annotations génomiques montrent que le catalogue des ARN codants pour des protéines est bien connu et apporte peu de nouveaux ARN codants. Par contre, elles permettent la découverte de nouveaux ARN non-codants inconnus qui augmente considérablement le catalogue des ARN non-codants connus [6, 33, 47]. Cette grande augmentation du nombre d'ARN non-codants inconnus montre l'importance des ARN non-codants dans le transcriptome humain.

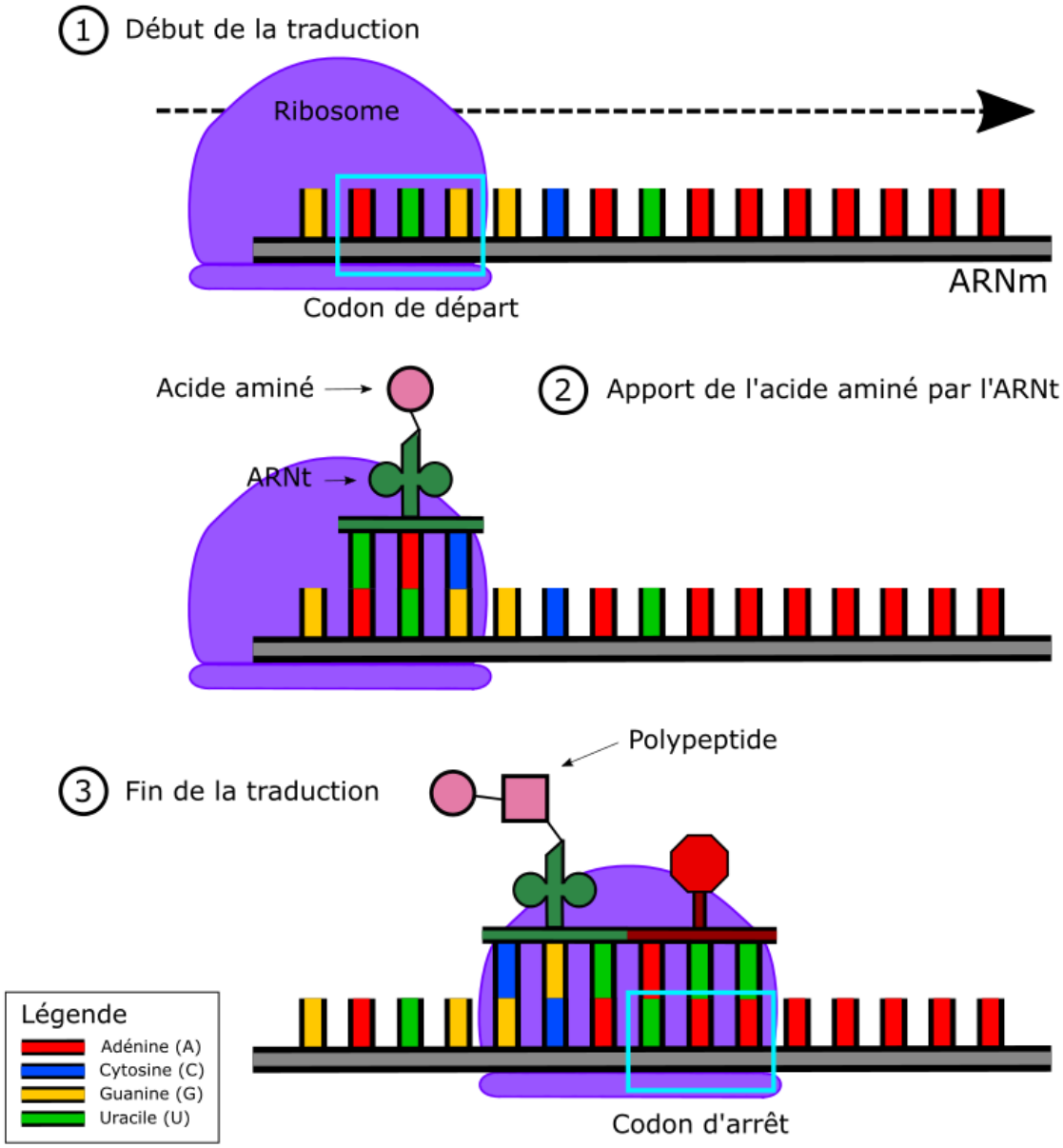

FIGURE 1.5 – La traduction de l'ARN débute lorsque le ribosome rencontre le codon de départ AUG (*étape 1*). Par la suite, un ARNt apporte un acide aminé en liant l'anticodon de cet ARNt sur le codon de l'ARNm (*étape 2*). Finalement, lorsque le ribosome rencontre un codon d'arrêt, la traduction se termine et l'enchaînement d'acides aminés, soit le polypeptide, est relâché (*étape 3*).

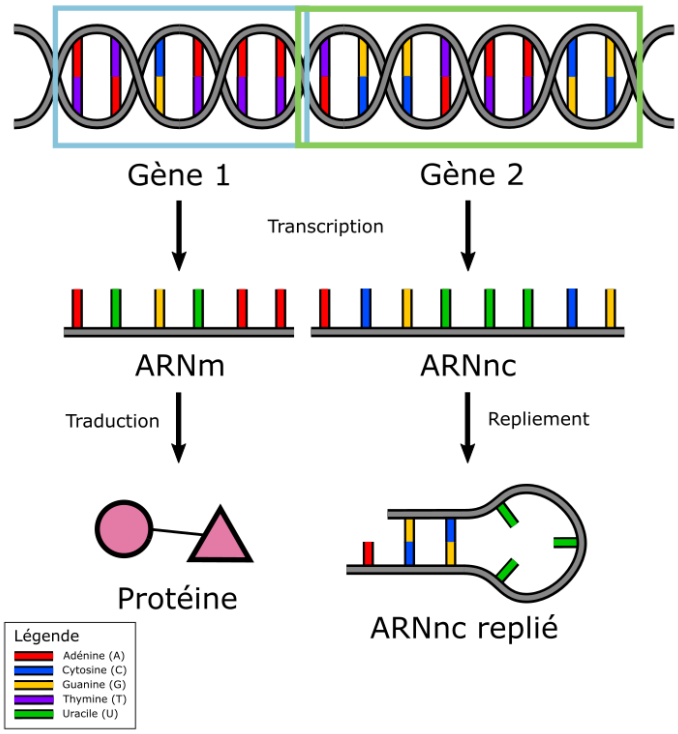

FIGURE 1.6 – ARNm et ARNnc

1.2 Les ARN non-codants

La transcription de l'ADN en ARN produit un ARN codant (ARNc) ou un ARN non-codant (ARNnc) à partir d'un gène. (*Voir Figure 1.6.*) Un ARNnc est un ARN fonctionnel qui n'encode pas de protéines [13, 51]. Dans une cellule, il existe plusieurs types d'ARNnc jouant de multiples rôles dans des processus biologiques, notamment dans la régulation de la cellule et le contrôle de l'expression d'un gène. Ainsi, il est possible de classifier les ARNnc en différentes catégories selon leur longueur et leur fonction dans la cellule [13, 39]. En se basant sur la longueur des séquences d'ARN, il existe deux principaux groupes. Le premier regroupe les séquences d'ARNnc ayant une longueur supérieure à 200 nucléotides, appelé les longs ARN non-codants (lARNnc) [13]. Le second regroupe les séquences d'ARNnc ayant une longueur inférieure ou égale à 200 nucléotides, appelé les petits ARN non-codants (pARNnc) [13]. Les lARNnc ont généralement des fonctions biologiques au sein de la cellule qui consistent à agir

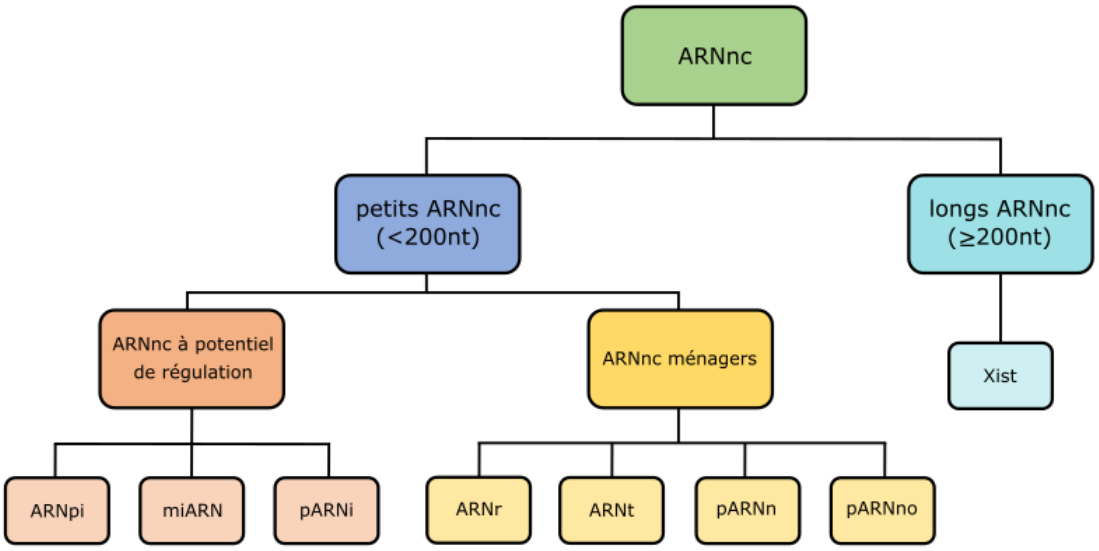

FIGURE 1.7 – Classification générale des ARNnc

comme des régulateurs cis- ou trans- dans processus biologiques [29]. Un régulateur cis- permet de faire de la régulation sur le même gène de l'ADN alors qu'un régulateur trans- permet de faire de la régulation sur des gènes distants de l'ADN. Par exemple, le Xist (*X-inactive specific transcript*) est un ARNnc se situant sur le chromosome X chez les mammifères placentaires permettant d'inactiver une des copies du chromosome X de la femelle [49]. Dans cette maîtrise, on s'intéresse particulièrement aux pARNnc et à la prédiction de leurs structures qui sont intrinsèquement reliées à leurs fonctions. Le groupe des pARNnc se subdivise en deux sous-groupes. Le premier sous-groupe correspond aux ARN non-codants permettant de maintenir le bon fonctionnement de la cellule appelés ARNnc ménagers (*housekeeping*) alors que le second sous-groupe correspond aux ARN non-codants permettant de contrôler l'expression d'un ou plusieurs gènes dans la cellule appelés ARNnc à potentiel de régulation (*regulatory potential*) [12]. (*Voir Figure 1.7.*)

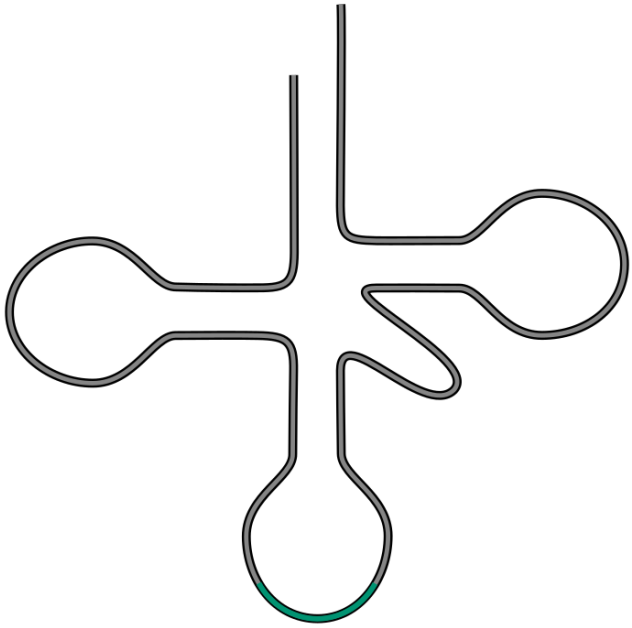

FIGURE 1.8 – Structure d'un ARNt

1.2.1 Les ARN non-codants ménagers

Le sous-groupe des ARNnc ménagers est composé d'ARN tels que les ARNr, les ARNt, les petits ARN nucléaires (pARNn) et les petits ARN nucléolaires (pARNno) [25]. (*Voir Figure 1.7.*) Les ARNr se lient avec les protéines ribosomales pour former la petite sous-unité et la grande sous-unité du ribosome et former un complexe ribonucléoprotéique. Ils sont les facteurs physiques et mécaniques du ribosome permettant de faire la traduction avec un ARNt et un ARNm. Les ARNt permettent de transporter un acide aminé afin de créer une chaîne d'acides aminés lors de la traduction d'un ARNm en protéine. (*Voir Figure 1.8.*) Les pARNn permettent de traiter l'ARN pré-messager dans le noyau de la cellule, réguler des facteurs de transcription et maintenir les télomères [15, 22]. Les pARNno sont divisés en deux classes, les «boîtes C/D» (*C/D box*), impliquées dans la méthylation, un procédé permettant d'ajouter un groupement méthyl sur un substrat ou une substitution d'un atome par un groupement méthyl ; les «boîtes H/ACA» (*H/ACA box*), associés avec la pseudouridine, un isomère du nucléoside uridine permettant de faire des modifications d'ARN [27, 45].

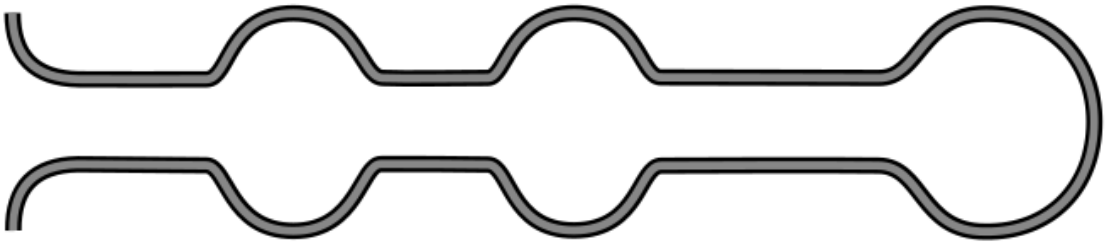

FIGURE 1.9 – Structure d'un miARN

1.2.2 Les ARN non-codants à potentiel de régulation

Le sous-groupe des ARNnc à potentiel de régulation est composé d'ARN tels que les miARN, les petits ARN interférant (pARNi) et les ARN piwi (ARNpi) [25]. (*Voir Figure 1.7.*) Les miARN permettent de faire la régulation négative et de faire la régulation post-transcriptionnelle de l'expression d'un gène [2]. (*Voir Figure 1.9.*) Les pARNi sont une classe d'ARN à double brin permettant d'interférer dans l'expression de gènes spécifiques avec la complémentarité des nucléotides de la séquence, en dégradant l'ARNm après la transcription pour empêcher la traduction de ces gènes [26]. Les ARNpi participent à l'épigénétique, qui permet de réguler l'expression des gènes sans modification et avec la molécule d'ADN, et la post-transcription négative des éléments transposables [41]. Un élément transposable est d'une séquence d'ADN permettant de créer ou d'inverser une mutation.

1.2.3 La structure d'un ARN non-codant

La structure d'une séquence d'ARNnc résulte d'un repliement de la séquence d'ARN par la complémentarité des bases azotées composant celle-ci. Les appariements de bases peuvent être canoniques de type Watson-Crick (A-U) et (C-G), ou non-canoniques de type (*wobble pairing*) (G-U) [7]. (*Voir Figure 1.10.*) La structure secondaire d'un ARNnc correspond à sa conformation dans un plan à deux dimensions alors que la structure tertiaire d'un ARNnc correspond à sa conformation dans un espace à trois dimensions au sein de la cellule. La quantité d'énergie libre d'une structure d'ARN dépend de la conformation de cette structure. En effet, plus la structure est compacte, soit qu'elle comporte beaucoup d'appariements de bases, plus l'énergie

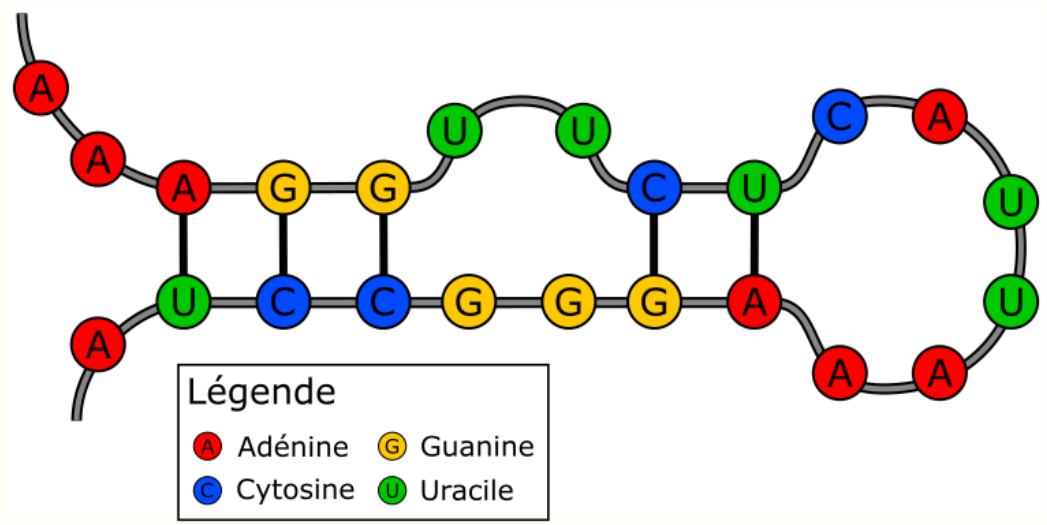

FIGURE 1.10 – Structure secondaire d'un ARNnc

libre est petite. Inversement, moins la structure est compacte, soit qu'elle comporte très peu d'appariements de bases et beaucoup de bases libres, plus l'énergie libre est grande. La fonction d'un ARNnc est généralement intrinsèquement liée à sa structure. C'est le cas par exemple de la forme de trèfle bien connue pour les ARNt. (*Voir Figure 1.8.*) Une famille d'ARNnc est un ensemble d'ARNnc descendant d'un même gène ancestral. Ces ARNnc sont dits homologues et ont généralement des rôles et des structures similaires. Par exemple, une famille d'ARNt contient des ARNnc ayant des structures secondaires de formes similaires à un trèfle.

1.3 Prédiction de la structure des ARN non-codants

Dans le cadre de cette maîtrise, le problème exploré est la prédiction de structures secondaires d'ARNnc. La prédiction de la structure secondaire des ARNnc est une étape importante pour la classification et l'analyse fonctionnelle des ARN. En effet, la structure secondaire d'une séquence d'ARN définit généralement le rôle de cet ARNnc au sein de la cellule. Le besoin de développer des méthodes de prédiction efficaces provient principalement de la taille importante et croissante de données de séquençage d'ARN générées pour diverse espèces du vivant. De plus, ce besoin provient du po-

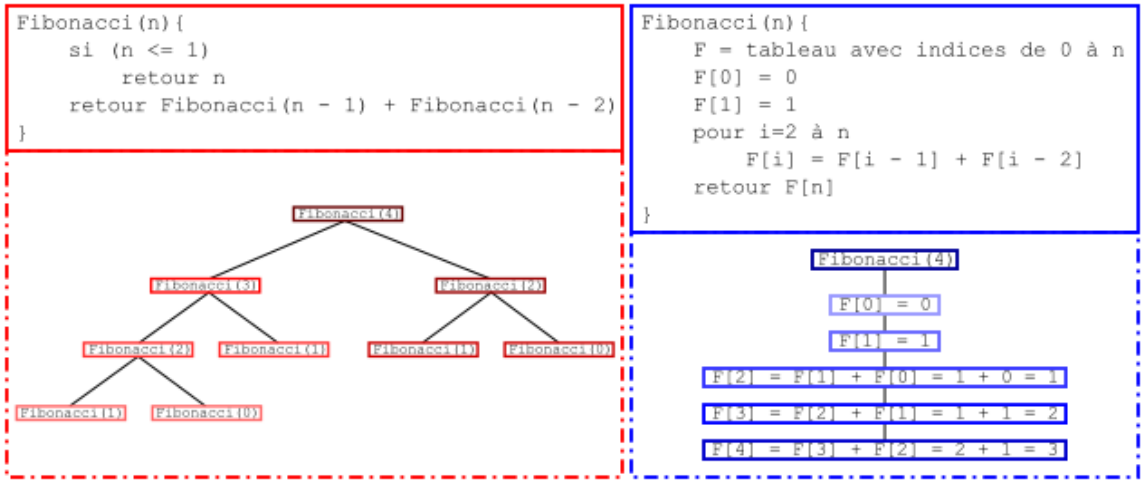

FIGURE 1.11 – Algorithme de Fibonacci par récursivité et algorithme de Fibonacci par programmation dynamique

tentiel de pouvoir faire des analyses de la structure des ARNnc, soient de l'évolution des structures, l'analyse structurelle d'interactions des ARNnc avec leur substrat ou d'autres molécules pour former des complexes, tels que le ribosome. Plusieurs algorithmes de prédiction de structures secondaires d'ARNnc utilisent la programmation dynamique. La programmation dynamique permet de résoudre un problème en calculant de façon itérative des solutions pour des sous-problèmes du problème initial de plus en plus larges jusqu'à résoudre le problème initial. Chaque solution d'un sous-problème est calculée à partir de solutions de sous-problèmes plus petits. De plus, lorsqu'une sous-solution est calculée, elle est stockée en mémoire dans une table appelée table de programmation dynamique pour y accéder plus tard en temps constant. Par exemple, pour calculer un terme quelconque de la série de Fibonacci en utilisant la programmation dynamique, la complexité en temps de l'algorithme est linéaire alors qu'en utilisant la récursivité, la complexité en temps est exponentielle. (*Voir Figure 1.11.*) Suivant le nombre de séquences d'ARN données en entrée, les méthodes de prédiction de structures secondaires d'ARNnc se divisent en deux principales classes. Ces deux classes sont la prédiction pour une seule séquence d'ARN non-codant et la prédiction pour un ensemble de séquences d'ARN non-codants.

1.3.1 Prédiction pour une seule séquence d'ARN non-codant

La première classe de méthodes contient les méthodes pour la prédiction de la structure d'une seule séquence d'ARNnc. Ces méthodes consistent à prédire une structure secondaire pour une seule séquence à la fois. L'algorithme de Nussinov, un algorithme utilisant la programmation dynamique, permet de trouver une structure secondaire pour une séquence d'ARNnc qui maximise le nombre de bases azotées appariées dans la structure et a une complexité en temps $O(L^3)$, où L est la longueur de la séquence en entrée [31]. Ainsi, considérant une séquence d'ARN S de longueur L avec les caractères de $S_1, ..., S_L$ et une fonction $a(i, j)$ calculant le score de l'appariement entre deux bases. Dans le cas où deux bases sont complémentaires, soient S_i et S_j, $a(i, j) = 1$ sinon $a(i, j) = 0$. Le score maximal $d(i, j)$ représentant le nombre de bases appariées pour une sous-séquence $S_i, ..., S_j$ est calculé par programmation dynamique comme suit :

Formules d'initialisation :

$$\begin{aligned} d(i, i-1) &= 0 \\ d(i, i) &= 0 \end{aligned}$$

Formule de récurrence :

$$d(i,j) = \max \begin{cases} d(i+1, j) \\ d(i, j-1) \\ d(i+1, j-1) + a(i,j) \\ \max_{i<k<j}[d(i,k) + d(k+1, j)] \end{cases}$$

Chaque sous-solution est issue du nombre d'appariements maximum pour une ou deux sous-séquences de la séquence S. Ainsi, les divisions de la séquence en sous-séquences non-chevauchantes à chaque étape permettent d'éviter les pseudo-noeuds causés par des appariements chevauchants. Cependant, le critère d'optimalité de l'algorithme de Nussinov est trop simpliste pour donner des prédictions de structure secondaire précises [11]. En effet, ce critère d'optimalité ne considère pas les limites et contraintes spatiales d'une molécule dans l'espace de recherche.

L'algorithme de Zuker, un autre algorithme utilisant la programmation dynamique, permet de trouver une structure secondaire optimale pour une séquence d'ARNnc en minimisant l'énergie libre de cette structure et a une complexité en temps de $O(L^4)$, où L est la longueur de la séquence en entrée [53]. L'énergie de la meilleure structure sur i, j, soit $W(i, j)$, est calculée par programmation dynamique comme suit :

Formule d'initialisation :

$$W(i, j) = 0$$

Formule de récurrence :

$$W(i, j) = \min \begin{cases} W(i, j-1) \\ \min_{i \leq k < j-m} W(i, k-1) + V(k, j) \end{cases}$$

où m représente la distance minimale entre 2 bases appariées

L'énergie de la meilleure structure sur i, j sachant que i, j sont appariées, soit $V(i, j)$, est calculée par programmation dynamique comme suit :

Formule d'initialisation :

$$V(i, j) = \infty$$

Formule de récurrence :

$$V(i, j) = \min \begin{cases} eH(i, j) \\ V(i+1, j-1) + eS(i, j) \\ \min_{i < i' < j' < j} V(i', j') + eL(i, j, i', j') \\ \min_{k, i < i_1 < j_1 < \ldots < i_k < j_k < j} eM(i, j, i_1, j_1, ..., i_k, j_k) + \sum_{1 \leq k' \leq k} V(i_{k'}, j_{k'}) \end{cases}$$

Dans la formule de récurrence pour $V(i, j)$, $eH(i, j)$ représente le coût de l'épingle à cheveux, $eS(i, j)$ représente le coût d'une tige, $eL(i, j, i', j')$ représente le coût d'une boucle et $eM(i, j, i_1, j_1, ..., i_k, j_k)$ représente le coût d'une boucle multiple. Ces éléments structuraux sont définis dans le chapitre 2.

L'énergie libre d'une structure d'ARN est déterminée par la somme des énergies pour les appariements de bases azotées et les énergies pour les bases azotées libres composant cette structure. En effet, les appariements de bases azotées permettent de réduire l'énergie libre de la structure ainsi que de la stabiliser alors que les bases azotées libres qui ne sont pas appariées par une base azotée complémentaire augmentent l'énergie libre de la structure et la rendent moins stable. D'autres outils ont été développés pour la prédiction d'une seule séquence en utilisant des paramètres thermodynamiques, comme RNAStructure[37] et Sfold[9], des méthodes statistiques avec poids, comme Simfold[1] et ContextFold[50] ou des modèles probabilistes, comme G1-G8[10] et TORNADO[36], qui sont des grammaires hors-contexte [35].

1.3.2 Prédiction pour un ensemble de séquences d'ARN non-codants

La seconde classe de méthodes contient les méthodes pour la prédiction de la structure de plusieurs séquences d'ARN homologues. Elles utilisent une approche comparative qui consiste à prédire une structure secondaire consensus pour l'ensemble des séquences d'ARNnc composant une famille d'ARN homologues. L'approche comparative comporte quatre stratégies.

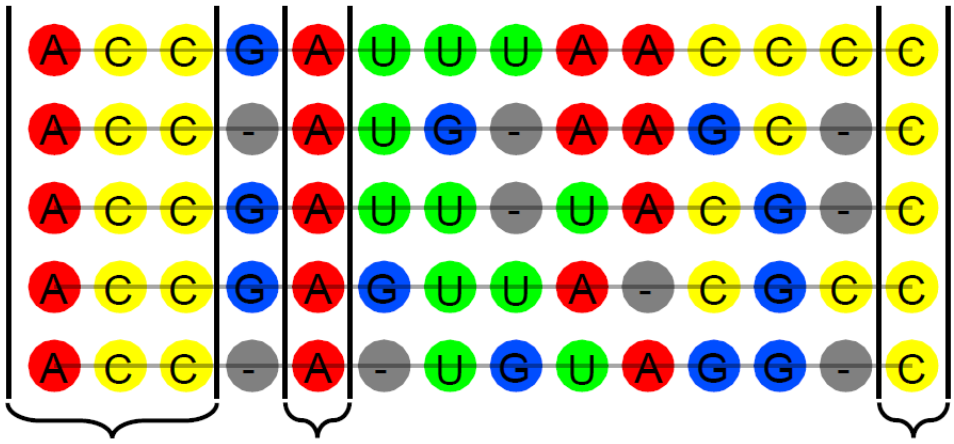

FIGURE 1.12 – Exemple d'un alignment multiples de 5 séquences d'ARN ainsi que les différentes régions conservées

Stratégie «aligner-et-replier»

La première stratégie appelée «aligner-et-replier» (*align-and-fold*) consiste à résoudre simultanément l'alignement des séquences d'ARNnc et le repliement des séquences sous forme de structure secondaire en utilisant un alignement multiple de type séquence-structure. (*Voir Figure 1.14.*) Un alignement entre deux ou plusieurs séquences permet de faire ressortir les régions homologues ou similaires de cet ensemble de séquences en plaçant les séquences les unes au-dessus des autres. Dans un alignement de séquences, les longueurs des séquences en entrée peuvent varier. (*Voir Figure 1.12.*)

Pour la stratégie «aligner-et-replier», plusieurs outils s'inspirent de l'algorithme exact de Sankoff qui permet de calculer l'alignement et le repliement de structures secondaires d'ARN. Cet algorithme utilise une approche par programmation dynamique basée sur une fonction objective représentée par un compromis entre l'énergie libre minimale d'une structure secondaire et le coût de l'alignement multiple de séquences d'ARN [40]. Dans cette même stratégie, l'outil FoldalignM[43] permet de faire la prédiction de structures secondaires en faisant simultanément le repliement et l'alignement par paires de séquences d'ARN. Cet outil effectue un clustering des séquences en se basant sur un alignement des séquences pour calculer des scores de similarités par paire, suivi d'alignements multiples des structures des candidats trouvés

à partir des analyses par paires. La version par paire (*pairwise*) utilisée par Foldalign emploie un système de score tenant compte de l'énergie de repliement, la substitution de paires de bases et des nucléotides à maximiser. L'alignement multiple de structures est calculé par les matrices de probabilité de base-paire en utilisant Foldalign, soit la version précédant FoldalignM. Le clustering de FoldalignM permet de regrouper les candidats qui sont similaires dans la structure et la séquence basé sur les scores par paires tous-contre-tous (*all-against-all*) [43]. L'outil TurboFoldII[42] permet de faire la prédiction de structures secondaires en faisant simultanément le repliement et l'alignement de séquences d'ARN par probabilités des positions des paires de nucléotides dans les séquences en incorporant les informations de l'identité de la séquence et des structures secondaires. L'alignement multiple de séquences est obtenu de ces probabilités en utilisant une transformation de cohérence probabiliste et un arbre guide hiérarchique et les probabilités de coïncidences postérieures sont obtenues avec un modèle de Markov caché pour les alignements par paires [42].

Stratégie «aligner-puis-replier»

La seconde stratégie appelée «aligner-puis-replier» (*align-then-fold*) consiste à résoudre séparément l'alignement des séquences d'ARNnc et le problème du repliement des séquences. L'alignement des séquences d'ARN se fait en premier, puis la solution de cet alignement est utilisée pour résoudre le problème du repliement des séquences. (*Voir Figure 1.14.*)

Pour la stratégie «aligner-puis-replier», l'outil RNAalifold[4] permet de faire la prédiction d'une structure secondaire consensus par une minimisation de l'énergie libre de la structure pour l'ensemble de la famille à la suite d'un alignement multiple des séquences d'ARN homologues de la famille [24]. De plus, cet outil introduit une manipulation plus rationnelle de l'alignement de gaps et utilise un modèle de covariance permettant de pondérer les matrices [4]. L'outil CentroidAlifold[19] permet de faire la prédiction d'une structure secondaire consensus par une précision maximale attendue (*maximum expected accuracy*) de la structure. Ce principe de précision maximale attendue donne de puissants estimateurs pour des problèmes d'estimation en bio-informatique tels que la prédiction de structures secondaires d'ARN et de la prédiction d'une structure secondaire commune d'un alignement multiple de séquences

d'ARN [19]. De plus, cet outil utilise une distribution de probabilité de structures secondaires communes du résultat de l'alignement de séquences d'ARN en entrée et une distribution de probabilité de structures secondaires pour chacune des séquences d'ARN dans l'alignement [19].

Stratégie «replier-puis-aligner»

La troisième stratégie appelée «replier-puis-aligner» (*fold-then-align*) consiste à résoudre séparément le problème du repliement des séquences et l'alignement des structures secondaires d'ARNnc. Le problème du repliement des séquences d'ARN se fait en premier, puis la solution de ce repliement est utilisée pour résoudre l'alignement des structures. (*Voir Figure 1.14.*)

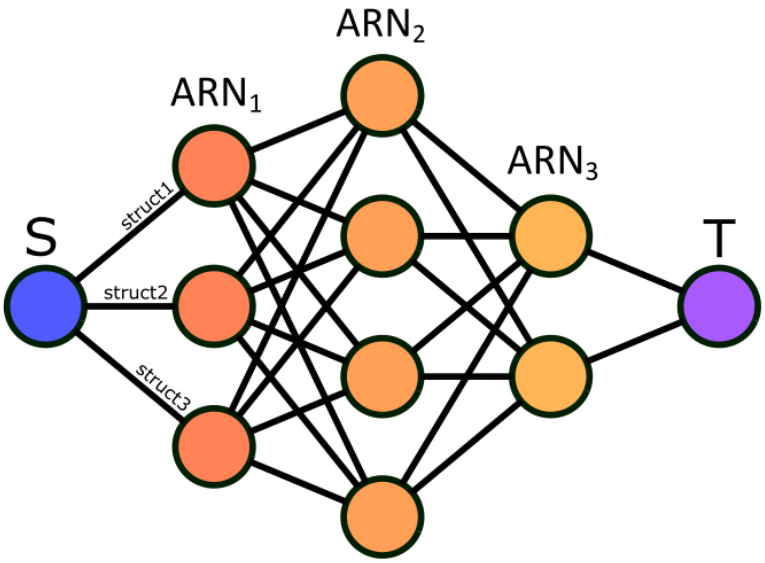

FIGURE 1.13 – Exemple de graphe

Pour la stratégie «replier-puis-aligner», l'outil RNAspa[18] permet de faire la prédiction d'une structure secondaire pour chacune de séquences d'ARN composant la famille en trouvant le chemin le plus court, soit celui composé des structures les plus similaires, dans un graphe dont les sommets sont les structures secondaires sous-optimales minimisant l'énergie libre. Il est important de préciser que ces structures secondaires sont sous-optimales dans le sens où elles ne minimisent pas nécessairement l'énergie libre de la molécule. En effet, il se pourrait qu'une structure secondaire minimisant l'énergie libre qui est optimale ne soit pas la structure réelle de l'ARN, alors qu'une structure sous-optimale l'est. En considérant la figure 1.13, la source, soit le rond bleu, est les séquences d'ARN en entrée. Les trois colonnes de ronds orange sont les structures secondaires sous-optimales pour les séquences d'ARN en entrée. Les arêtes du graphe, soient les traits reliant un rond avec un autre, sont les coûts de transition entre un état vers un autre. Le puit, soit le rond mauve, est l'ensemble regroupant les structures secondaires des séquences d'ARN en entrée.

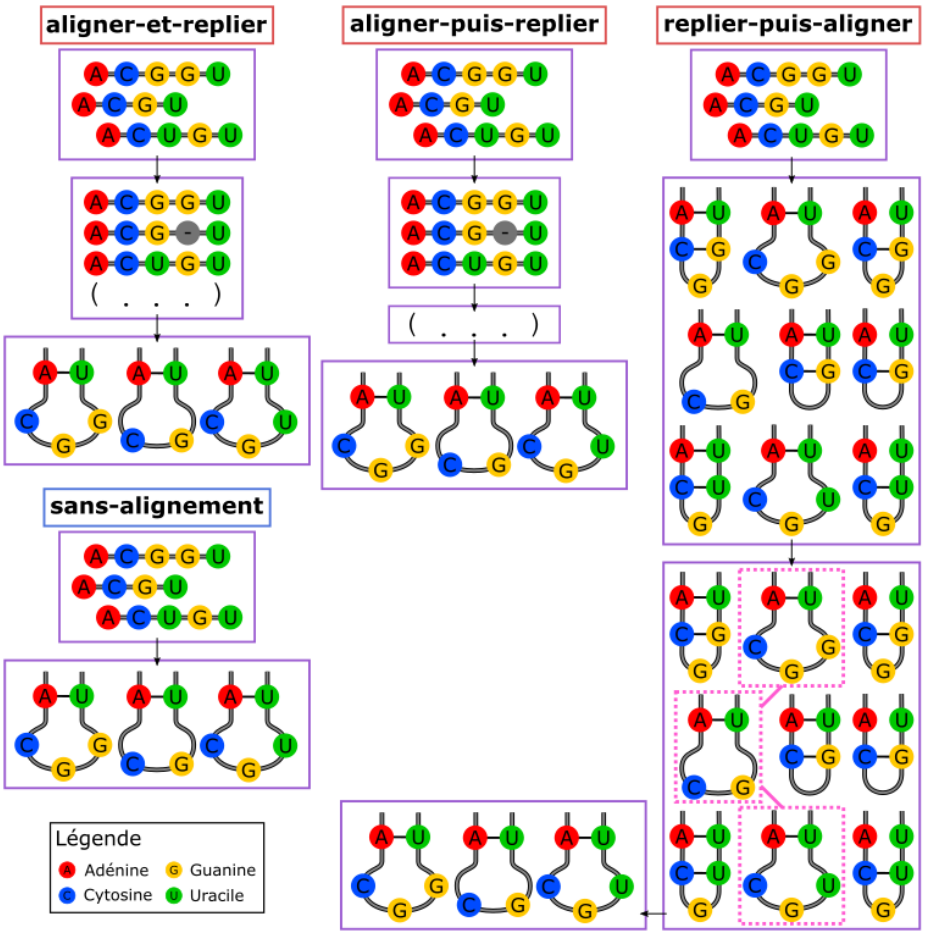

FIGURE 1.14 – Différentes stratégies de prédiction pour un ensemble de séquences d'ARNnc

Stratégie «sans-alignement»

La quatrième et dernière stratégie des méthodes comparatives pour la prédiction de structures secondaires de séquences d'ARN homologues est «sans-alignement» (*alignment-free*). Cette stratégie consiste à faire la prédiction de structures secondaires des séquences sans se baser sur l'alignement des séquences ou des structures. (*Voir Figure 1.14.*)

Pour la stratégie «sans-alignement», l'outil RNAcast[34] permet de déterminer la structure secondaire consensus commune pour l'ensemble des séquences d'ARN de la famille par une représentation sous forme abstraite des structures secondaires. L'outil aliFreeFold[14] permet de prédire une structure secondaire représentative pour l'ensemble des séquences d'ARN composant la famille en utilisant un ensemble de structures secondaires sous-optimales et une représentation vectorielle des éléments structuraux composant chaque sous-structure.

1.3.3 Avantages et limites des différentes stratégies de prédiction

Les méthodes de prédiction pour une seule séquence sont principalement limitées par la précision du modèle de minimisation de l'énergie libre de la structure et par le fait que la séquence peut avoir différentes structures secondaires selon les conditions d'environnement biotique, par exemple la température [8, 30, 44]. De plus, ces méthodes ne permettent de retrouver que 60% à 70% des vrais appariements de bases diminuant jusqu'à 40% pour des longues séquences [8, 30]. En revanche, le principal avantage des méthodes de prédiction pour une seule séquence est la rapidité, soit un temps d'exécution très faible. Pour cette raison, elles sont souvent utilisées comme étape préliminaire dans certaines méthodes comparatives pour la prédiction de la structure de plusieurs séquences.

Comparativement aux méthodes pour une seule séquence, les méthodes comparatives pour la prédiction de la structure de plusieurs séquences d'ARN homologues produisent des structures secondaires plus précises [32]. Ces méthodes exploitent la conservation de structure au sein d'une famille d'ARN homologues pour prédire la structure consensus de la famille. Alors que la précision des structures secondaires augmente, le coût en temps d'exécution augmente aussi. L'algorithme exact de Sankoff pour la stratégie «aligner-et-replier» a une complexité en temps de $O(n^{3N})$ et une complexité en mémoire de $O(n^{2N})$ où n est la longueur maximale des séquences d'ARN et N est le nombre de séquences d'ARN [40]. L'intuition derrière la complexité en temps de cet algorithme est que pour chaque séquence, l'ensemble de ses segments est considéré, soit $O(n^2)$ segments pour chaque séquence. La combinaison des segments

des N séquences requière $O(n^{2N})$ en espace pour stocker les résultats pour chaque combinaison. Puis, pour chaque combinaison, chaque segment peut être divisé en deux sous-segments consécutifs de $O(n)$ différentes façons, soit $O(n^N)$ pour les N segments d'une combinaison. La complexité en temps résultante est donc en $O(n^{3N})$. L'algorithme de l'alignement et le repliement pour deux séquences, par exemple, d'ARN est décrit comme suit :

Formules de récurrence :

$$M(i,j;k,l) = \max \begin{cases} M(i,j-1;k,l-1) + \sigma(A_j, B_l) \\ M(i,j-1;k,l) + \gamma \\ M(i,j;k,l-1) + \gamma \\ \max_{j',l'} M(i,j'-1;k,l'-1) + D(j',j;l',l) \end{cases}$$

$$D(i,j;k,l) = M(i+1,j-1;k+1,l-1) + \tau(i,j,k,l)$$

Dans les formules de récurrence ci-haut, $\sigma(A_j, B_l)$ représente le score élémentaire pour aligner les bases non-appariées A_j et B_l alors que $\tau(i,j,k,l)$ représente le score élémentaire élémentaire pour aligner les bases appariées A_i, A_j, B_k et B_l. Cet algorithme n'est pas applicable à grande échelle à cause de sa complexité exponentielle. Ainsi, des heuristiques et des algorithmes d'approximation, ont été développés suivant la stratégie «aligner-et-replier» afin de réduire les temps de calcul au prix d'une réduction de la précision de la prédiction de structures secondaires.

Les méthodes de la stratégie «aligner-puis-replier» sont contraintes par la précision des algorithmes d'alignement multiple de séquences qui commence à diminuer drastiquement lorsque les séquences de la famille sont dissimilaires, soit moins de 60% de pourcentage d'identité de séquence [5, 17]. L'avantage de ces méthodes est la rapidité.

Quant aux méthodes de la stratégie «replier-puis-aligner», elles sont contraintes par le temps d'alignement des structures générées pour les séquences. L'avantage de ces méthodes est que celles-ci ne sont pas limitées par la précision de l'alignement des séquences. Alors que les méthodes des stratégies «aligner-et-replier» et «replier-puis-aligner» prédisent des structures secondaires d'ARN avec une plus grande précision que les méthodes de la stratégie «aligner-puis-replier», cette dernière est plus rapide en temps d'exécution que les deux premières.

Enfin, les méthodes de la stratégie «sans-alignement» permettent de prédire des structures secondaires pour des séquences d'ARN homologues sans considérer des alignements de séquences ou de structures. Elles ne sont pas sensibles au taux de similarité des séquences. De plus, la complexité en temps de ces méthodes est généralement linéaire par rapport au nombre de séquences. Cependant, la précision des méthodes actuelles basées sur cette stratégie est encore inférieure à celles des méthodes de la stratégie «aligner-et-replier».

1.4 Hypothèse et objectif de la maîtrise

L'hypothèse de cette maîtrise est que, l'extension de l'outil de prédiction de structures secondaires d'ARN homologues «sans-alignement» aliFreeFold permettrait de prédire la structure secondaire pour chacune des séquences d'une famille d'ARN et permettrait d'augmenter la précision des structures secondaires prédites.

L'objectif principal de cette maîtrise est de développer un outil de prédiction de structures secondaires d'ARN homologues utilisant la stratégie «sans-alignement», en particulier une extension d'aliFreeFold, ayant des performances comparables aux méthodes des stratégies «aligner-et-replier» et «replier-puis-aligner» et un temps de calcul faible comparativement à ces deux dernières stratégies.

1.5 Structure du mémoire

Ce mémoire comporte :

— Un chapitre d'introduction générale qui présente les différentes notions explorées à travers le mémoire et donne un résumé de l'état de l'art concernant la prédiction de structures secondaires d'ARNnc.

— Un chapitre qui présente la méthode aliFreeFoldMulti développée. Ce chapitre décrit en détails l'outil aliFreeFold, la méthode aliFreeFoldMulti et les apports de aliFreeFoldMulti par rapport à aliFreeFold. L'article intégral d'aliFreeFoldMulti rédigé en anglais est inclus.

— Un chapitre de conclusion qui présente un résumé de l'apport de l'outil aliFreeFoldMulti à la communauté scientifique dans le domaine de la prédiction de structures secondaires d'ARN homologues ainsi que les différentes perspectives pour l'amélioration de cette méthode.

Chapitre 2

Méthode «sans-alignement» de prédiction de structures secondaires

Dans ce chapitre, nous décrivons un nouvel algorithme de prédiction de structures secondaires d'ARN homologues «sans-alignement» aliFreeFoldMulti développé au sein du laboratoire CoBIUS de l'Université de Sherbrooke. La première section présente l'outil aliFreeFold et son fonctionnement. La seconde section décrit le nouvel algorithme aliFreeFoldMulti qui est une extension de l'outil aliFreeFold. L'algorithme aliFreeFold a été développé avant mon arrivée et j'ai effectué les développements de l'algorithme aliFreeFoldMulti ainsi que les analyses et les tests de présents dans l'article décrivant l'algorithme aliFreeFoldMulti au cours de mes travaux de maîtrise.

2.1 La méthode «sans-alignement» aliFreeFold

aliFreeFold se situe dans la classe «sans-alignement» des méthodes de prédiction pour des séquences d'ARN homologues. Cet outil utilise des structures sous-optimales générées pour les séquences en entrée. Il est basé sur la démonstration que pour une séquence contenant moins de 800 nucléotides, il y a toujours, à l'intérieur des premières 25 structures sous-optimales, une structure idéale ayant en moyenne 80% des paires

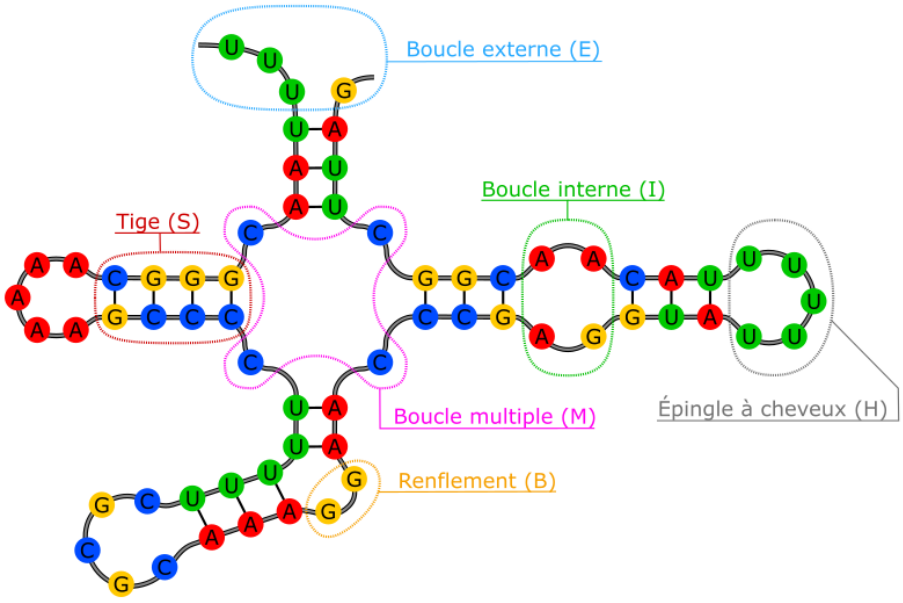

FIGURE 2.1 – Une structure secondaire d'ARN et ses différents éléments structuraux

de bases communes avec la structure optimale de référence [52]. aliFreeFold utilise une représentation vectorielle des structures secondaires d'ARN suivant le modèle des n-motifs [16]. (*Voir la section suivante pour la définition d'un n-motif.*)

2.1.1 Le modèle des n-motifs

La structure secondaire d'une séquence d'ARNnc est composée d'éléments structuraux élémentaires, tels que les tiges (*stem*), les renflements (*bulge*), les épingles à cheveux (*hairpin*), les boucles externes (*external loop*), les boucles internes (*internal loop*) et les boucles multiples (*multiple loop*). (*Voir Figure 2.1.*) Un n-motif est un élément structural ou un ensemble d'éléments structuraux adjacents [16]. Par exemple, un 0-motif correspond à un élément structurel comme une tige, un 1-motif correspond à un 0-motif et ses éléments adjacents comme une boucle externe et une tige et ainsi de suite.

La génération du modèle des n-motifs pour un ensemble de structures secondaires comporte trois étapes importantes. La première étape de ce modèle consiste à identifier tous les motifs présents dans chacune des structures secondaires. La deuxième

étape consiste à calculer un niveau d'importance relatif de chacun des n-motifs et de retenir les motifs permettant de couvrir et représenter l'ensemble des éléments structuraux des structures secondaires. Finalement, la dernière étape consiste à transformer cette représentation de structures secondaires en n-motifs en utilisant une décomposition en valeurs singulières permettant de capter des similarités intrinsèques entre des structures secondaires ne partageant pas beaucoup de n-motifs. De plus, cela permet de réduire le nombre de n-motifs dans la représentation finale de structures secondaires. Les n-motifs aberrants sont retirés de la représentation et les n-motifs sur-représentés sont réduits afin de diminuer l'impact de cette sur-représentation par rapport aux n-motifs moyennement représentés [16].

2.1.2 Fonctionnement d'aliFreeFold

aliFreeFold prend en entrée un ensemble de $s > 1$ séquences d'ARN homologues. Par la suite, aliFreeFold génère les 25 premières structures sous-optimales pour chacune des séquences d'ARN composant la famille à l'aide de RNAsubopt[24]. (*Voir l'étape 1 de la figure 1 dans l'article aliFreeFoldMulti.*) On obtient ainsi un total de $25s$ structures secondaires correspondant aux s séquences en entrée. Pour l'ensemble des structures sous-optimales obtenues, aliFreeFold génère une représentation sous forme de n-motifs formant ainsi une matrice de $25s$ lignes et m colonnes, telle que m est le nombre de n-motifs existant à travers l'ensemble des structures générées. (*Voir l'étape 2 de la figure 1 dans l'article aliFreeFoldMulti.*) Par la suite, aliFreeFold calcule un indice de conservation basé sur l'entropie pour chacune des colonnes de la matrice. (*Voir l'étape 3 de la figure 1 dans l'article aliFreeFoldMulti.*) On obtient ainsi une matrice ligne composée de m colonnes contenant les indices de conservation pour chacun des n-motifs. Plus la valeur de l'indice de conservation est grande, plus ce n-motif est conservé. Ensuite, chaque valeur dans la matrice représentant les structures sous forme de n-motifs, est multipliée par l'indice de conservation du n-motif correspondant. (*Voir l'étape 4 de la figure 1 dans l'article aliFreeFoldMulti.*) À partir de la nouvelle matrice pondérée, aliFreeFold calcule un centroïde global pour l'ensemble des structures sous-optimales de la famille en calculant la moyenne de chacune des colonnes. (*Voir l'étape 5 de la figure 1 dans l'article aliFreeFoldMulti.*) La structure

sous-optimale la plus proche de ce centroïde, soit la structure sous-optimale ayant la distance euclidienne minimale par rapport au centroïde, est extraite. (*Voir l'étape 6 de la figure 1 dans l'article aliFreeFoldMulti.*) Finalement, la sortie d'aliFreeFold est la séquence et la structure secondaire correspondant à la structure sous-optimale la plus proche du centroïde.

Les principaux avantages d'aliFreeFold sont qu'il peut prédire rapidement et précisément une structure secondaire consensus pour un ensemble de séquences d'ARN homologues sans étape d'alignement en se basant sur une représentation vectorielle des structures secondaires sous-optimales. En revanche, la principale limite d'aliFreeFold est qu'il ne produit qu'une seule structure secondaire, soit une structure secondaire consensus et représentative de la famille d'ARN. En effet, peu importe la taille de la famille, c'est-à-dire le nombre de séquences d'ARN homologues, aliFreeFold retourne toujours une seule structure secondaire pour la famille.

2.2 aliFreeFoldMulti, une extension d'aliFreeFold

Nous avons développé l'outil aliFreeFoldMulti[3] qui est une extension de l'outil aliFreeFold pour permettre la prédiction de structures secondaires pour toutes les séquences d'une famille d'ARN homologues. Cette extension de l'outil aliFreeFold comporte quatre nouvelles stratégies pour permettre de prédire la structure de toutes les séquences d'ARN d'une famille. Ce nouvel outil a fait l'objet d'une publication dans *NAR Genomics and Bioinformatics* en octobre 2020 [3].

2.2.1 Stratégie «centroïde»

La première stratégie appelée «centroïde» permet de générer une structure secondaire pour chacune des séquences d'ARN homologues en entrée. Dans cette stratégie, chaque groupe de 25 structures secondaires sous-optimales produites par RNAsubopt pour une séquence est comparé avec le centroïde obtenu dans la méthode aliFreeFold et la structure secondaire sous-optimale la plus proche du centroïde est retenue comme structure secondaire pour la séquence. (*Voir la figure 2 dans l'article aliFreeFoldMulti.*)

2.2.2 Stratégie «centroïde ajustée»

La deuxième stratégie appelée «centroïde ajusté» est une stratégie qui dérive de la précédente. Elle permet de générer une structure secondaire pour chacune des séquences d'ARN en tentant de satisfaire deux conditions d'optimalité, soient de minimiser la distance avec le centroïde et la distance entre les structures sous-optimales retenues pour les séquences. (*Voir la figure 2 dans l'article aliFreeFoldMulti.*)

2.2.3 Stratégie «plongement des tiges»

La troisième stratégie appelée «plongement de tiges» permet de calculer une structure secondaire pour chacune des séquences d'ARN homologues en utilisant la structure secondaire représentative de l'ensemble de la famille d'ARN, calculée par aliFreeFold, et les tiges des différentes structures secondaires sous-optimales générées par RNAsubopt. (*Voir la figure 2 dans l'article aliFreeFoldMulti.*) La tige est définie par quatre caractéristiques, soient la longueur du début, la longueur de la fin, la longueur de la tige et la longueur de la boucle. La longueur du début est le nombre de bases azotées du début de la séquence jusqu'au début de la tige, la première base azotée appariée de la tige. La longueur de la fin est le nombre de bases azotées de la fin de la tige jusqu'à la fin de la séquence, la dernière base azotée appariée de la tige. La longueur de la tige correspond au nombre d'appariements de paire de bases formant la tige. La longueur de la boucle est le nombre de bases azotées se retrouvant à l'intérieur de la tige entre les bases appariées de gauche et de droite. (*Voir la figure 3 de l'article aliFreeFoldMulti.*) Trois méthodes différentes de plongement des tiges des structures secondaires sous-optimales dans la structure secondaire représentative ont été développées. Elles diffèrent par l'ordre dans lequel les tiges sont traitées, soient d'abord les meilleures tiges ou de la gauche vers la droite ou encore de la droite vers la gauche. L'algorithme de plongement est récursif et itératif. Par exemple, dans le cas de la méthode qui considère les tiges de la gauche vers la droite, lorsqu'on traite une tige, avant de passer à la suivante, on doit vérifier s'il y a d'autres tiges à l'intérieur de la première tige. (*Voir la figure 4 de l'article aliFreeFoldMulti.*) Cette stratégie, comparativement aux autres développées, propose une structure secondaire qui est une combinaison de tiges des solutions optimales qui s'accordent le mieux avec la

structure consensus de la famille déterminée par aliFreeFold et non une des solutions des structures secondaires sous-optimales.

2.2.4 Stratégie «plus proche du sous-optimal»

Finalement, la quatrième stratégie appelée «plus proche du sous-optimal» permet de calculer une structure secondaire pour chacune des séquences d'ARN homologues basée sur le résultat de la stratégie «plongement des tiges». Ainsi, chaque structure secondaire inférée pour une séquence sera comparée avec les 25 structures sous-optimales de cette même séquence. La structure secondaire sous-optimale la plus proche de la structure inférée sera celle retenue pour la séquence. Dans cette stratégie, aliFreeFoldMulti retourne la structure secondaire sous-optimale générée par RNAsubopt la plus similaire de celle déterminée par la stratégie «plongement des tiges». (*Voir la figure 2 de l'article aliFreeFoldMulti.*)

2.2.5 Résumé des résultats

Le principal avantage d'aliFreeFoldMulti par rapport à aliFreeFold est la possibilité de prédire rapidement et précisément une structure secondaire pour chacune des séquences d'ARN composant la famille d'ARN donnée en entrée. De plus, aliFreeFoldMulti offre diverses stratégies permettant d'obtenir des prédictions de structures secondaires rapidement et précisément.

Les résultats obtenus sur la comparaison des stratégies montrent que les stratégies «centroïde», «centroïde ajusté» et «plus proche du sous-optimal» sont les meilleures stratégies d'aliFreeFoldMulti. De plus, les temps de calcul pour les différentes stratégies sont comparables. Les résultats obtenus sur la comparaison de aliFreeFoldMulti en utilisant la stratégie «centroïde» avec d'autres outils de prédiction de structures secondaires d'ARN homologues montrent que aliFreeFoldMulti a les meilleurs scores maximums, mais TurboFoldII obtient les meilleurs scores moyens. En revanche, en ce qui concerne les temps de calcul, aliFreeFoldMulti est largement plus rapide que tous les autres outils.

2.3 Article «aliFreeFoldMulti : alignment-free method to predict secondary structures of multiple RNA homologs»

L'article ci-dessous, rédigé en anglais et publié dans la revue *NAR Genomics and Bioinformatics*, élabore davantage sur l'outil aliFreeFoldMulti, son fonctionnement et les résultats de la comparaison entre les différentes stratégies et avec différents outils de prédiction de structures secondaires d'ARN, soient CentroidAlifold, FoldalignM, RNAalifold, RNAspa et TurboFoldII.

Résumé en français : Prédire la structure d'un ARN est crucial pour la compréhension du mécanisme d'action de l'ARN. Les approches comparatives pour la prédiction de structures d'ARN peuvent être classées en quatre stratégies. Les trois premières, «aligner-et-replier», «aligner-puis-replier» et «replier-puis-aligner», exploitent des alignements multiples de séquences et/ou de structures pour améliorer la précision de la prédiction de la structure d'ARN conservée. Les méthodes «aligner-et-replier» performent généralement mieux, mais sont aussi typiquement plus lentes que les deux autres approches. La quatrième stratégie «sans-alignement», consiste à la prédiction de la structure de l'ARN conservée sans s'appuyer sur l'alignement de séquences ou de structures. Cette stratégie a l'avantage d'être plus rapide, tout en prédisant des structures précises grâce à l'utilisation de représentations latentes des structures candidates pour chaque séquence. Cet article présente aliFreeFoldMulti, une extension de l'algorithme d'aliFreeFold. Ce dernier prédit une structure secondaire représentative de plusieurs ARN homologues en utilisant une représentation vectorielle de leurs structures sous-optimales. aliFreeFoldMulti améliore aliFreeFold en calculant en plus la structure conservée pour chaque séquence. aliFreeFoldMulti est évaluée en comparant ses performances de prédiction et son efficacité de temps avec un ensemble de méthodes de prédiction de la structure d'ARN. aliFreeFoldMulti a les temps de calcul les plus bas et les scores de précision maximum les plus élevés. Il atteint une précision de prédiction de structures moyenne comparable à celle d'autres méthodes, à l'exception de TurboFoldII qui est la meilleure en termes de précision moyenne mais avec les temps de calcul les plus élevés. Nous présentons aliFreeFoldMulti comme une

illustration du potentiel des approches «sans-alignement» pour fournir des méthodes rapides et précises de prédiction de la structure d'ARN.

Contribution des auteurs : J'ai conçu l'étude avec Jean-Pierre Séhi Glouzon et Aïda Ouangraoua. J'ai écrit le programme avec Valentin Carpentier et Aïda Ouangraoua. J'ai rédigé la documentation et conçu le serveur web avec Yanchun Qi. J'ai collecté les données avec Yoann Anselmetti. J'ai réalisé les expériences et généré les figures des résultats. J'ai rédigé la première version de l'article avec Yoann Anselmetti. Aïda Ouangraoua a révisé le manuscrit. L'article a été publié dans le journal *NAR Genomics and Bioinformatics* le 27 octobre 2020.

2.3. ARTICLE «ALIFREEFOLDMULTI : ALIGNMENT-FREE METHOD TO PREDICT SECONDARY STRUCTURES OF MULTIPLE RNA HOMOLOGS»

Published online 27 October 2020

NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 4 **1**
doi: 10.1093/nargab/lqaa086

aliFreeFoldMulti: alignment-free method to predict secondary structures of multiple RNA homologs

Marc-André Bossanyi, Valentin Carpentier, Jean-Pierre S. Glouzon, Aïda Ouangraoua* and Yoann Anselmetti

CoBIUS lab, Department of Computer Science, University of Sherbrooke, 2500 Boulevard de l'Université, Sherbrooke, QC J1K 2R1, Canada

Received October 01, 2020; Editorial Decision October 06, 2020; Accepted October 19, 2020

ABSTRACT

Predicting RNA structure is crucial for understanding RNA's mechanism of action. Comparative approaches for the prediction of RNA structures can be classified into four main strategies. The three first—align-and-fold, align-then-fold and fold-then-align—exploit multiple sequence alignments to improve the accuracy of conserved RNA-structure prediction. Align-and-fold methods perform generally better, but are also typically slower than the other alignment-based methods. The fourth strategy—alignment-free—consists in predicting the conserved RNA structure without relying on sequence alignment. This strategy has the advantage of being the faster, while predicting accurate structures through the use of latent representations of the candidate structures for each sequence. This paper presents aliFreeFoldMulti, an extension of the aliFreeFold algorithm. This algorithm predicts a representative secondary structure of multiple RNA homologs by using a vector representation of their suboptimal structures. aliFreeFoldMulti improves on aliFreeFold by additionally computing the conserved structure for each sequence. aliFreeFoldMulti is assessed by comparing its prediction performance and time efficiency with a set of leading RNA-structure prediction methods. aliFreeFoldMulti has the lowest computing times and the highest maximum accuracy scores. It achieves comparable average structure prediction accuracy as other methods, except TurboFoldII which is the best in terms of average accuracy but with the highest computing times. We present aliFreeFoldMulti as an illustration of the potential of alignment-free approaches to provide fast and accurate RNA-structure prediction methods.

INTRODUCTION

RNA-structure prediction is essential to better understand the biological mechanism of noncoding RNAs, which are involved in a vast part of the biochemical machinery in cells (1). Some examples are the transcription of DNA in RNA with RNA polymerases (2), the regulation of gene expression (3), the translation of RNA in proteins (4), but there are many other biological functions (5).

In the last two decades, several approaches have been devised to predict RNA secondary structures from a single RNA sequence or a set of homologous RNA sequences. Single-sequence approaches were developed first. They are mainly based on the computation of the minimum-free-energy (MFE) secondary structure of an RNA sequence (6–8). Various studies have shown that single-sequence approaches have limited accuracy, because several MFE secondary structures are possible for a given RNA sequence and biotic environment conditions can affect the stability of the MFE structure (7,9–10). Compared to single-sequence approaches, multiple-sequence approaches have been fruitful in improving the prediction of RNA secondary structures. They consist in predicting a consensus RNA secondary structure for a set of RNA homologs. Most multiple-sequence approaches are based on a comparative approach that combines multiple RNA sequence alignment and RNA folding prediction (11–13). Comparative approaches that exploit sequence alignment can be categorized into three main strategies. The first strategy, align-and-fold consists of methods that solve the sequence alignment and folding problems simultaneously by computing an optimal multiple sequence-structure alignment. The complexity of the exact solution for the simultaneous multiple RNA sequence alignment and folding problem on a set of homologous RNA sequences is in $O(n^{3N})$ in time and $O(n^{2N})$ in space, where n is the maximum length of the RNA sequences and N is the number of RNA sequences (14). Given that the computation of an exact solution solution is highly time-consuming, current methods that follow the align-and-fold strategy are based on greedy heuristics to find the common structure using multiple pairwise compar-

*To whom correspondence should be addressed. Tel: +1 819 821 8000 #62014; Fax: +1 819 821 8200; Email: aida.ouangraoua@usherbrooke.ca

2.3. ARTICLE «ALIFREEFOLDMULTI : ALIGNMENT-FREE METHOD TO PREDICT SECONDARY STRUCTURES OF MULTIPLE RNA HOMOLOGS»

isons (e.g. Foldalign (15,16), TurbofoldII (17), DynalignII (18), SPARSE (19)).

The second and third strategies referred as align-then-fold and fold-then-align consist of methods that solve the alignment and folding problems sequentially and use the solution of the first problem as a proxy to solve the second problem. The advantage of align-then-fold methods is speed, but their drawback is that the quality of the structure prediction depends on the quality of the sequence alignment which reflects poorly the structural homology with dissimilar sequences (e.g. RNAalifold (20), CentroidFold (21), Transat (22), CentroidAlifold (23)). Fold-then-align methods predict a set of low-free-energy secondary structures for each RNA sequence and then align the predicted structures to find the lowest free energy structure common to all sequences (e.g. RNAspa (24)). Their advantage is not being limited by the accuracy of sequence alignment. Their drawback is being highly time-consuming in aligning all low-free-energy structures. Thus, align-and-fold and fold-then-align methods generally predict more accurate RNA secondary structures than align-then-fold, but the first two are typically slower than the last one, which leaves room for the development of fast methods yielding accurate structure prediction. In response to this need, a fourth strategy named alignment-free has been developed and does not rely on any time-consuming sequence or structure alignment computations. Methods using this strategy consist in predicting a set of low-free energy secondary structures for each RNA sequence, and using a latent representation of the secondary structures to explore their homology and predict a consensus RNA secondary structure (e.g. RNAcast (25), aliFreeFold (26)).

Recently, we developed the aliFreeFold algorithm (26) that predicts a consensus secondary structure for a set of RNA homologous sequences using an alignment-free strategy. aliFreeFold consists in computing suboptimal MFE secondary structures for each RNA sequence using RNAsubopt (8) and the Zuker *et al*. method (27). It then computes a vector representation of structures based on the n-motifs model (28). The n-motifs model represents an RNA secondary structure as a vector of counts of elementary structural motifs. The vector representation of suboptimal structures helps to capture conservation signals of structural features across the suboptimal structures, and to extract a single representative secondary structure that contains conserved structural features. This paper presents aliFreeFoldMulti which is an extension of the aliFreeFold algorithm. aliFreeFoldMulti improves on the original aliFreeFold algorithm by predicting secondary structures for all sequences of a family of RNA homologs, instead of a single consensus structure for the family. It includes several strategies to predict the secondary structures of all homologous RNA sequences. To assess the performance of aliFreeFoldMulti, the accuracy of structure predictions and the computing time were compared with those of the current best performing prediction methods, including align-and-fold methods (FoldalignM (16), TurbofoldII (17)), align-then-fold methods (RNAalifold (20), CentroidAlifold (23)) and a fold-then-align method (RNAspa (24)). The results show that TurboFoldII has higher average prediction accuracy than all methods, when all predicted structures for an RNA family are considered. However, when we consider the best predicted structure in each family, aliFreeFoldMulti has the highest maximum accuracy. In terms of time efficiency, aliFreeFoldMulti is faster than the other methods. Like aliFreeFold, aliFreeFoldMulti effectively captures conservation signals to achieve fast, and accurate predictions. The source code of aliFreeFoldMulti is freely available under the GPL license at https://github.com/UdeS-CoBIUS/aliFreeFoldMulti. A web server is available at https://alifreefold.cobius.usherbrooke.ca.

MATERIALS AND METHODS

aliFreeFold

The input for the original aliFreeFold algorithm (26) is a set of homologous RNA sequences and the output is a representative consensus secondary structure for the set of RNA sequences using an alignment-free strategy (see Figure 1 for an overview of the original aliFreeFold algorithm). The method comprises five main steps. In Step 1, it starts by generating the first 25 suboptimal structures for each sequence using RNAsubopt (8). In Step 2, aliFreeFold represents each suboptimal structure using the n-motif representation model such that a n-motif is an elementary RNA structural motif, such as a hairpin, stem, bulge, or internal or multiple loops with the adjacent motifs (28). Each suboptimal structure is represented by a vector of counts of n-motifs occurring in the structure. This yields a matrix representation of the set of suboptimal structures such that lines represent the suboptimal structures generated for all sequences, columns represents the n-motifs occurring in the structures and each cell (i,j) contains the number of occurrences of the j^{th} n-motif in the i^{th} suboptimal structure. In Step 3, aliFreeFold computes the entropy-based conservation index for each n-motif on the whole set of suboptimal structures generated. In Step 4, using the conservation indexes of n-motifs, the n-motif representation of the set of suboptimal structures is transformed using the conservation indexes of n-motifs into a weighted n-motif representation giving more importance to conserved n-motifs. In Step 5, the centroid of all the suboptimal structures represented by the weighted n-motifs representation is computed as the mean vector of the weighted n-motif representation, and the distance between the centroid and each suboptimal structure is computed. Lastly, in Step 6, the representative structure is defined as the structure that has the most common structural features with the suboptimal structures of homologous sequences. It is computed as the suboptimal structure closest to the centroid in terms of distance.

aliFreeFoldMulti

aliFreeFoldMulti improves upon the original aliFreeFold algorithm by providing secondary structure predictions for all sequences of an input family of RNA homologs instead of a single consensus structure for the RNA family. It includes four strategies which have been developed to extend the aliFreeFold algorithm in order to predict all secondary structures for a set of RNA homologs. Each of the four strategies is described below in more detail (see Figure 2 for an overview of the four strategies).

2.3. ARTICLE «ALIFREEFOLDMULTI : ALIGNMENT-FREE METHOD TO PREDICT SECONDARY STRUCTURES OF MULTIPLE RNA HOMOLOGS»

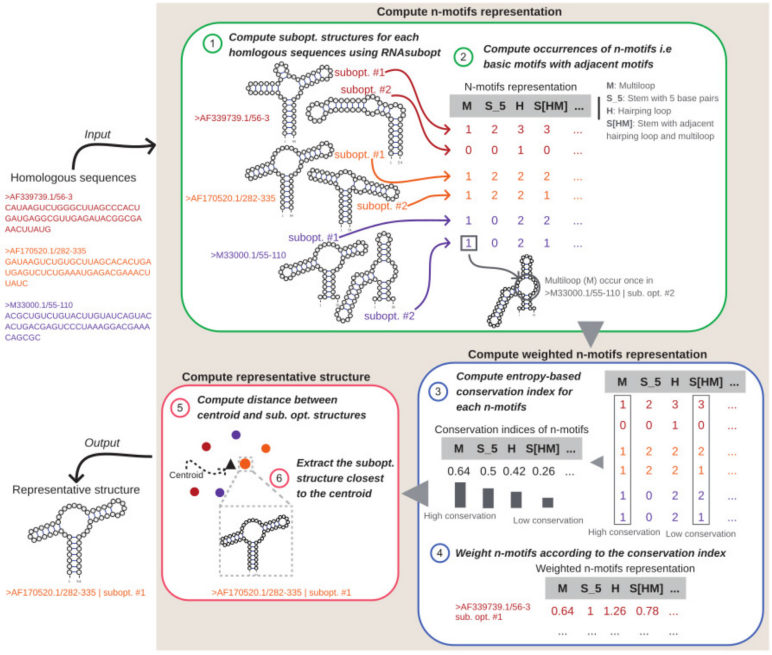

Figure 1. Overview of the aliFreeFold approach (Figure 1 from (26)).

Centroid strategy. This strategy is the most direct extension of the original aliFreeFold algorithm, maintaining the first five steps in aliFreeFold. The last step of the centroid strategy consists in returning the suboptimal structure for each RNA sequence that is the closest to the centroid in terms of distance. The rationale behind this strategy is that the results of the original aliFreeFold algorithm have shown that the centroid effectively summarizes the conserved structural features of a set of homologous RNA. Thus, we expect the centroid strategy to yield a set of homologous secondary structures that share the conserved structural features captured by the centroid.

Adjusted-centroid strategy. This strategy was derived from the centroid strategy. It aims at computing a set of homologous secondary structures that are both close to the centroid and close to each other. In addition to computing the distance between the centroid and each suboptimal structure, the adjusted-centroid strategy computes the distance between each pair of suboptimal structures. The sum of the distances to the closest suboptimal structures of the other RNA sequences is computed for each structure predicted by the centroid strategy for a RNA sequence. Then, the method chooses the predicted structure that minimizes this sum of distances, and its set of closest suboptimal structures as the set of homologous RNA structures for the input RNA sequences. The rationale of this strategy is that the input RNA sequences are expected to have the most similar RNA structures.

Stem-embedding strategy. This strategy aims at using the representative structure computed by the original aliFreeFold algorithm as a proxy to infer the secondary structures of other homologous sequences. The first six aliFreeFold steps are used to compute a representative secondary structure for the input set of homologous RNA sequences. The computed representative structure is a suboptimal structure of one of the input RNA sequences denoted by S_{rep}. The last step consists in computing, for each input RNA sequence S, a structure-preserving embedding of the set of stems of the representative structure S_{rep} in the set of stems of all 25 suboptimal structures of the sequence S. Given an input RNA sequence S different from S_{rep}, let X be the set of stems of the representative structure S_{rep}, and Y be the set of stems of all 25 suboptimal structures of the sequence S. A structure-preserving embedding of X in Y is an injective map f from X to Y such that, for any two stems $s1$ and $s2$ in X, if $s1$ is located inside (resp. before) $s2$, then $f(s1)$ is also located inside (resp. before) $f(s2)$. The embedding f of X in Y is computed with a heuristic algorithm that aims at minimizing the sum of distances between the stems of X and their images in Y by f. The distance $d(x, y)$ between

2.3. ARTICLE «ALIFREEFOLDMULTI : ALIGNMENT-FREE METHOD TO PREDICT SECONDARY STRUCTURES OF MULTIPLE RNA HOMOLOGS»

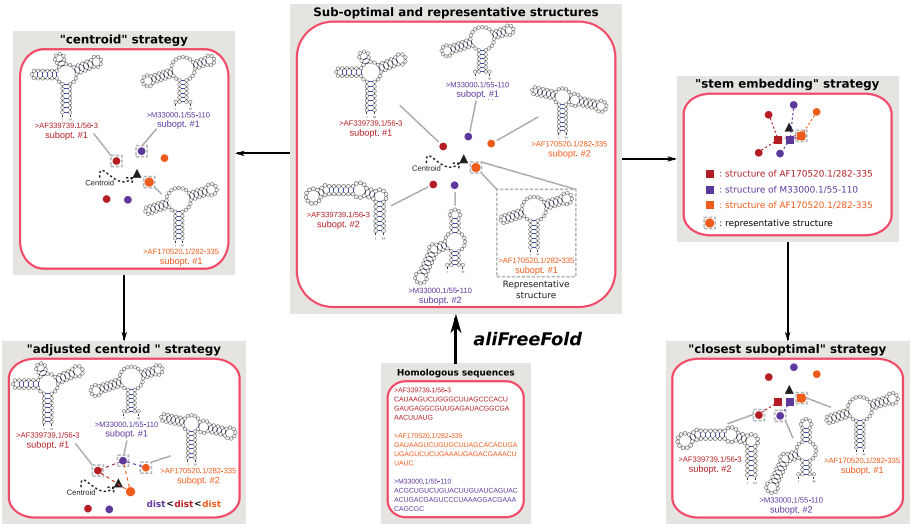

Figure 2. Overview of the four aliFreeFoldMulti strategies. aliFreeFoldMulti takes a set of RNA homologs as input. aliFreeFold samples 25 suboptimal structures for each RNA sequence and computes a representative structure. (i) In the centroid strategy, aliFreeFoldMulti defines the structure of each sequence as its suboptimal structure that is the closest to the centroid. (ii) In the adjusted-centroid strategy, aliFreeFoldMulti searches for a set of suboptimal structures that minimize both the distances to the centroid and the sum of pairwise distances between each other. (iii) In the stem-embedding strategy, aliFreeFoldMulti looks, for each sequence, for a set of stems of its suboptimal structures that forms a secondary structure and are the most similar to the stems of the representative structure (computed by aliFreeFold). (iv) In the closest-suboptimal strategy, aliFreeFoldMulti defines the structure of each sequence as its suboptimal structure that is the closest to the structure computed with the stem-embedding strategy.

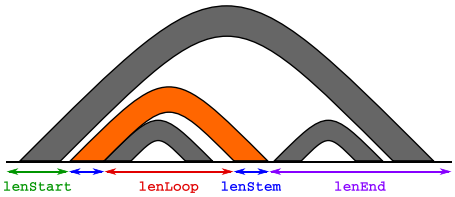

Figure 3. Arc diagram of a RNA secondary structure illustrating a stem (in orange), and the four information (*lenStart*, *lenLoop*, *lenStem* and *lenEnd*) used to compute the alignment score between two stems.

any stem x in X and y in X makes it possible to compare the location of stems x and y in their respective RNA sequences, and is defined based on the following information on the location of a stem s in its RNA sequence S (see Figure 3 for an illustration): $lenStart(s)$ is the length between the start of the RNA sequence and the first 5′ nucleotide of the stem s; $lenLoop(s)$ is the length between the last 5′ nucleotide of the stem s and the first 3′ nucleotide of the stem s, corresponding to the length of the 'loop' inside the stem; $lenEnd(s)$ is the length between the last 3′ nucleotide of the stem s and the end of the RNA sequence; and $lenStem$ is the number of pairs of nucleotides composing the stem. Based on this information computed for each stem of X and Y, the distance $d(x, y)$, for any $(x, y) \in X \times Y$, is computed with this formula:

$$\begin{aligned} d(x,y) = &(lenStart(x) - lenStart(y))^2 \\ &+(lenLoop(x) - lenLoop(y))^2 \\ &+(lenEnd(x) - lenEnd(y))^2 \\ &+(lenStem(x) - lenStem(y))^2 \end{aligned}$$

Based on the pairwise distances computed between stems of X and stems of Y, a greedy heuristic recursive algorithm is used to infer an embedding f of X in Y which minimizes the sum of distances between stems of X and their images in Y by f (see Figure 4 for an illustration of the three versions of the heuristic recursive algorithm). At each stage of the algorithm, a stem in x in X is selected, an optimal image $f(x)$ in Y is chosen to minimize $d(x, f(x))$, and the algorithm is recursively applied on subsets of X and Y, corresponding to the stems located respectively before x and $f(x)$, after x and $f(x)$, or nested in x and $f(x)$. The three versions named '*start*,' '*end*' and '*best*' of the greedy heuristic recursive algorithm have been developed. The three versions differ in the strategy used in each stage of the algorithm to select the stem x in X for which an optimal image in $f(x)$ in Y is chosen. The 'start' version consists in selecting the stem x from X minimizing $lenStart(x)$, i.e. which is the closest to

2.3. ARTICLE «ALIFREEFOLDMULTI : ALIGNMENT-FREE METHOD TO PREDICT SECONDARY STRUCTURES OF MULTIPLE RNA HOMOLOGS»

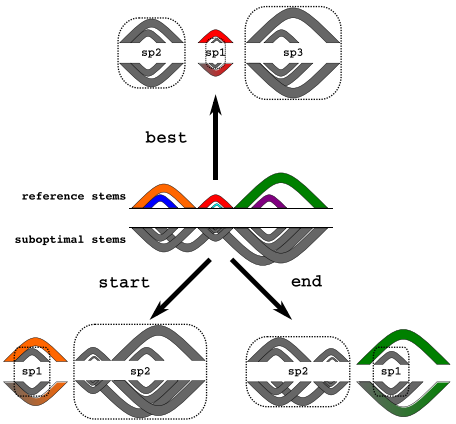

Figure 4. Illustration of the first step of the three different recursive algorithms of the stem-embedding strategies. Reference stems are the stems of the representative structure predicted by aliFreeFold. Suboptimal stems are the whole set of stems contained in the 25 suboptimal structures of the target RNA sequence. The 'start' strategy begins with the leftmost reference stem. Then, the sub-problems sp1 and sp2 are considered. The 'end' strategy is similar to the 'start' strategy, except that it begins with the rightmost reference stem. The 'best' strategy starts with the closest stems in the reference and the suboptimal sets. Then, recursively, the sub-problems sp1, sp2 and sp3 are considered.

the start of the sequence. The 'end' version consists in selecting the stem x from X minimizing $lenEnd(x)$, i.e. which is the closest to the end of the sequence. Lastly, the 'best' version consists in selecting a stem x from X not nested in any other stem of X and minimizing the distance to any stem in Y. The greedy heuristic algorithm is recursively applied, at each stage of the method, on sub-problems delineated by subsets of X and Y defined by the stem x selected in X and its optimal image in $f(x)$ in Y, until the considered subset of X or Y is empty. In order to make the greedy heuristic less sensitive to erroneous locally optimal choices, at each stage of the method, the three most optimal images for the stem x selected in X are tested and the image of x that yields the best global optimum is kept. Lastly, the structure predicted for an RNA sequence S is the structure comprising the images by f of the set of stems X.

Closest suboptimal strategy. The last strategy included in aliFreeFoldMulti is an extension of the stem-embedding strategy. It aims at computing a set of homologous secondary structures that are suboptimal structures close to the structures predicted by the stem-embedding strategy. It computes the suboptimal structure for each input RNA sequence that is the closest to the structure predicted by the stem-embedding strategy, in terms of the distance computed with the weighted n-motif representation.

Experimental setup

Datasets.

Small dataset. To evaluate the performance of aliFreeFoldMulti on case-study RNA-families, we used a dataset composed of 30 noncoding RNA families obtained from the BRALIBASE II (29) and MXSCARNA dataset (30). These two datasets were previously built and used in (29) and (30) to benchmark multiple sequence alignment programs upon structural RNAs. Each family is composed of a set of homologous sequences, each associated with a corresponding secondary structure. In each family, the redundant sequences were removed to leave a single copy of each sequence. Families differ in the number of homologous sequences, the average PID and the average sequence length, respectively, ranging from 16 to 98 sequences, from $\sim$ 58% PID to $\sim$ 98% PID and from $\sim$48 nt to $\sim$463 nt length. Further characteristics of the dataset are described in Additional File 1, Supplementary Table S1.

Large dataset. For a large-scale evaluation of the methods, we extracted a larger dataset from the Rfam database (version 14.1). Out of the 3016 ncRNA families available in Rfam, we selected all families composed of 10–100 RNA sequences, with maximum sequence length of 1000 nt. This resulted in 1125 families. Among these families, we discarded 221 families for which there is no consensus secondary structure, or that contain pseudoknots in their structure. We also removed 27 families for which nucleotide sequences contain character of the extended IUPAC code (i.e. RYSWKMBDHV). Out of the remaining 877 RNA-families, we finally discarded 14 families that yielded 'out-of-memory' errors for the FoldAlignM method, or 'infinite loop' errors for the RNAspa method. The final dataset is composed of 863 RNA families that have an average number of 26.89 ($\pm$18.66) sequences per family, and an average sequence length of 110.52 ($\pm$62.07) nt. Complete statistics for the 863 families are available in Additional File 2, Supplementary Table S1.

Compared methods. We selected six RNA secondary structure prediction methods representing the different strategies of comparative methods, for comparison with aliFreeFoldMulti in terms of prediction accuracy and computing time.

i. FoldalignM (15,16) and TurboFoldII (17) use the align-and-fold strategy. FoldalignM implements a multi-threading version of the Sankoff algorithm (14) with heuristics relying on a maximum length of the alignment γ, and a maximum difference between any two aligned subsequences δ. This allows for reducing the time complexity of the Sankoff algorithm from $O(L^6)$ to $O(L^2\gamma^2\delta^2)$, where L is the sequence length. TurboFoldII is a probabilistic approach that iteratively estimates base pairing probabilities for each sequence based on the thermodynamic nearest-neighbor model and posterior nucleotide co-incidence probabilities obtained using a hidden Markov model (HMM) for pairwise alignments. After several iterations of refinement, posterior co-incidence probabilities are used to compute the multiple sequence alignment and updated base-pair proba-

2.3. ARTICLE «ALIFREEFOLDMULTI : ALIGNMENT-FREE METHOD TO PREDICT SECONDARY STRUCTURES OF MULTIPLE RNA HOMOLOGS»

bilities are used to predict RNA structure for each sequence.

ii. CentroidAlifold (23) and RNAalifold (20) use the align-then-fold strategy. Their input is a multiple alignment of homologous RNA sequences. CentroidAlifold is an algorithm based on maximum expected accuracy. It maximizes the expected gain under a probability distribution of secondary structures for each RNA sequence. RNAalifold computes a consensus structure according to the partition function and base-pairing probability matrix using RIBOSUM scoring matrices in addition to the computation of MFE structure. CentroidAlifold and RNAalifold infer a single consensus structure for an RNA-family, but not a structure for each RNA sequence. In order to allow the comparison with methods predicting a structure for each sequence, we used the 'refold.pl' script from the ViennaRNA package (8) to obtain a secondary structure for each sequence.

iii. RNAspa (24) uses the fold-then-align approach. It uses the RNAsubopt method (8) from the ViennaRNA package to first sample suboptimal structures for each sequence. The set of suboptimal structures for each RNA sequence is represented as layer of disconnected vertices. RNAspa computes a similarity score alignment for all pairs of alternative structures of two adjacent layers, producing a directed acyclic graph with edges weighted by the similarity scores. RNAspa then predicts a secondary structure for each RNA sequence by finding the shortest path by traversal from the top to the bottom layer.

RNAcast (25) an alignment-free approach was not included in the analysis because aliFreeFold (26) outperforms it. Moreover, RNAcast ran out of memory above the threshold of 182 nt average sequence length, and it is no longer supported for recent Linux distributions. A webserver is available for RNAcast, but retrieving the execution time is not possible.

Evaluation criteria for the prediction accuracy. We use the performance metrics below to assess the accuracy of predicted RNA structures. The positive predictive value (PPV) represents the proportion of the predicted base pairs that are retrieved in the reference structure. The sensitivity (SENS) gives the ratio of the known base pairs of the reference structure found in the predicted ones. The Matthews correlation coefficient (MCC) summarizes the SENS and the PPV (31). PPV and SENS scores range between 0 and 1. A PPV score of 1 (resp. 0) means that all (resp. no) base pairs in the reference structure are found in the predicted structure. A SENS score of 1 (resp. 0) means that all (resp. no) base pairs in the predicted structure are found in the reference structure. MCC scores range between −1 and 1. A MCC score of 1 (resp. −1) means that the overall prediction is accurate (resp. inaccurate). SENS, PPV and MCC scores are computed as follows:

$$\text{SENSITIVITY} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{POSITIVE PREDICTIVE VALUE} = \frac{\text{TP}}{\text{TP} + (\text{FP} - \epsilon)}$$

$$\text{MATTHEW CORRELATION COEFFICIENT} = \frac{(\text{TP} * \text{TN}) - ((\text{FP} - \epsilon) * \text{FN})}{\sqrt{(\text{TP} + (\text{FP} - \epsilon))(\text{TP} + \text{FN})(\text{TN} + (\text{FP} - \epsilon))(\text{TN} + \text{FN})}}$$

where the true positives (TPs), the true negatives (TNs), the false negatives (FNs) and the false positives (FPs) represent, respectively, the number of correctly predicted base pairs, the number of nucleotide couples correctly identified as not paired, the number of base pairs in the reference not predicted, and the number of wrongly predicted base pairs. ϵ represents the number of base pairs in the predicted structures that are compatible with base pairs in the reference.

RESULTS

Performances of aliFreeFoldMulti strategies

The first evaluation consisted in assessing the accuracy and computing time of RNA secondary predictions obtained with the various aliFreeFoldMulti strategies and sub-strategies. The different strategies were applied on the small and large datasets of RNA families. For each sequence of each family, the MCC, PPV and SENS scores between the predicted and expected structures were computed. For each score (i.e. MCC, PPV and SENS) and each aliFreeFoldMulti strategy (i.e. centroid, adjusted-centroid, stem-embedding start, stem-embedding end, stem-embedding best, closest suboptimal start, closest suboptimal end and closest suboptimal best), Figure 5 gives two boxplots representing the maximum and average score distributions for the large dataset. The sub-strategies 'start,' 'end' and 'best' yielded similar results for each of the strategies stem-embedding and closest suboptimal. Therefore, we did not consider sub-strategies in the sequel, and we only discuss the global results of the strategies stem-embedding and closest suboptimal. Supplementary Figure S1A and B in Additional File 1 show the maximum and average score distributions for the small dataset, and the execution times of each strategy for increasing sequence lengths.

Centroid and adjusted-centroid are the best strategies for aliFreeFoldMulti. Based on Figure 5, we conclude that centroid and adjusted-centroid are the best strategies for aliFreeFoldMulti. The results show that these strategies yielded the best results, i.e. the highest maximum and average MCC scores. The adjusted-centroid strategy yielded results that are similar to the centroid strategy but with a slightly higher average MCC score, but a slightly lower maximum MCC score. The closest suboptimal strategy obtains performance scores (MCC, PPC and SENS) that are slightly lower than the two centroid-based strategies. The stem-embedding strategy had the highest PPV scores, but also the lowest SENS scores, which result in the lowest MCC scores. This means that the stem-embedding strategy found artificial structures that contain, on average, fewer incorrect pairs of nucleotides but also a lower number of expected pairs of nucleotides.

The accuracies of all strategies correlate with aliFreeFold accuracy. Given the high variances of scores within all strategies, we split the large dataset (respectively the small dataset) of RNA-families into three datasets according to

2.3. ARTICLE «ALIFREEFOLDMULTI : ALIGNMENT-FREE METHOD TO PREDICT SECONDARY STRUCTURES OF MULTIPLE RNA HOMOLOGS»

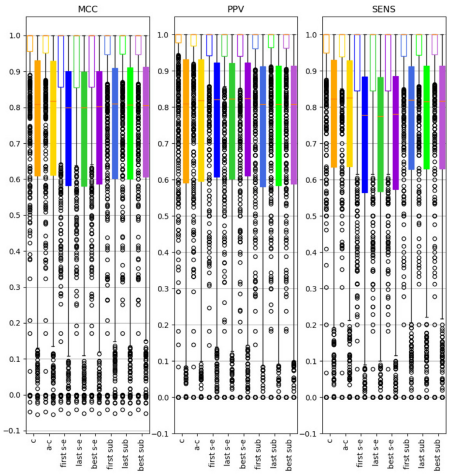

Figure 5. Boxplots of the MCC, PPV and SENS scores compared to expected structures on the large dataset of ncRNA families to assess the prediction accuracy of aliFreeFoldMulti strategies. The *x*-axis displays the four aliFreeFoldMulti strategies: centroid (c), adjusted-centroid (a–c), stem-embedding (s–e), and closest-suboptimal (sub). For stem-embedding and closest-suboptimal there are three different results corresponding to the substrategy used: 'start,' 'end' or 'best.' For each strategy, the left/empty (resp. right/full) boxplot represents the distribution of the maximum (resp. average) score. Structure prediction strategies and sub-strategies are described under 'Materials and Methods' section.

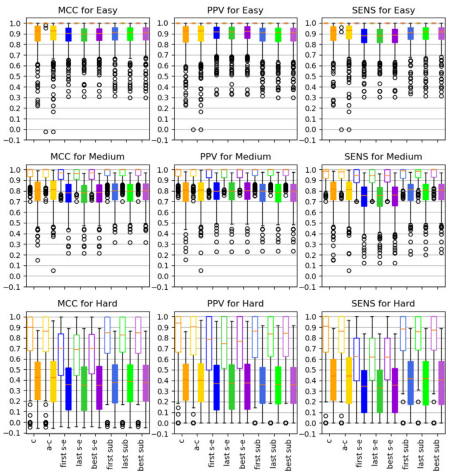

Figure 6. Boxplots of the MCC, PPV and SENS scores to assess the prediction accuracy of aliFreeFoldMulti strategies for the three datasets 'Easy,' 'Medium,' and 'Hard' of the large RNA-families dataset. The *x*-axis displays the four aliFreeFoldMulti strategies: centroid (c), adjusted-centroid (a–c), stem-embedding (s–e), and closest-suboptimal (sub). For each strategy, the left/empty (resp. right/full) boxplot represents the distribution of the maximum (resp. average) score. Structure prediction strategies and sub-strategies are described under 'Materials and Methods' section.

the accuracy obtained using the initial aliFreeFold algorithm. The first dataset named 'Easy' consists of the 357 (respectively 10) families for which the MCC score between the representative RNA structure predicted by aliFreeFold and the expected structure equaled 1. The second dataset referred to as 'Medium' consists of the 289 (respectively 13) families for which the MCC score fell between 0.7 and 1 (excluded). The third dataset labeled 'Hard' consists of the 217 (respectively 7) remaining families for which the MCC was <0.7. Figure 6 shows the boxplots representing maximum and average score (MCC, PPV and SENS) distributions in families for each dataset and each strategy on the large dataset. As expected, the splitting of the initial dataset into three datasets drastically reduced the variance of the MCC, PPV and SENS statistics in each dataset. The results in Figure 6 show that all strategies achieved higher accuracy with the 'Easy' dataset (MCC median: ∼0.9) than with the 'Medium' (MCC median: ∼0.8) and 'Hard' (MCC median: ∼0.4) dataset. We observed similar results for the small dataset in Additional File 1, Supplementary Figure S2.

A strong decrease of the sensitivity of the stem-embedding strategy. A comparison of the various strategies reveals that the stem-embedding strategy performed almost as well as the other strategies for the 'Easy' and 'Medium' datasets, but it predicted significantly less accurate structures for the 'Hard' dataset. This is explained by a strong decrease of the SENS score, especially for the maximum score.

The time efficiency of all strategies were comparable. Supplementary Figure S1B in Additional File 1 shows that the execution times of all strategies are very similar. Most of the time spent is for the computation of the RNA-family representative structure (aliFreeFold algorithm).

Performances of aliFreeFoldMulti and the five selected methods

The second evaluation consisted in comparing the prediction results of the best-performing aliFreeFoldMulti strategy (the centroid strategy) with five existing RNA folding methods: FoldalignM (15,16), TurboFoldII (17), CentroidAlifold (23), RNAalifold (20) and RNAspa (24). For each family, RNAspa, FoldalignM and TurboFoldII take a FASTA file containing the RNA sequences of the family as input. CentroidAlifold and RNAalifold require a multiple sequence alignment of the family as input. For the latter two, we used the same multiple sequence alignments of RNA families computed with MAFFT (32) with parameters that consider RNA folding. The MCC, PPV and SENS scores between the predicted and expected structures of each sequence of each family were computed for each method. Figure 7A provides two boxplots representing the maximum and average score distributions in families for

2.3. ARTICLE «ALIFREEFOLDMULTI : ALIGNMENT-FREE METHOD TO PREDICT SECONDARY STRUCTURES OF MULTIPLE RNA HOMOLOGS»

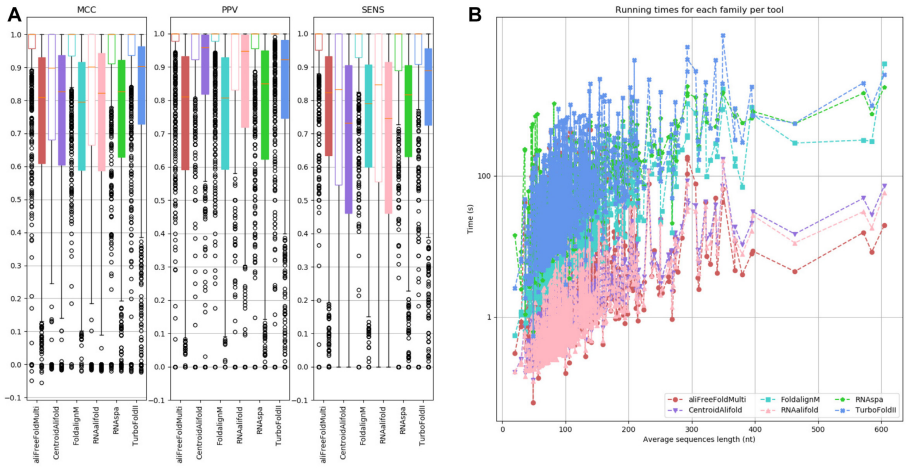

Figure 7. (**A**) Boxplots of the MCC, PPV and SENS scores compared to expected structures on the large dataset of RNA families to assess the prediction accuracy of the six methods: aliFreeFoldMulti, CentroidAlifold, RNAalifold, RNAspa, FoldalignM, and TurboFoldII. For each method, the left/empty (resp. right/full) boxplot represents the distribution of the maximum (resp. average) score. (**B**) Running time analysis (for average sequence length in families).

each score (i.e. MCC, PPV and SENS) and each method for the large dataset. Figure 7B shows the execution times of each method for increasing average sequence lengths in the large dataset. Supplementary Figure S3A and B in Additional File 1 provide the same results for the small dataset.

aliFreeFoldMulti achieves the highest maximum MCC scores and the lowest computing times. TurboFoldII obtains the highest average MCC and SENS scores, while aliFreeFoldMulti obtains the highest maximum MCC, PPV and SENS scores. The two align-then-fold methods CentroidAlifold and RNAalifold obtain the highest maximum and average PPV scores, but also the lowest maximum and average SENS scores with a high variance (Figure 7A). The methods can be separated in three groups in terms of execution time. The first group consists of the 'align-and-fold' approaches (TurboFoldII and FoldalignM) and the 'fold-then-align' approach (RNAspa) which are the most time consuming. The second group consists of the align-then-fold methods (RNAalifold and CentroidAlifold). The third category contains aliFreeFoldMulti, which is the fastest (Figure 7B).

The accuracies of all methods correlates with aliFreeFold accuracy. Figure 8 shows the boxplots representing maximum and average scores (MCC, PPV and SENS) distributions in families for each dataset described in the previous section (i.e. 'Easy', 'Medium' and 'Hard') and each method. We observe that, for all methods, the MCC, PPV and SENS scores decreased unidirectionally from the 'Easy' to the 'Hard' datasets. For all datasets, TurboFoldII always has the highest average MCC scores, and aliFreeFoldMulti always has the highest maximum MCC scores.

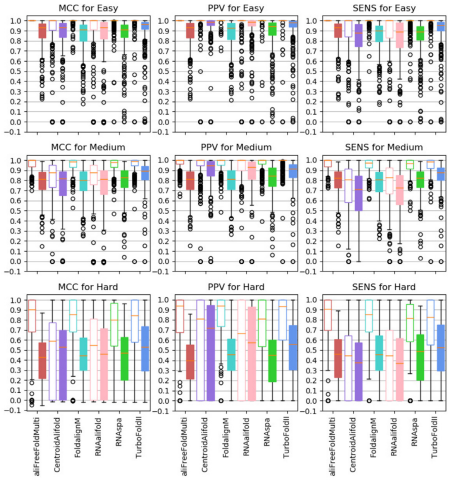

Figure 8. Boxplots of the MCC, PPV and SENS scores to assess the prediction accuracy of aliFreeFoldMulti and the other five selected RNA structure prediction methods for the three datasets 'Easy,' 'Medium' and 'Hard' of the large RNA-families dataset. The *x*-axis displays the six methods. For each method, the left/empty (resp. right/full) boxplot represents the distribution of the maximum (resp. average) score.

2.3. ARTICLE «ALIFREEFOLDMULTI : ALIGNMENT-FREE METHOD TO PREDICT SECONDARY STRUCTURES OF MULTIPLE RNA HOMOLOGS»

DISCUSSION

Summary of results

Centroid and adjusted-centroid are the best strategy for aliFreeFoldMulti, while stem-embedding is the worst. Analysis of the results for the various aliFreeFoldMulti strategies shows that the simplest solutions, the centroid and the adjusted-centroid strategies, yielded the best results. In particular, the centroid strategy yielded the highest maximum MCC, PPV and SENS scores. The other strategies developed with the aim to improve accuracy by searching out similar structures of homologous RNA (adjusted-centroid strategy) or by searching for structures that are similar to the representative structure (stem-embedding and closest-suboptimal strategies) did not outperform the centroid strategy. Yet, the results are enlightening. In particular, the stem-embedding strategy yielded the worst results. It returned artificial structures that might combine stems from different suboptimal structures. The low performance of the stem-embedding strategy shows the importance to use suboptimal structures, and suggests that the predicted structure should always be chosen from within the set of suboptimal structures.

aliFreeFoldMulti: high maximum MCC scores and low computing times. Comparison of the six RNA structure prediction methods shows that aliFreeFoldMulti achieved the highest maximum accuracy scores and the best time efficiency. TurbofoldII outperformed aliFreeFoldMulti in terms of average accuracy scores, but it requested more time. The median (resp. average) execution time for TurboFoldII is 214.8 (resp. 600.9) seconds, compared to 20.4 seconds (resp. 48.5) for aliFreeFoldMulti for the 30 noncoding RNA-families dataset. Like TurboFoldII, FoldalignM and RNAspa were among the most time-consuming methods.

Analysis of the Three RNA-family subsets: 'Easy,' 'Medium' and 'Hard'

Splitting the RNA-families datasets into three subsets made it possible to reduce the high variance in the RNA folding accuracy of the various strategies of aliFreeFoldMulti and the other five methods assessed. The accuracy of RNA folding with aliFreeFoldMulti and the five methods selected correlated with the accuracy of aliFreeFold. We conducted further analyses to understand the causes of the different performances of the methods on the three RNA family subsets, with the aim to find new directions for the improvement of aliFreeFoldMuli.

Number of sequences and average sequence length do not fully explain the difference between 'Easy,' 'medium,' and 'Hard' RNA-family subsets. To better characterize the three RNA-family subsets, we analyzed the distribution of the average sequence length and the number of sequences for the three subsets from the small and large RNA-families datasets (Supplementary Figures S5 and 6, Additional File 1). The three subsets can be partially distinguished based on average sequence length. The 'Easy' families had an average sequence length shorter than the 'Medium' and 'Hard' families (Supplementary Figure S6, Additional File 1). As for the number of sequences, all three datasets had a similar median number of sequences per family (Supplementary Figure S6, Additional File 1). Since average sequence length is the most discriminating criterion, we plotted the distribution of RNA-sequence length of the 30 families from the small dataset, ordered from the best to the lowest MCC score for the RNA consensus structure predicted by aliFreeFold (see Supplementary Figure S7, Additional File 1). We observed no correlation between sequence length and MCC values. Moreover, we can observe that some families with relatively high sequence lengths have high MCC values, such as 'RF00168+Lysine' (37 sequences; median sequence length: ∼175 nt and MCC = 1.0) and 'RF00012 + U3' (17 sequences; median sequence length: ∼225 nt and MCC = 0.923). Therefore, number of sequences and average sequence length criterion were not sufficient to fully characterize the three subsets.

Distribution of pairwise distances between the suboptimal structures partially explains the difference between the 'Easy,' 'Medium' and 'Hard' subsets. We conducted an additional analysis to determine if the distribution of pairwise distances between the sampled suboptimal structures could better characterize the three subsets 'Easy', 'Medium' and 'Hard'. For each sequence, we computed the average of the pairwise distances between the 25 suboptimal structures sampled. We then plotted the distribution of this average for each family ordered from the best to the lowest MCC value for the RNA representative structure predicted by aliFreeFold for the small and large RNA-families datasets (Supplementary Figures S8 and 9, Additional File 1). Results show that the pairwise distances between suboptimal structures for the 'Hard' dataset are in average higher than for the 'Easy' and 'Medium' datasets. This suggests that the more variability there is between suboptimal structures, the less accurate the prediction of the structure.

Accuracy of the best RNA suboptimal structure explains the difference between the 'Easy,' 'Medium' and 'Hard' subsets. aliFreeFoldMulti is based on the hypothesis that, in a sample of the 25 suboptimal structures for each sequence, there is at least one suboptimal structure that has, on average, 80% correct base pairs (27). Additional results produced on the small and large datasets of RNA families, show that this hypothesis did not hold true for all RNA-families. Supplementary Figures S10 and 11 in Additional File 1 plot the distribution of the best MCC score of the suboptimal structures per sequence for each family from the small and large datasets. For the 'Easy' subset, most of the families had a median of the maximum MCC scores distribution of 1.0 with a very low variance. For 'Medium' families, the median fell between 0.7 and 1.0. For 'Hard' families, the median ranged between 0.4 and 0.9. Most of the RNA sequences in the 'Hard' families had lower than the expected 80% correct base pairs (Additional File 1: Supplementary Figure S10), which explains the low average scores of aliFreeFoldMulti for these families. However, we observe that for each family, there is at least one sequence with one suboptimal structure that has more than 80% correct base pairs. This explains the high maximum accuracy scores of aliFreeFoldMulti, and the previously reported outperformance of al-

2.3. ARTICLE «ALIFREEFOLDMULTI : ALIGNMENT-FREE METHOD TO PREDICT SECONDARY STRUCTURES OF MULTIPLE RNA HOMOLOGS»

iFreeFold for predicting representative consensus structures (26). Thus, we can conclude that the prediction accuracy of aliFreeFold and aliFreeFoldMulti is strongly related to the accuracy of the set of suboptimal structures generated. On the other hand, supplementary Figure S10 in Additional File 1 also shows that for the 'Easy' families, aliFreeFoldMulti predicted RNA suboptimal structures that were not the most accurate generated. For most sequences in the 'Easy' dataset, the highest MCC score was 1.0, while the average MCC score of aliFreeFoldMulti was ∼0.85 (Supplementary Figure S4, Additional File 1). Therefore, there is still room for improvement in the prediction accuracy of aliFreeFoldMulti, while preserving its low computing times.

Conclusion and perspectives

We described an alignment-free method named aliFeeFoldMulti and its four strategies to predict secondary structures of multiple RNA homologs. aliFeeFoldMulti is an extension of the aliFreeFold algorithm that was previously developed to predict a representative secondary structure of multiple RNA homologs by using a vector representation of their suboptimal structures. Among the strategies developed in aliFeeFoldMulti, we showed that the centroid-based strategies were the best to predict secondary structures for all sequences of a RNA family. Yet, the analysis of the other two strategies, namely the stem-embedding and the closest-suboptimal strategies, allowed to highlight the importance of the use of suboptimal structures rather than artificial structures. The comparison of the performances of aliFreeFoldMulti to the five selected other RNA structure prediction methods showed that aliFreeFoldMulti is the fastest and best performing method in terms of maximum MCC score. In terms of average MCC scores, TurboFoldII is the best performing methods, while aliFreeFoldMulti achieve performances that are comparable to the four others approaches. The splitting of the initial RNA-families dataset into three datasets based on the MCC score of the consensus structure predicted by aliFreeFold allowed to show that all methods had the same dynamic on the RNA structure prediction accuracy.

The results herein show that there is a significant potential for improving aliFreeFoldMulti to obtain more accurate predictions of RNA structure by using a more appropriate approach for exploring the set of suboptimal structures. This improved exploration of suboptimal structures would lead to more accurate results in average while maintaining low computation times. We showed that the selection of the first 25 suboptimal structures is not always sufficient to obtain the most accurate predictions in average (Supplementary Figures S10 and 11, Additional file 1). Therefore, we also need to define intermediate criteria and methods to better characterize RNA families in order to define suboptimal structure sampling strategies according to the characteristics of each RNA family. Another future direction is to refine the aliFreeFoldMulti strategy in order to always determine the most accurate suboptimal structure among the set of suboptimal structures generated.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

The authors thank the members of the CoBiUS lab at the University of Sherbrooke for their helpful constructive discussion.

FUNDING

This work was supported by the Canada Research Chair (CRC Tier 2 Grant 950-230577), and the The Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant RGPIN-2017-05552).
Conflict of interest statement. None declared.

REFERENCES

1. Mattick,J.S. (2001) Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.*, **2**, 986–991.
2. Werner,F. (2007) Structure and function of archaeal RNA polymerases. *Mol. Microbiol.*, **65**, 1395–1404.
3. Serganov,A. and Patel,D.J. (2007) Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nat. Rev. Genet.*, **8**, 776–790.
4. Moore,P.B. and Steitz,T.A. (2011) The roles of RNA in the synthesis of protein. *CSH Perspect. Biol.*, **3**, a003780.
5. Mattick,J.S. (2003) Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, **25**, 930–939.
6. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
7. Mathews,D.H., Moss,W.N. and Turner,D.H. (2010) Folding and finding RNA secondary structure. *CSH Perspect. Biol.*, **2**, a003665.
8. Lorenz,R., Bernhart,S.H., Höner zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
9. Trotta,E. (2014) On the normalization of the minimum free energy of RNAs by sequence length. *PLoS One*, **9**, e113380.
10. Doshi,K.J., Cannone,J.J., Cobaugh,C.W. and Gutell,R.R. (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, **5**, 105.
11. Lalwani,S., Kumar,R. and Gupta,N. (2014) Sequence-structure alignment techniques for RNA: a comprehensive survey. *Adv. Life Sci.*, **4**, 21–35.
12. Puton,T., Kozlowski,L.P., Rother,K.M. and Bujnicki,J.M. (2013) CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. *Nucleic Acids Res.*, **41**, 4307–4323.
13. Wright,E.S. (2020) RNAconTest: comparing tools for noncoding RNA multiple sequence alignment based on structural consistency. *RNA*, **26**, 531–540.
14. Sankoff,D. (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Dyn. Syst.*, **45**, 810–825.
15. Sundfeld,D., Havgaard,J.H., de Melo,A.C.M.A. and Gorodkin,J. (2016) Foldalign 2.5: multithreaded implementation for pairwise structural RNA alignment. *Bioinformatics*, **32**, 1238–1240.
16. Torarinsson,E., Havgaard,J.H. and Gorodkin,J. (2007) Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, **23**, 926–932.
17. Tan,Z., Fu,Y., Sharma,G. and Mathews,D.H. (2017) TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Res.*, **45**, 11570–11581.
18. Fu,Y., Sharma,G. and Mathews,D.H. (2014) Dynalign II: common secondary structure prediction for RNA homologs with domain insertions. *Nucleic Acids Res.*, **42**, 13939–13948.
19. Will,S., Otto,C., Miladi,M., Möhl,M. and Backofen,R. (2015) SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. *Bioinformatics*, **31**, 2489–2496.

2.3. ARTICLE «ALIFREEFOLDMULTI : ALIGNMENT-FREE METHOD TO PREDICT SECONDARY STRUCTURES OF MULTIPLE RNA HOMOLOGS»

20. Bernhart,S.H., Hofacker,I.L., Will,S., Gruber,A.R. and Stadler,P.F. (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.
21. Sato,K., Hamada,M., Asai,K. and Mituyama,T. (2009) CentroidFold: a web server for RNA secondary structure prediction. *Nucleic Acids Res.*, **37**, W277–W280.
22. Wiebe,N.J.P. and Meyer,I.M. (2010) Transat—a method for detecting the conserved helices of functional RNA structures, including transient, pseudo-knotted and alternative structures. *PLOS Comput. Biol.*, **6**, e1000823.
23. Hamada,M., Sato,K. and Asai,K. (2011) Improving the accuracy of predicting secondary structure for aligned RNA sequences. *Nucleic Acids Res.*, **39**, 393–402.
24. Horesh,Y., Doniger,T., Michaeli,S. and Unger,R. (2007) RNAspa: a shortest path approach for comparative prediction of the secondary structure of ncRNA molecules. *BMC Bioinformatics*, **8**, 366.
25. Reeder,J. and Giegerich,R. (2005) Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics*, **21**, 3516–3523.
26. Glouzon,J.-P.S. and Ouangraoua,A. (2018) aliFreeFold: an alignment-free approach to predict secondary structure from homologous RNA sequences. *Bioinformatics*, **34**, i70–i78.
27. Zuker,M., Jaeger,J.A. and Turner,D.H. (1991) A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nucleic Acids Res.*, **19**, 2707–2714.
28. Glouzon,J.-P.S., Perreault,J.-P. and Wang,S. (2017) The super-n-motifs model: a novel alignment-free approach for representing and comparing RNA secondary structures. *Bioinformatics*, **33**, 1169–1178.
29. Gardner,P.P., Wilm,A. and Washietl,S. (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.*, **33**, 2433–2439.
30. Tabei,Y., Kiryu,H., Kin,T. and Asai,K. (2008) A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, **9**, 33.
31. Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimic. Biophys. Acta*, **405**, 442–451.
32. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.

aliFreeFoldMulti: Additional File 1

Table S1: Statistics of the small dataset of the 30 RNA families used in the experimentation.

Family	Nb of seq.	Avg. PID	Avg. seq. length
BRALIBASE II [Gardner et al., 2005]			
g2intron	70	65.882±8.062	83.129±21.266
rRNA	98	63.347±9.254	117.551±2.483
tRNA	72	58.118±10.699	72.306±2.891
U5	80	70.518±11.145	118.162±4.853
MXSCARNA [Tabei et al., 2008]			
RF00002-5-8S_rRNA	46	71.941±9.16	154.065±6.198
RF00003-U1	42	67.681±10.378	158.048±7.73
RF00004-U2	51	72.533±8.144	185±17.233
RF00008-Hammerhead_3	54	73.695±12.659	55.593±6.356
RF00011-RNaseP_bact_b	19	68.409±7.725	391.105±20.08
RF00012-U3	17	67.099±10.473	246.529±48.758
RF00015-U4	25	74.138±10.812	141.4±8.718
RF00017-SRP_euk_arch	49	58.704±9.567	294.49±11.518
RF00019-Y	16	73.112±10.184	94.688±11.898
RF00023-tmRNA	36	60.047±6.007	374.611±22.09
RF00024-Telomerase	24	69.91±9.578	463.125±32.332
RF00025-Telomerase	16	69.203±10.443	168.938±16.027
RF00031-SECIS	49	54.205±8.691	64.592±3.278
RF00037-IRE	37	67.658±17.802	28.757±1.442
RF00045-U17	23	75.837±8.466	214.174±7.088
RF00050-RFN	28	69.952±5.845	150.179±11.845
RF00162-S_box	17	72.446±6.275	128.941±17.594
RF00163-Hammerhead_1	21	98.078±1.195	115.095±5.281
RF00164-s2m	37	78.798±9.912	42.919±0.682
RF00167-Purine	35	62.371±5.52	99.571±0.884
RF00168-Lysine	37	59.646±5.828	179.838±6.56
RF00169-SRP_bact	55	62.417±8.332	95.309±8.613
RF00181-sno_14q_I_II	42	70.603±9.011	73.976±4.027
RF00233-Tymo_tRNA	28	71.698±10.734	82.643±2.87
RF00236-ctRNA_pGA1	17	77.224±12.086	80.294±1.724
RF00436-UnaL2	60	78.47±9.396	54.4±1.861

References

[Gardner et al., 2005] Gardner, P. P., Wilm, A., and Washietl, S. (2005). A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Research*, 33(8):2433–2439.

[Tabei et al., 2008] Tabei, Y., Kiryu, H., Kin, T., and Asai, K. (2008). A fast structural multiple alignment method for long RNA sequences. *BMC Bioinformatics*, 9(1):33.

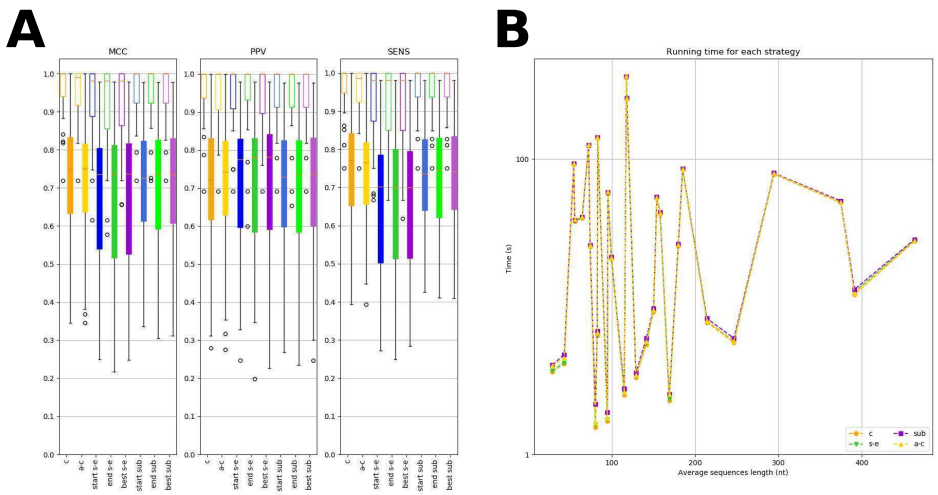

Figure S1: (A) Boxplots of the MCC, PPV, and SENS scores compared to expected structures on the small RNA-families dataset to assess the prediction accuracy of aliFreeFoldMulti strategies. The x-axis displays the four aliFreeFoldMulti strategies: centroid (c), adjusted-centroid (a-c), stem-embedding (s-e), and closest-suboptimal (sub). For stem-embedding and closest-suboptimal there are three different results corresponding to the substrategy used: "start," "end," or "best." For each strategy, the left/empty (resp. right/full) boxplot represents the distribution of the maximum (resp. average) score. **(B)** Running time analysis for average sequence length in families. Structure prediction strategies and sub-strategies are described under MATERIALS AND METHODS.

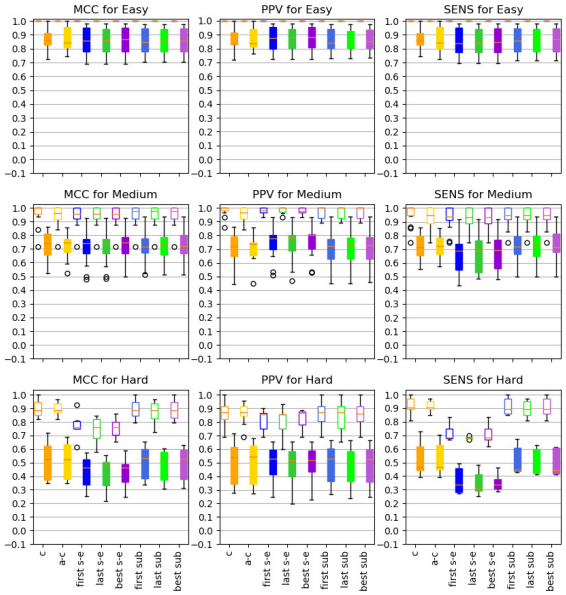

Figure S2: Boxplots of the MCC, PPV, and SENS scores to assess the prediction accuracy of aliFreeFoldMulti strategies for the three RNA-family datasets "Easy," "Medium," and "Hard" for the small RNA-families dataset. The x-axis displays the four aliFreeFoldMulti strategies: centroid (c), adjusted-centroid (a-c), stem-embedding (s-e), and closest-suboptimal (sub). For each strategy, the left/empty (resp. right/full) boxplot represents the distribution of the maximum (resp. average) score. Structure prediction strategies and sub-strategies are described under MATERIALS AND METHODS.

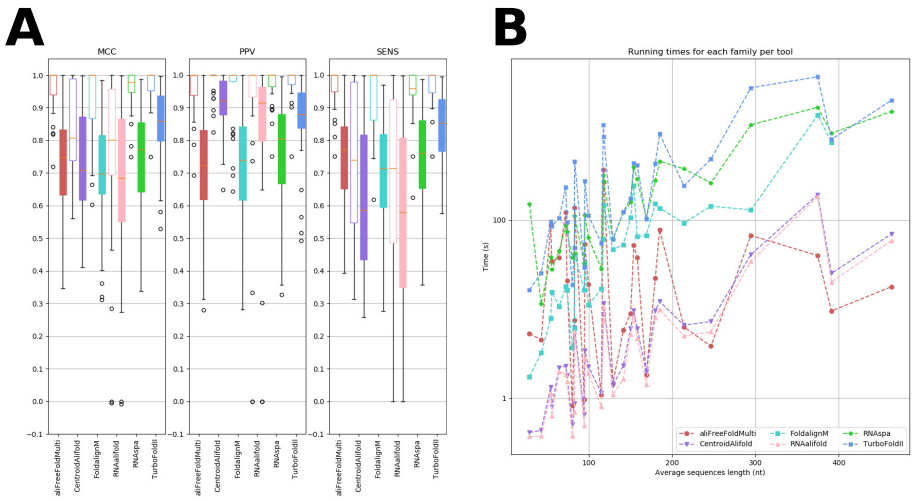

Figure S3: (A) Boxplots of the MCC, PPV, and SENS scores to assess the prediction accuracy of of aliFreeFold-Multi, CentroidAlifold, RNAalifold, RNAspa, FoldalignM, and TurboFoldII on the small RNA-families dataset. The x-axis displays the six methods. For each method, the left/empty (resp. right/full) boxplot represents the distribution of the maximum (resp. average) score. **(B)** Running time analysis for average sequence length in families.

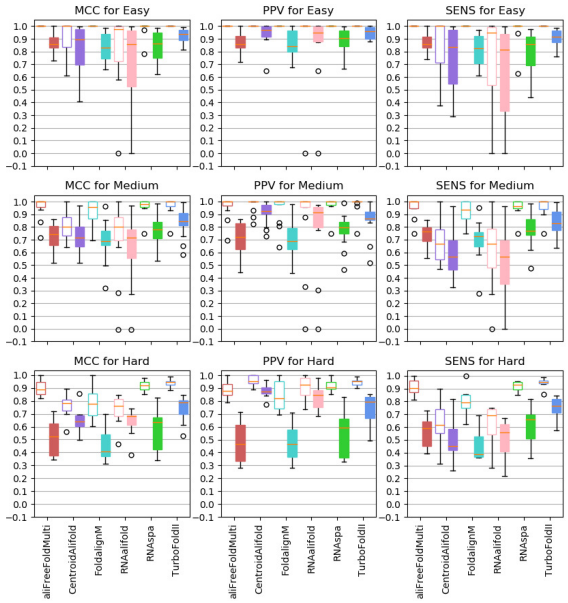

Figure S4: Boxplots of the MCC, PPV, and SENS scores to assess the prediction accuracy of aliFreeFoldMulti and the other five selected RNA structure prediction methods for the three RNA-family datasets ("Easy," "Medium," and "Hard") on the small RNA-families dataset. For each method, the left/empty (resp. right/full) boxplot shows the distribution of the maximum (resp. average) score for each family.

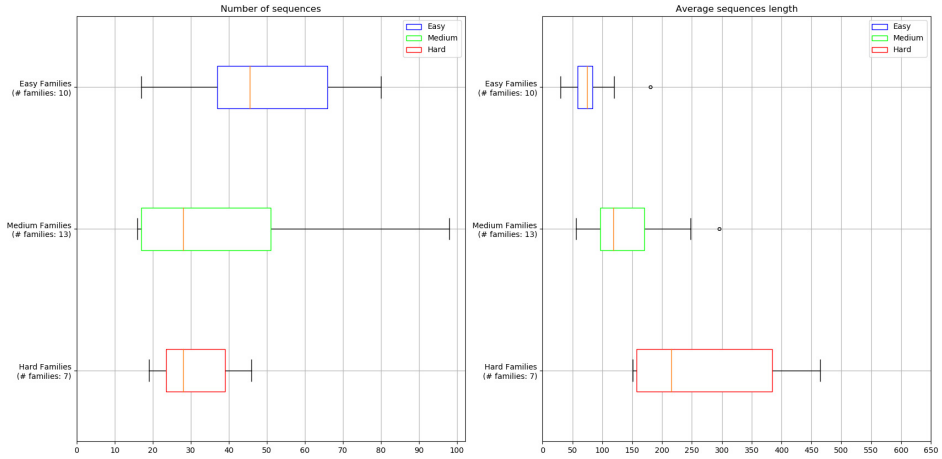

Figure S5: Boxplots representing the number of sequences (left) and average sequence length (right) in the three RNA-family datasets (”Easy,” ”Medium,” and ”Hard”) for the small RNA-families dataset.

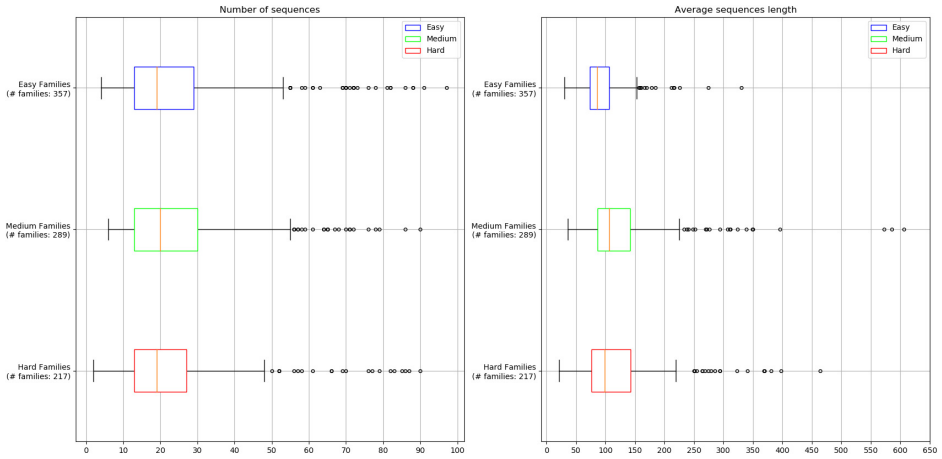

Figure S6: Boxplots representing the number of sequences (left) and average sequence length (right) in the three RNA-family datasets (”Easy,” ”Medium,” and ”Hard”) for the large RNA-families dataset.

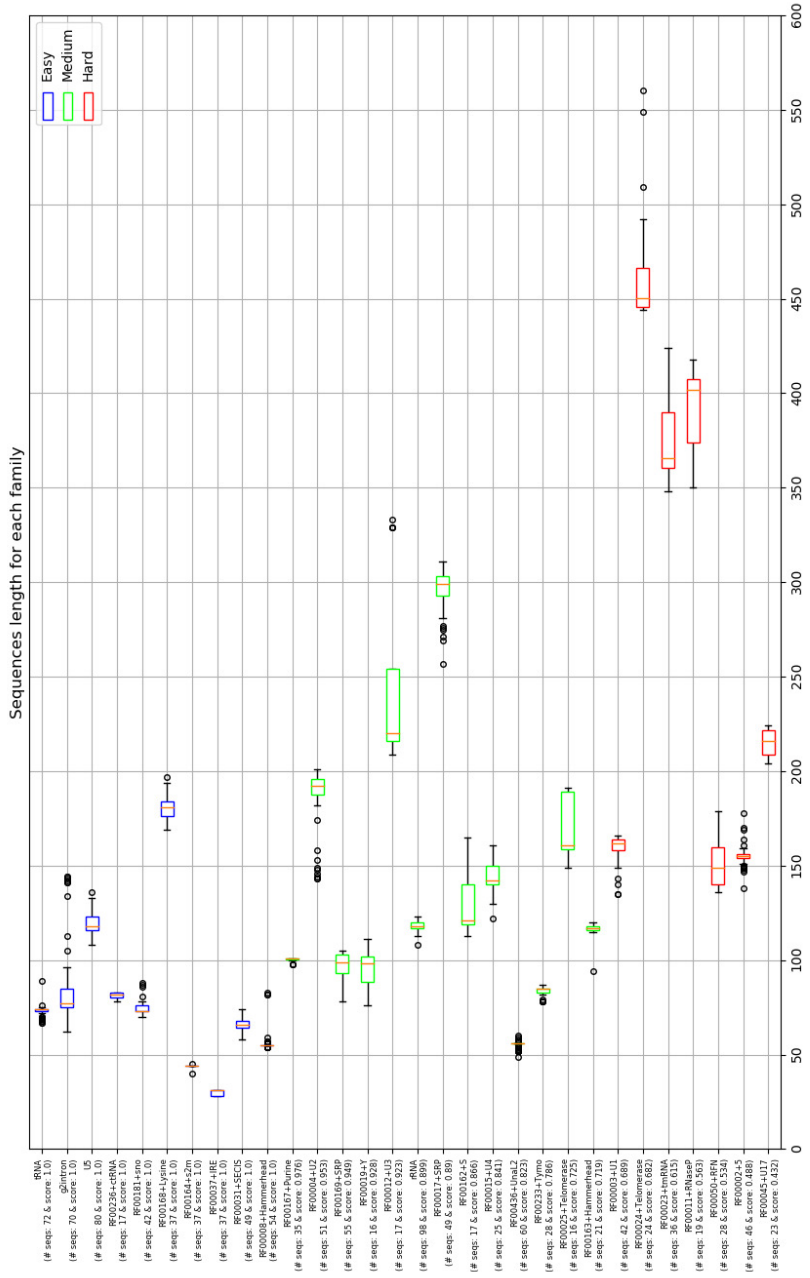

Figure S7: Boxplots representing the sequence length distribution of RNA sequences for the small RNA-families dataset ordered by decreasing MCC score. Each family is annotated with the number of sequences and the MCC score.

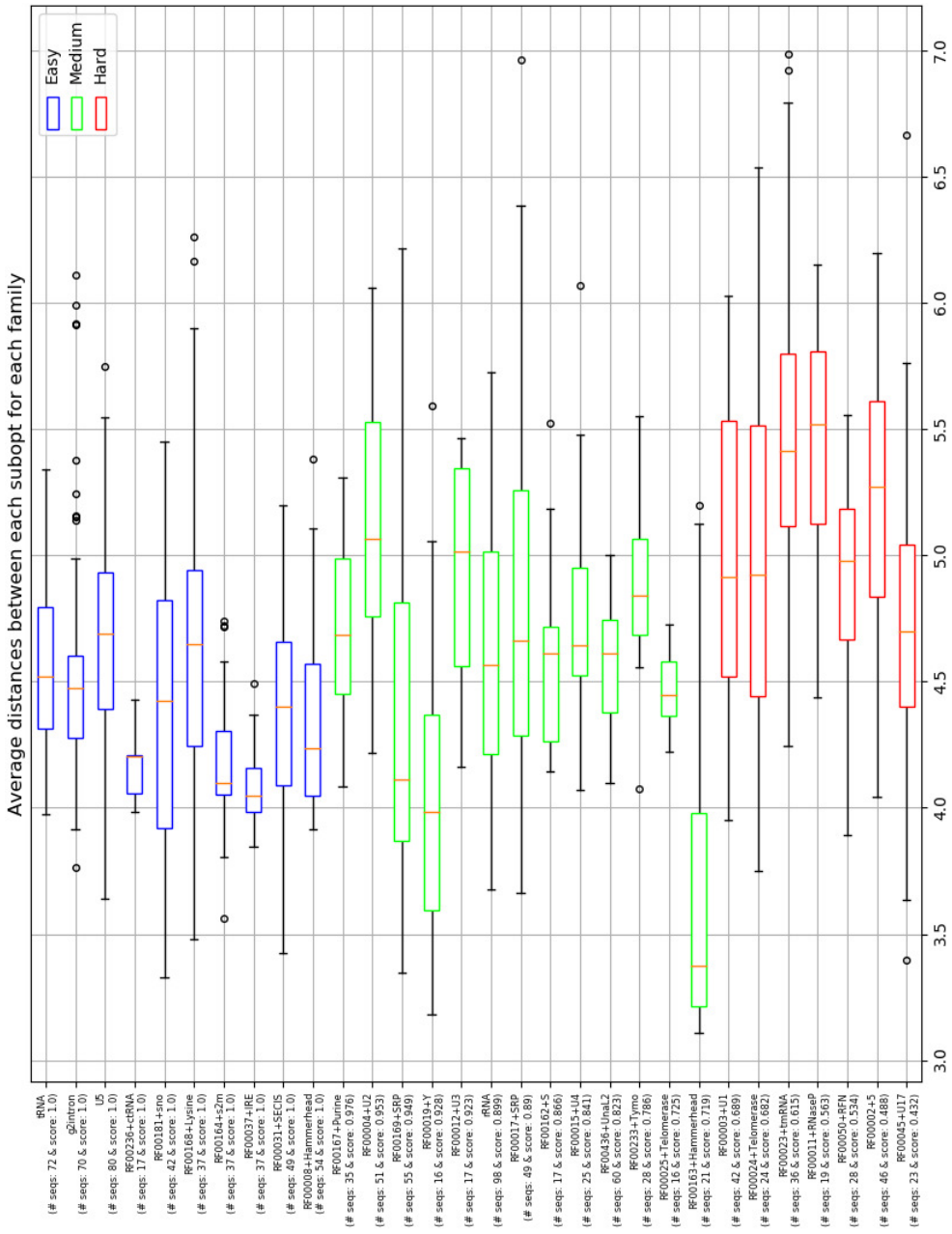

Figure S8: Boxplots representing the average pairwise distances between the 25 suboptimal structures, for each RNA sequence of the small RNA-families dataset, ordered by decreasing MCC score. Each family is annotated with the number of sequences and the MCC score.

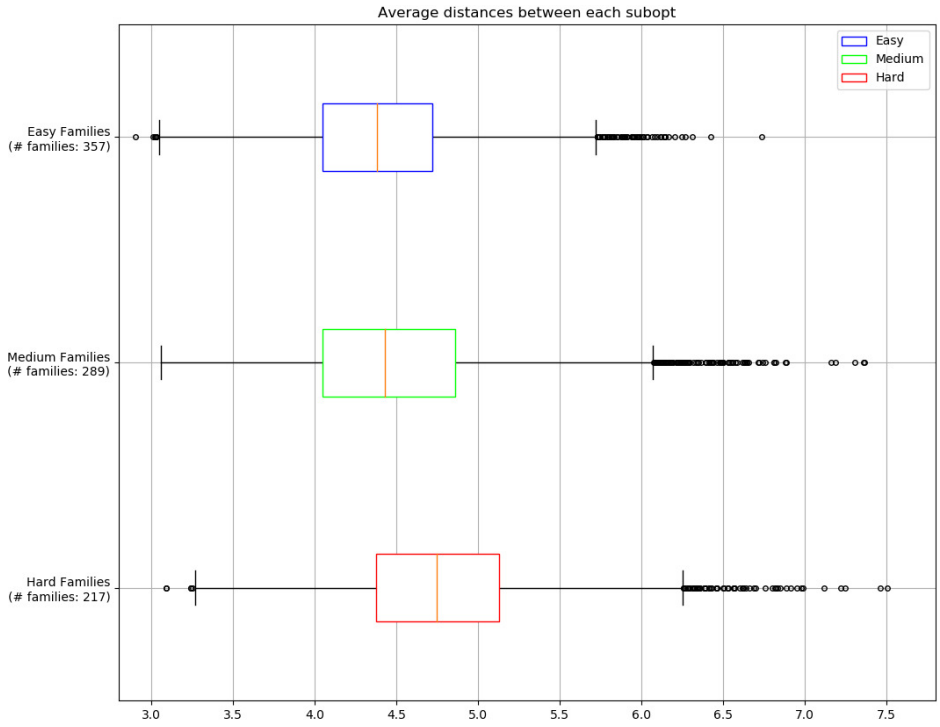

Figure S9: Boxplots representing the average pairwise distances between the 25 suboptimal structures, for each RNA sequence of the three RNA-family datasets ("Easy," "Medium," and "Hard") for the large RNA-families dataset.

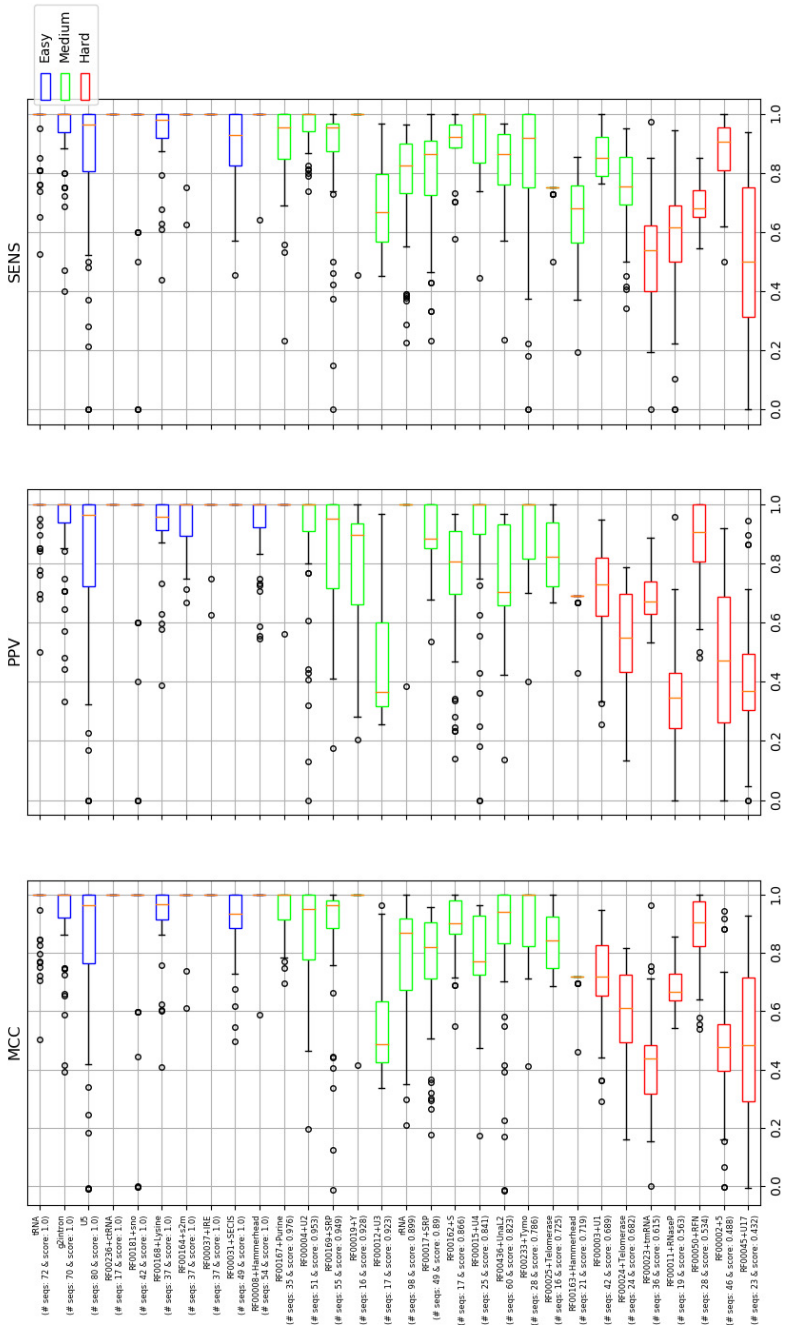

Figure S10: Boxplots representing the distribution of the maximum MCC, PPV and SENS scores for each sequence among the 25 suboptimal structures, in each family of the small RNA-families dataset. "Easy" families appear in blue, "Medium" families in green, and "Hard" families in red.

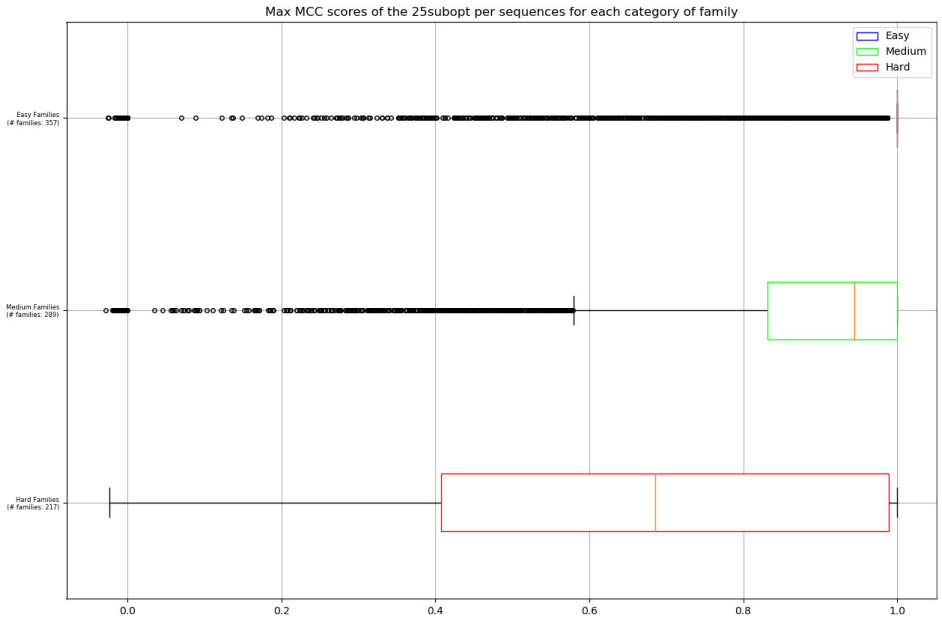

Figure S11: Boxplots representing the distribution of the maximum MCC score for each sequence among the 25 suboptimal structures, for the large RNA-families dataset. "Easy" families appear in blue, "Medium" families in green, and "Hard" families in red.

Conclusion

Au cours de cette maîtrise, nous avons développé, aliFreeFoldMulti, un algorithme permettant de faire la prédiction de structures secondaires de plusieurs ARN homologues. aliFreeFoldMulti est une extension de l'outil aliFreeFold [14]. aliFreeFold permet de prédire la structure secondaire représentative pour un ensemble de séquences d'ARN homologues en utilisant un ensemble de structures secondaires sous-optimales calculé par RNAsubopt[24] et une représentation vectorielle sous forme de n-motifs de chaque structure sous-optimale. Quant à aliFreeFoldMulti, cet outil prédit la structure secondaire de chacune des séquences de l'ensemble de séquences d'ARN homologues en utilisant quatre stratégies. aliFreeFoldMulti présente plusieurs avantages et quelques limites par rapport aux outils existants. Parmi les différents avantages, aliFreeFoldMulti peut prédire rapidement et précisément une structure secondaire pour chacune des séquences d'une famille d'ARNnc donnée en entrée. De plus, aliFreeFoldMulti offre plusieurs stratégies dont les meilleures stratégies, soient celles qui donnent de meilleurs scores, sont «centroïde», «centroïde ajusté» et «plus proche du sous-optimal». (*Voir la figure 5 de l'article aliFreeFoldMulti.*) Comparativement aux autres outils de prédiction de structures secondaires pour plusieurs séquences d'ARN, aliFreeFoldMulti est l'outil le plus rapide et retourne les meilleurs scores maximums. Cependant, la principale limite d'aliFreeFoldMulti est qu'il ne retourne pas les meilleurs scores en moyenne. (*Voir la figure 7 de l'article aliFreeFoldMulti.*)

En observant et analysant plus en détails les différents résultats obtenus pour aliFreeFoldMulti, il est possible de mieux comprendre les comportements et les limites de cet outil. En se basant sur la comparaison entre le score MCC de la structure secondaire consensus représentative résultant de aliFreeFold et le score MCC de la

structure secondaire consensus attendue d'après la base de données Rfam pour chacune des familles d'ARN données en entrée, le jeu de données d'évaluation a été divisé en trois sous-ensembles de familles. Le sous-ensemble «Facile» (*Easy*) regroupe les familles ayant un score MCC parfait, c'est-à-dire 1.00. Le sous-ensemble «Moyen» (*Medium*) regroupe les familles ayant un score MCC entre 0.70 et 0.99 inclusivement. Le sous-ensemble «Difficile» (*Hard*) regroupe les familles ayant un score MCC inférieur à 0.70. Ainsi, sur l'ensemble des 863 familles d'ARN utilisées pour l'évaluation, 357 se retrouvent dans le sous-ensemble «Facile», 289 dans le sous-ensemble «Moyen» et 217 dans le sous-ensemble «Difficile». Suite à ce découpage en 3 sous-ensembles, il est possible d'extraire davantage de conclusions sur le comportement de l'outil aliFreeFoldMulti, en ce qui concerne la comparaison des différentes stratégies offertes par aliFreeFoldMulti, soient «centroïde», «centroïde ajusté», «plongement des tiges» et «plus proche du sous-optimal». Il est possible d'extraire également des conclusions sur la comparaison d'aliFreeFoldMulti par rapport aux autres outils de prédiction de structures secondaires d'ARN homologues.

De façon générale, les différents scores, c'est-à-dire MCC, PPV et SENS, sont meilleurs pour les familles du sous-ensemble «Facile» que les familles du sous-ensemble «Moyen» et que les familles du sous-ensemble «Difficile». Donc, les résultats de aliFreeFoldMulti sont corrélés à ceux de aliFreeFold. Ensuite, lorsqu'on observe les scores des structures sous-optimales générés par RNAsubopt, on observe que les scores sont plus élevés et comportent le moins de variation pour les familles du sous-ensemble «Facile» et que les scores sont plus faibles et comportent le plus de variation pour les familles du sous-ensemble «Difficile». Donc, la qualité des résultats et des scores d'aliFreeFoldMulti semble corréler à la qualité des structures secondaires sous-optimales calculées par l'outil RNAsubopt. (*Voir la figure S11 du matériel supplémentaire de l'article aliFreeFoldMulti.*)

En ce qui concerne les résultats de la comparaison des quatre stratégies d'aliFreeFoldMulti, les meilleures stratégies sont «centroïde», «centroïde ajusté» et «plus proche du sous-optimal» et la pire stratégie est «plongement des tiges» en se basant sur les scores de l'ensemble des familles et les sous-ensembles des familles. Donc, les structures secondaires fictives donnent des résultats moins bons et moins précis que les structures secondaires sous-optimales générées par RNAsubopt. (*Voir la figure 5*

de l'article aliFreeFoldMulti.)

Pour les résultats de la comparaison d'aliFreeFoldMulti avec les cinq outils de prédiction de structures secondaires d'ARN homologues, aliFreeFoldMulti est celui ayant le meilleur score MCC maximal et les temps de calcul les plus faibles. CentroidAlifold et RNAalifold sont les outils ayant les meilleurs scores PPV moyens. Ces outils sont des méthodes «aligner-puis-replier» pour la prédiction de structures secondaires d'ARN. TurboFoldII est l'outil ayant les meilleurs scores MCC et SENS moyens, mais les temps de calcul les plus élevés. Cet outil est une méthode «aligner-et-replier» pour la prédiction de structures secondaires d'ARN. Ces résultats permettent de conclure qu'il faut améliorer les stratégies d'échantillonnage d'aliFreeFoldMulti et de recherche de la meilleure structure secondaire sous-optimale. (*Voir la figure 7 de l'article aliFreeFoldMulti.*)

Finalement, en ce qui concerne les perspectives d'amélioration de l'outil aliFreeFoldMulti, cet outil est basé sur l'hypothèse qu'il est possible de trouver parmi les 25 premières structures secondaires sous-optimales, une structure ayant 80% de similarité d'appariement des bases avec la structure secondaire réelle [52]. Cependant, avec les résultats d'aliFreeFoldMulti, on constate à l'aide de la figure S11 que cette hypothèse n'est pas toujours valide, car pour les familles du sous-ensemble «Difficile» la médiane se retrouve à environ 70% qui est inférieur à 80%. Ainsi, afin d'améliorer l'outil de prédiction de structures secondaire d'ARN homologues aliFreeFoldMulti, il faudrait déterminer une meilleure stratégie de génération et d'analyse des structures secondaires sous-optimales. Par exemple, il serait possible d'augmenter l'échantillon utilisé pour les structures secondaires sous-optimales calculées par RNAsubopt, soit d'augmenter le nombre de structures secondaires générées par RNAsubopt pour une séquence.

Bibliographie

[1] M. Andronescu, A. Condon, H. H. Hoos, D. H. Mathews, et K. P. Murphy, « Efficient parameter estimation for RNA secondary structure prediction, » *Bioinformatics*, vol. 23, no. 13, pp. i19–i28, 2007.

[2] D. P. Bartel, « MicroRNAs : genomics, biogenesis, mechanism, and function, » *cell*, vol. 116, no. 2, pp. 281–297, 2004.

[3] M.-A. Bossanyi, V. Carpentier, J.-P. S. Glouzon, A. Ouangraoua, et Y. Anselmetti, « aliFreeFoldMulti : alignment-free method to predict secondary structures of multiple RNA homologs, » *NAR Genomics and Bioinformatics*, vol. 2, no. 4, p. lqaa086, 2020.

[4] S. H. Bernhart, I. L. Hofacker, S. Will, A. R. Gruber, et P. F. Stadler, « RNAalifold : improved consensus structure prediction for RNA alignments, » *BMC bioinformatics*, vol. 9, no. 1, p. 474, 2008.

[5] A. Bremges, S. Schirmer, et R. Giegerich, « Fine-tuning structural RNA alignments in the twilight zone, » *BMC bioinformatics*, vol. 11, no. 1, pp. 1–8, 2010.

[6] E. P. Consortium *et al.*, « Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, » *nature*, vol. 447, no. 7146, p. 799, 2007.

[7] F. Crick, « Codon-anticodon pairing : the wobble hypothesis, » 1966.

[8] K. J. Doshi, J. J. Cannone, C. W. Cobaugh, et R. R. Gutell, « Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction, » *BMC bioinformatics*, vol. 5, no. 1, p. 105, 2004.

[9] Y. Ding, C. Y. Chan, et C. E. Lawrence, « RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble, » *Rna*, vol. 11, no. 8, pp. 1157–1166, 2005.

[10] R. D. Dowell et S. R. Eddy, « Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction, » *BMC bioinformatics*, vol. 5, no. 1, pp. 1–14, 2004.

[11] R. Durbin, S. R. Eddy, A. Krogh, et G. Mitchison, *Biological sequence analysis : probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[12] M. G. Dozmorov, C. B. Giles, K. A. Koelsch, et J. D. Wren, « Systematic classification of non-coding RNAs by epigenomic similarity, » dans *BMC bioinformatics*, vol. 14, no. S14. Springer, 2013, p. S2.

[13] J. K. Dhanoa, R. S. Sethi, R. Verma, J. S. Arora, et C. S. Mukhopadhyay, « Long non-coding RNA : its evolutionary relics and biological implications in mammals : a review, » *Journal of animal science and technology*, vol. 60, no. 1, p. 25, 2018.

[14] J.-P. S. Glouzon et A. Ouangraoua, « aliFreeFold : an alignment-free approach to predict secondary structure from homologous RNA sequences, » *Bioinformatics*, vol. 34, no. 13, pp. i70–i78, 2018.

[15] C. Guthrie et B. Patterson, « Spliceosomal snRNAs, » *Annual review of genetics*, vol. 22, no. 1, pp. 387–419, 1988.

[16] J.-P. S. Glouzon, J.-P. Perreault, et S. Wang, « The super-n-motifs model : a novel alignment-free approach for representing and comparing RNA secondary structures, » *Bioinformatics*, vol. 33, no. 8, pp. 1169–1178, 2017.

[17] P. P. Gardner, A. Wilm, et S. Washietl, « A benchmark of multiple sequence alignment programs upon structural RNAs, » *Nucleic acids research*, vol. 33, no. 8, pp. 2433–2439, 2005.

[18] Y. Horesh, T. Doniger, S. Michaeli, et R. Unger, « RNAspa : a shortest path approach for comparative prediction of the secondary structure of ncRNA molecules, » *BMC bioinformatics*, vol. 8, no. 1, p. 366, 2007.

[19] M. Hamada, K. Sato, et K. Asai, « Improving the accuracy of predicting secondary structure for aligned RNA sequences, » *Nucleic acids research*, vol. 39, no. 2, pp. 393–402, 2011.

[20] M. Ibba et D. Söll, « Aminoacyl-tRNA synthesis, » *Annual review of biochemistry*, vol. 69, no. 1, pp. 617–650, 2000.

[21] M. Karin, « Too many transcription factors : positive and negative interactions. » *The New Biologist*, vol. 2, no. 2, pp. 126–131, 1990.

[22] A. Krämer, « The structure and function of proteins involved in mammalian pre-mRNA splicing, » *Annual review of biochemistry*, vol. 65, no. 1, pp. 367–409, 1996.

[23] D. S. Latchman, « Transcription factors : an overview, » *The international journal of biochemistry & cell biology*, vol. 29, no. 12, pp. 1305–1312, 1997.

[24] R. Lorenz, S. H. Bernhart, C. H. Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, et I. L. Hofacker, « ViennaRNA Package 2.0, » *Algorithms for molecular biology*, vol. 6, no. 1, p. 26, 2011.

[25] M. Losko, J. Kotlinowski, et J. Jura, « Long noncoding RNAs in metabolic syndrome related disorders, » *Mediators of inflammation*, vol. 2016, 2016.

[26] A. Laganà, D. Veneziano, F. Russo, A. Pulvirenti, R. Giugno, C. M. Croce, et A. Ferro, « Computational design of artificial RNA molecules for gene regulation, » dans *RNA Bioinformatics*. Springer, 2015, pp. 393–412.

[27] S. Massenet, E. Bertrand, et C. Verheggen, « Assembly and trafficking of box C/D and H/ACA snoRNPs, » *RNA biology*, vol. 14, no. 6, pp. 680–692, 2017.

[28] E. Martin et R. Hine, *A Dictionary of Biology*. Oxford University Press, 2008. Disponible à https://www.oxfordreference.com/view/10.1093/acref/9780199204625.001.0001/acref-9780199204625

[29] H. Ma, Y. Hao, X. Dong, Q. Gong, J. Chen, J. Zhang, et W. Tian, « Molecular mechanisms and function prediction of long noncoding RNA, » *The Scientific World Journal*, vol. 2012, 2012.

[30] D. H. Mathews, W. N. Moss, et D. H. Turner, « Folding and finding RNA secondary structure, » *Cold Spring Harbor perspectives in biology*, vol. 2, no. 12, p. a003665, 2010.

[31] R. Nussinov, G. Pieczenik, J. R. Griggs, et D. J. Kleitman, « Algorithms for loop matchings, » *SIAM Journal on Applied mathematics*, vol. 35, no. 1, pp. 68–82, 1978.

[32] T. Puton, L. P. Kozlowski, K. M. Rother, et J. M. Bujnicki, « CompaRNA : a server for continuous benchmarking of automated methods for RNA secondary structure prediction, » *Nucleic acids research*, vol. 41, no. 7, pp. 4307–4323, 2013.

[33] A. F. Palazzo et E. S. Lee, « Non-coding RNA : what is functional and what is junk ? » *Frontiers in genetics*, vol. 6, p. 2, 2015.

[34] J. Reeder et R. Giegerich, « Consensus shapes : an alternative to the Sankoff algorithm for RNA consensus structure prediction, » *Bioinformatics*, vol. 21, no. 17, pp. 3516–3523, 2005.

[35] E. Rivas, « The four ingredients of single-sequence RNA secondary structure prediction. A unifying perspective, » *RNA biology*, vol. 10, no. 7, pp. 1185–1196, 2013.

[36] E. Rivas, R. Lang, et S. R. Eddy, « A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more, » *RNA*, vol. 18, no. 2, pp. 193–212, 2012.

[37] J. S. Reuter et D. H. Mathews, « RNAstructure : software for RNA secondary structure prediction and analysis, » *BMC bioinformatics*, vol. 11, no. 1, pp. 1–9, 2010.

[38] R. G. Roeder, « The role of general initiation factors in transcription by RNA polymerase II, » *Trends in biochemical sciences*, vol. 21, no. 9, pp. 327–335, 1996.

[39] G. Romano, D. Veneziano, M. Acunzo, et C. M. Croce, « Small noncoding RNA and cancer, » *Carcinogenesis*, vol. 38, no. 5, pp. 485–491, 2017.

[40] D. Sankoff, « Simultaneous solution of the RNA folding, alignment and protosequence problems, » *SIAM journal on applied mathematics*, vol. 45, no. 5, pp. 810–825, 1985.

[41] M. C. Siomi, K. Sato, D. Pezic, et A. A. Aravin, « PIWI-interacting small RNAs : the vanguard of genome defence, » *Nature reviews Molecular cell biology*, vol. 12, no. 4, pp. 246–258, 2011.

[42] Z. Tan, Y. Fu, G. Sharma, et D. H. Mathews, « TurboFold II : RNA structural alignment and secondary structure prediction informed by multiple homologs, » *Nucleic acids research*, vol. 45, no. 20, pp. 11 570–11 581, 2017.

[43] E. Torarinsson, J. H. Havgaard, et J. Gorodkin, « Multiple structural alignment and clustering of RNA sequences, » *Bioinformatics*, vol. 23, no. 8, pp. 926–932, 2007.

[44] E. Trotta, « On the normalization of the minimum free energy of RNAs by sequence length, » *PloS one*, vol. 9, no. 11, p. e113380, 2014.

[45] N. J. Watkins et M. T. Bohnsack, « The box C/D and H/ACA snoRNPs : key players in the modification, processing and the dynamic folding of ribosomal RNA, » *Wiley Interdisciplinary Reviews : RNA*, vol. 3, no. 3, pp. 397–414, 2012.

[46] J. D. Watson et F. H. Crick, « Molecular structure of nucleic acids : a structure for deoxyribose nucleic acid, » *Nature*, vol. 171, no. 4356, pp. 737–738, 1953.

[47] K. C. Wang et H. Y. Chang, « Molecular mechanisms of long noncoding RNAs, » *Molecular cell*, vol. 43, no. 6, pp. 904–914, 2011.

[48] C. R. Woese et G. E. Fox, « Phylogenetic structure of the prokaryotic domain : the primary kingdoms, » *Proceedings of the National Academy of Sciences*, vol. 74, no. 11, pp. 5088–5090, 1977.

[49] A. Wutz et J. Gribnau, « X inactivation Xplained, » *Current opinion in genetics & development*, vol. 17, no. 5, pp. 387–393, 2007.

[50] S. Zakov, Y. Goldberg, M. Elhadad, et M. Ziv-Ukelson, « Rich parameterization improves RNA structure prediction, » *Journal of Computational Biology*, vol. 18, no. 11, pp. 1525–1542, 2011.

[51] Y. Zhang, H. Huang, D. Zhang, J. Qiu, J. Yang, K. Wang, L. Zhu, J. Fan, et J. Yang, « A review on recent computational methods for predicting noncoding RNAs, » *BioMed research international*, vol. 2017, 2017.

[52] M. Zuker, J. A. Jaeger, et D. H. Turner, « A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison, » *Nucleic acids research*, vol. 19, no. 10, pp. 2707–2714, 1991.

[53] M. Zuker et P. Stiegler, « Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information, » *Nucleic acids research*, vol. 9, no. 1, pp. 133–148, 1981.