Université de Montréal

Représentation et recherche de motifs cycliques et structuraux d'ARN connus dans les structures secondaires

par
Caroline Louis-Jeune

Département de biochimie
Faculté de médecine

Mémoire présenté à la Faculté des études supérieures et post-doctorales
en vue de l'obtention du grade de maîtrise
en bio-informatique

avril, 2009

Université de Montréal
Faculté des études supérieures et post-doctorales

Ce mémoire intitulé :

Représentation et recherche de motifs cycliques et structuraux d'ARN connus dans les structures secondaires

présenté par :
Caroline Louis-Jeune

a été évaluée par un jury composé des personnes suivantes :

Pascal Chartrand, président-rapporteur
François Major, directeur de recherche
Sylvie Hamel, membre du jury

Résumé

L'acide désoxyribonucléique (ADN) et l'acide ribonucléique (ARN) sont des polymères de nucléotides essentiels à la cellule. À l'inverse de l'ADN qui sert principalement à stocker l'information génétique, les ARN sont impliqués dans plusieurs processus métaboliques. Par exemple, ils transmettent l'information génétique codée dans l'ADN. Ils sont essentiels pour la maturation des autres ARN, la régulation de l'expression génétique, la prévention de la dégradation des chromosomes et le ciblage des protéines dans la cellule. La polyvalence fonctionnelle de l'ARN résulte de sa plus grande diversité structurale.

Notre laboratoire a développé *MC-Fold*, un algorithme pour prédire la structure des ARN qu'on représente avec des graphes d'interactions inter-nucléotidiques. Les sommets de ces graphes représentent les nucléotides et les arêtes leurs interactions. Notre laboratoire a aussi observé qu'un petit ensemble de cycles d'interactions à lui seul définit la structure de n'importe quel motif d'ARN. La formation de ces cycles dépend de la séquence de nucléotides et *MC-Fold* détermine les cycles les plus probables étant donnée cette séquence.

Mon projet de maîtrise a été, dans un premier temps, de définir une base de données des motifs structuraux et fonctionnels d'ARN, *bdMotifs*, en terme de ces cycles. Par la suite, j'ai implanté un algorithme, *MC-Motifs*, qui recherche ces motifs dans des graphes d'interactions et, entre autres, ceux générés par *MC-Fold*. Finalement, j'ai validé mon algorithme sur des ARN dont la structure est connue, tels que les ARN ribosomaux (ARNr) 5S, 16S et 23S, et l'ARN utilisé pour prédire la structure des riborégulateurs.

Le mémoire est divisé en cinq chapitres. Le premier chapitre présente la structure chimique, les fonctions cellulaires de l'ARN et le repliement structural du polymère. Dans le deuxième chapitre, je décris la base de données *bdMotifs*. Dans le troisième chapitre, l'algorithme de recherche *MC-Motifs* est introduit. Le quatrième chapitre présente les

résultats de la validation et des prédictions. Finalement, le dernier chapitre porte sur la discussion des résultats suivis d'une conclusion sur le travail.

Mots-clés : ARN, structure secondaire, motif, cycle

Abstract

Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are polymers of nucleotides essential for the survival of the cell. Contrary to DNA, whose main role is to store genetic information, RNA is involved in multiple metabolic processes. For example, RNA is involved in the transfer of information from DNA to protein, the processing and modification of other RNAs, the regulation of gene expression, the end-maintenance of chromosomes, and the sorting of proteins within the cell. This functional versatility of RNA comes from its structural diversity.

Our laboratory developed *MC-Fold*, an algorithm that predicts RNA structures by representing them with nucleotide interaction graphs. The nodes in these graphs represent the nucleotides, and the edges the interactions between them. Our laboratory also observed that a limited number of interaction cycles can define the structure of any RNA motif. The formation of these cycles is determined by the nucleotide sequence and *MC-Fold* determines the most likely cycles based on that sequence.

In this Master Degree project, I first built a database of structural and functional RNA motifs, *bdMotifs*, based on their constituent cycles. Then, I implemented an algorithm, *MC-Motifs*, which detects motifs within interaction graphs generated either by *MC-Fold* or by any other method. Finally, I validated my algorithm on known RNA structures such as the 5S, 16S and 23S ribosomal RNA (rRNA) and predicted structure of riboswitches.

The Master thesis is divided into five chapters. The first chapter presents the chemical structure of RNA, its cellular functions and the structural folding of the polymer. In the second chapter, the database *bdMotifs* is described. In the third chapter, the *MC-Motifs* algorithm is introduced. In the fourth chapter, I present the results of *MC-Motifs*. Finally, in the last chapter, I discuss theses results and I give a conclusion on the project.

Keywords : RNA, secondary structure, motif, cycle

Table des matières

RÉSUMÉ . . . . . **iii**

ABSTRACT . . . . . **v**

TABLE DES MATIÈRES . . . . . **vii**

LISTE DES TABLEAUX . . . . . **xi**

LISTE DES FIGURES . . . . . **xii**

DÉDICACE . . . . . **xv**

REMERCIEMENTS . . . . . **xvi**

CHAPITRE 1 : INTRODUCTION **1**

1. Les acides ribonucléiques (ARN), leurs fonctions et leurs structures **1**

1.1 Les acides ribonucléiques 1

1.2 Les fonctions cellulaires des ARN 3

1.2.1 Synthèse des protéines 3

1.2.2 Régulation de l'expression génétique 5

1.2.3 Conclusion 6

1.3 Structure de l'ARN 7

1.4 Niveaux d'organisation structurale 7

1.4.1 Structure primaire 8

1.4.2 Structure secondaire 8

1.4.3 Structure tertiaire 9

1.4.4 Conclusion 10

1.5 Paires de bases non classiques de l'ARN 11

1.5.1 Historique 11

1.5.2 Introduction de la nomenclature Leontis-Westhof 12
1.5.2.1 Les faces d'une base 12
1.5.2.2 L'orientation des liens glycosidiques 13
1.5.3 Utilité de la nomenclature Leontis-Westhof 14
1.6 Les graphes d'interactions 15
1.7 Les motifs cycliques de l'ARN 18
1.8 But du mémoire 20

CHAPITRE 2 : BASE DE DONNÉES DE MOTIFS DE L'ARN 22

2.1 La tétraboucle GNRA 22
2.1.1 Génération du graphe représentant la tétraboucle GNRA 23
2.1.2 Résultats de recherche de la tétraboucle GNRA par *MC-Search* . . . 24
2.2 La boucle T 25
2.2.1 Génération du graphe représentant la boucle T 26
2.2.2 Résultats de recherche de la boucle T par *MC-Search* 26
2.3 Le motif sarcine-ricine 27
2.3.1 Génération du graphe représentant le sarcin-ricin 28
2.3.2 Résultats de recherche de sarcin-ricin par *MC-Search* 29
2.4 Le kink-turn 30
2.4.1 Génération du graphe représentant le kink-turn 31
2.4.2 Résultats de recherche de kink-turn par *MC-Search* 32
2.5 Le motif C 33
2.5.1 Génération du graphe représentant le motif C 35
2.5.2 Résultats de recherche de motif C par *MC-Search* 37
2.6 Le motif « UA_handle » 38

2.6.1 Génération du graphe représentant le « UA_handle » 39
2.6.2 Résultats de recherche de « UA_handle » par *MC-Search* 41
2.7 Conclusion 43

CHAPITRE 3 : RECHERCHE DE MOTIFS STRUCTURAUX ET FONCTIONNELS DANS UNE STRUCTURE SECONDAIRE . . 46

3.1 Prétraitement de la chaîne de points et parenthèses 47
3.2 Algorithme de recherche de motifs structuraux et fonctionnels 48
3.3 Les pseudonœuds 53
3.4 Conclusion 56

CHAPITRE 4 : VALIDATION ET UTILISATION DE *MC-Motifs* 57

4.1 Validation de *MC-Motifs* 57
4.2 Prédiction de motifs dans une famille d'ARN 59
4.2.1 Riborégulateur FMN de *Vibrio cholerae* 61
4.2.2 Riborégulateur FMN de *Mesorhizobium loti* 65

CHAPITRE 5 : DISCUSSION ET CONCLUSION 72

5.1 Utilité des motifs cycliques 72
5.2 Avantages de *MC-Motifs* 73
5.3 Inconvénients de *MC-Motifs* 77
5.4 Conclusion 78

BIBLIOGRAPHIE 80

ANNEXE A : ALGORITHME DE *pdb2NCM.pl* I

ANNEXE B : LES MODULES DE *MC-Motifs* XV

ANNEXE C : PSEUDOCODE POUR CONVERTIR UNE CHAÎNE DE POINTS

ET PARENTHÈSES EN CHAÎNE DE MOTIFS CYCLIQUES . . . XVII

ANNEXE D : LA BASE DE DONNÉES *bdMotifs* XIX

Liste des tableaux

Table 1.1	**Douze familles d'appariement distinctes**	**14**
Table 2.1	**Exemples de séquences des motifs GNRA détectés par *MC-Search***	**24**
Table 2.2	**Séquences des boucles T détectées par *MC-Search***	**26**
Table 2.3	**Séquences des motifs sarcine-ricine détectés par *MC-Search***	**29**
Table 2.4	**Séquences des motifs kink-turn détectés par *MC-Search***	**33**
Table 2.5	**Séquence du motif C de type 5x2 trouvé par *MC-Search***	**37**
Table 2.6	**Séquences du motifs C de type 5x3 trouvés par *MC-Search***	**37**
Table 2.7	**Séquence du motif C de type 6x4 trouvé par *MC-Search***	**38**
Table 2.8a	**Exemples de séquences des motifs « UA_handle » de type I trouvés par *MC-Search***	**42**
Table 2.8b	**Exemples de séquences des motifs « UA_handle » de type II trouvés par *MC-Search***	**42**
Table 2.8c	**Exemples de séquences des motifs « UA_handle » de type III trouvés par *MC-Search***	**43**
Table 2.9	**Exemple du contenu de la base de données *bdMotifs***	**43**

Liste des figures

Figure 1.1	**Composants moléculaires des quatre nucléotides de l'ARN**	**1**
Figure 1.2	**Structure chimique de la séquence d'ARN AUGC**	**2**
Figure 1.3	**Schéma de la synthèse des protéines dans une cellule procaryote**	**4**
Figure 1.4	**Différences structurales entre une double hélice et un ARN de transfert**	**7**
Figure 1.5	**Éléments structuraux secondaires et classiques de l'ARN 16S de *Thermus thermophilus* (1J5E.pdb)**	**9**
Figure 1.6	**Exemple d'éléments structuraux tertiaires**	**10**
Figure 1.7	**Repliement hiérarchique de l'ARNr 5S *Haloarcula marismortui***	**11**
Figure 1.8	**Appariements possibles d'une base d'ARN**	**13**
Figure 1.9	**Orientation des liens glycosidiques**	**13**
Figure 1.10	**Graphe d'interactions du 5S de l'ARNr *Haloarcula marismortui***	**15**
Figure 1.11	**Graphe structurel agrandi de 5'-GAGUA-GAAA-3'**	**16**
Figure 1.12	**Descripteur pour une tige-boucle 5'-ACUGU-3'**	**17**
Figure 1.13	**L'alphabet structural**	**19**
Figure 1.14	**Structure secondaire de la boucle E prédite par *MC-Fold***	**20**
Figure 2.1	**Exemple de la tétraboucle GNRA dans l'ARNr 23S d'*Haloarcula marismortui***	**23**
Figure 2.2	**Descripteur de la tétraboucle GNRA**	**24**
Figure 2.3	**La boucle T**	**25**
Figure 2.4	**Graphe d'interactions et descripteur de la boucle T**	**26**
Figure 2.5	**Motif structural sarcine-ricine de l'ARNr 23S d'*Haloarcula marismortui***	**28**
Figure 2.6	**Descripteur du motif sarcine-ricine**	**29**
Figure 2.7	**Motif structural de kink-turn de l'ARNr 23S d'*Haloarcula marismortui***	**31**
Figure 2.8	**Descripteurs pour localiser des kink-turn**	**32**
Figure 2.9	**Exemples des variants du motif C**	**35**

Figure 2.10	**Descripteurs de motif C**	**36**
Figure 2.11	**Exemples de motif « UA_handle »**	**39**
Figure 2.12a	**Graphes d'interactions et descripteurs du motif « UA_handle » de type I**	**40**
Figure 2.12b	**Graphes d'interactions et descripteur du motif « UA_handle » de type II**	**40**
Figure 2.12c	**Graphes d'interactions et descripteurs du motif « UA_handle » de type III**	**41**
Figure 3.1	**Interface de saisie de données de *MC-Motifs***	**47**
Figure 3.2	**Prétraitement de la chaîne de points et parenthèses**	**48**
Figure 3.3	**Traitement de la chaîne de motifs cycliques par *MC-Motifs***	**49**
Figure 3.4	**Algorithme de recherche de sarcine-ricine par *MC-Motifs***	**50**
Figure 3.5	**Algorithme de recherche de la tétraboucle GNRA *MC-Motifs***	**52**
Figure 3.6	**Motifs sarcine-ricine et GNRA trouvés dans la tige-boucle II par *MC-Motifs***	**53**
Figure 3.7	**Domaine « T-loop PK »**	**54**
Figure 4.1	**Motifs trouvés dans le domaine I de l'ARNr 16S de *Thermus thermophilus* par *MC-Motifs***	**58**
Figure 4.2	**Motifs trouvés dans l'ARNr 5S d'*Haloarcula marismortui* par *MC-Motifs***	**59**
Figure 4.3	**Prédiction du noyau structural du riborégulateur FMN**	**60**
Figure 4.4	**Motifs structuraux détectés par *MC-Motifs* dans l'ARN du riborégulateur FMN de *Vibrio cholerae***	**61**
Figure 4.5	**Régions ayant peu d'appariements dans l'ARN de *Vibrio cholerae***	**62**
Figure 4.6	**Structures secondaires des régions P3, P4 et P6 de *Vibrio cholerae* générées par *MC-Fold***	**62**
Figure 4.7	**Motifs structuraux détectés par *MC-Motifs* dans les régions P3, P4 et P6 de *Vibrio cholerae***	**63**
Figure 4.8	**Structure secondaire du noyau structural de l'ARN FMN de *Vibrio cholerae***	**65**
Figure 4.9	**Motifs structuraux détectés par *MC-Motifs* dans l'ARN du**	**66**

riborégulateur FMN de *Mesorhizobium loti*

Figure 4.10 **Régions ayant peu d'appariements dans l'ARN de *Mesorhizobium loti*** **66**

Figure 4.11 **Structures secondaires des régions P3, P4 et P6 de *Mesorhizobium loti* générées par *MC-Fold*** **67**

Figure 4.12 **Motifs structuraux détectés par *MC-Motifs* dans les régions P3, P4 et P6 de *Mesorhizobium loti*** **68**

Figure 4.13 **Structure secondaire du noyau structural de l'ARN FMN de *Mesorhizobium loti*** **70**

Figure 5.1 **Interactions tertiaires prédites dans la structure secondaire du noyau structural de l'ARN FMN de *Mesorhizobium loti*** **74**

Figure 5.2 **Interactions tertiaires prédites dans la structure secondaire du noyau structural de l'ARN FMN de *Vibrio cholerae*** **76**

Figure 5.3 **Motif « lonepair triloop » dans l'hélice six du domaine I de l'ARNr 16S *Thermus thermophilus*** **77**

Ce mémoire est dédicacé à mes parents, mon frère et mes amies (Patricia, Johanne, Vanessa et Lynda) qui m'ont encouragé durant mes études de maîtrise.

Remerciements

Tout d'abord, je veux remercier mon directeur de recherche, Dr. François Major, pour m'avoir introduit au domaine passionnant de la bioinformatique, pour ses conseils sur ce sujet de recherche et son soutien tout au long de mes études de maîtrise. Je remercie également les membres du laboratoire : Marc Parisien, Ali Mokdad, Véronique Lisi et Karine Saint-Onge pour leurs nombreux conseils. J'ai eu le plaisir de travailler avec vous. Finalement, je tiens à remercier Patrick Gendron, membre de l'équipe bioinformatique de l'Institut de recherche en immunologie et cancérologie (IRIC), pour son expertise technique.

Je remercie chaleureusement le Dr. Pascal Chartrand et le Dr. Sylvie Hamel d'avoir accepté d'être membre du jury d'évaluation de ce mémoire.

CHAPITRE 1

INTRODUCTION

LES ACIDES RIBONUCLÉIQUES (ARN), LEURS FONCTIONS ET LEURS STRUCTURES

1.1 Les acides ribonucléiques

Les acides ribonucléiques (ARN) sont des polymères formés par l'assemblage de quatre différents types de monomères nommés nucléotides (Saenger, 1984). Un nucléotide consiste en un groupement phosphate lié à un ribose auquel est rattaché une des quatre bases azotées: les pyrimidines uracile (U) et cytosine (C), et les purines adénine (A) et guanine (G) (*Figure 1.1*).

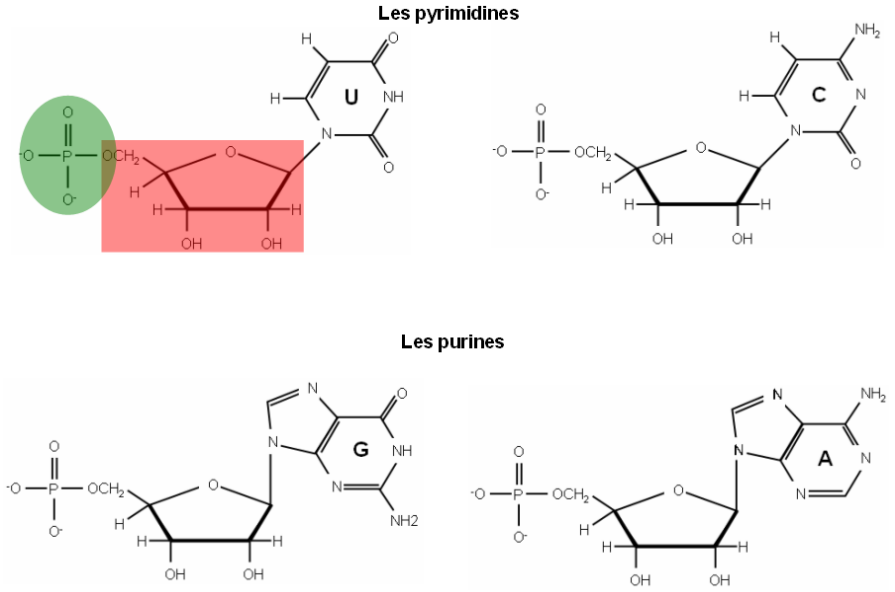


Figure 1.1: Composants moléculaires des quatre nucléotides de l'ARN. Chaque nucléotide est composé d'un groupement phosphate (vert), d'un ribose (rouge) et d'une base azotée.

Les nucléotides sont liés entre eux dans une chaîne polynucléotidique (séquence) par des liens phosphodiesters entre le groupement 3'-hydroxyle du ribose d'un nucléotide et le groupement phosphate rattaché au groupement 5'-hydroxyle d'un autre nucléotide (*Figure 1.2*). Ces liens phosphodiesters forment un squelette où les riboses et les groupements phosphate sont alternés et les bases sont unies au squelette. L'ARN possède un groupement phosphate libre à l'extrémité 5' et un groupement hydroxyle à l'extrémité 3'. La convention d'écriture des séquences se fait du 5' vers le 3'.

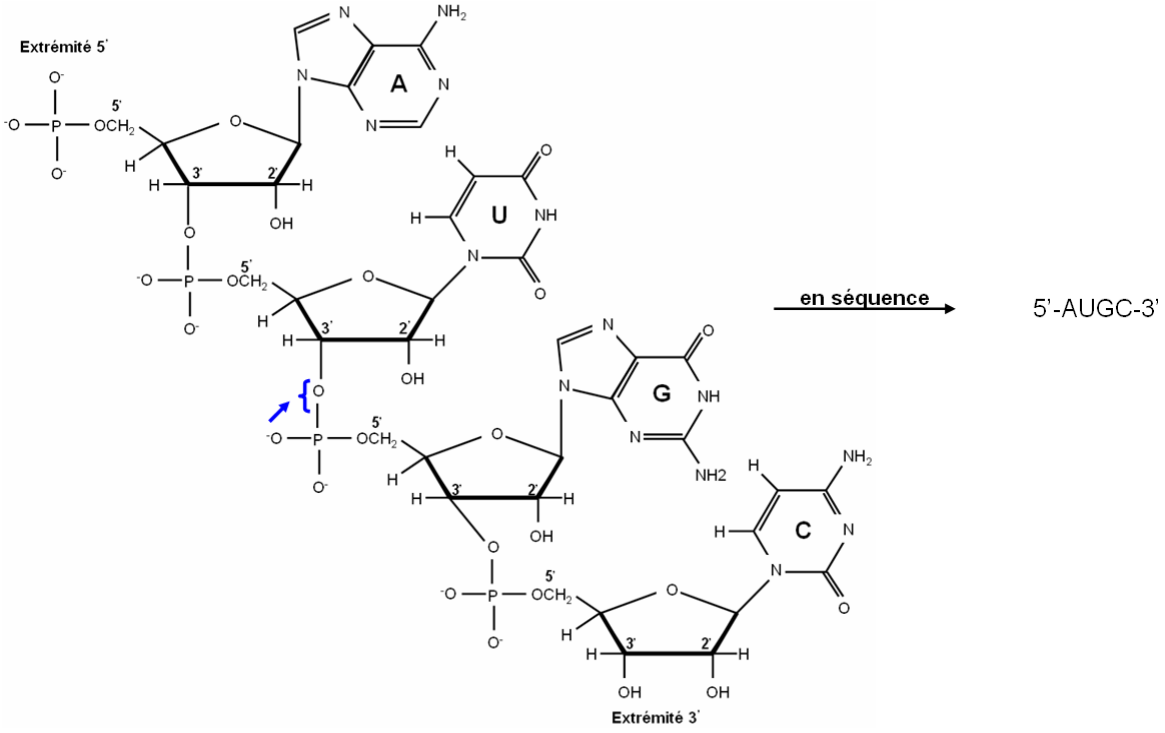


Figure 1.2: Structure chimique de la séquence d'ARN AUGC. Les nucléotides sont liés entre eux par des liens phosphodiesters indiqués ici par une flèche. Les numéros correspondent aux positions des atomes du ribose. La direction de la séquence se fait dans le sens 5' → 3'.

1.2 Les fonctions cellulaires des ARN

Les molécules d'ARN participent dans plusieurs processus métaboliques essentiels pour la cellule. Dans cette section, les ARN impliqués dans la synthèse des protéines et dans la régulation de l'expression génétique sont présentés, car ce sont les principaux ARN utilisés dans mon projet de recherche.

1.2.1 Synthèse des protéines

Les ARN permettent que l'information génétique provenant de l'ADN soit exprimée sous forme de protéines. Le transfert d'information entre ADN et protéines se fait en deux étapes: la transcription et la traduction (Crick, 1970; Lewin, 1997).

Durant la transcription, l'information stockée dans l'ADN est copiée sous la forme d'ARN messager (ARNm). On l'appelle « messager » car il porte l'information génétique de l'ADN vers les ribosomes, un organite de la cellule. Dans les ribosomes, l'ARNm est utilisé comme guide pour fabriquer les protéines au cours du deuxième processus qu'est la traduction. Lors de la traduction, un autre type d'ARN, l'ARN de transfert (ARNt), interprète l'information de l'ARNm en reconnaissant les codons (triplets de nucléotides). Il transporte l'acide aminé, une unité structurale de base des protéines, correspondant au codon vers les ribosomes.

Les ribosomes sont des complexes ribonucléoprotéiques constitués de protéines et d'ARN : les ARN ribosomaux (ARNr). Ils sont formés d'une grande et d'une petite sous-unités désignées 50S et 30S respectivement chez les procaryotes (ex : bactéries), 60S et 40S respectivement chez les eucaryotes (ex : animaux, levures ou plantes) (Marintchev & Wagner, 2004). Chaque sous-unité ribosomale contient des molécules d'ARNr de taille variable. Chez les bactéries, la sous-unité 30S contient une seule molécule d'ARNr 16S. La grande sous-unité 50S contient deux molécules d'ARNr : ARNr 23S et ARNr 5S. Chez

les eucaryotes, la petite sous-unité 40S contient une molécule d'ARNr 18S. La grande sous-unité 60S contient trois molécules d'ARNr : ARNr 5.8S, ARNr 28S et ARNr 5S.

Les deux sous-unités ribosomales effectuent différentes tâches au cours de la traduction. Par exemple, chez les procaryotes, les ARNr 16S des sous-unités 30S aident à l'initiation de la synthèse des protéines en facilitant la fixation des ARNm et des ARNt sur le ribosome; alors que les ARNr 23S des sous-unité 50S assemblent et relient les uns aux autres les acides aminés qui forment les protéines (*Figure 1.3*) (Noller et al., 1992; Cech, 2000; Puglisi et al., 2000; Marintchev & Wagner, 2004; Noller et al., 2005).

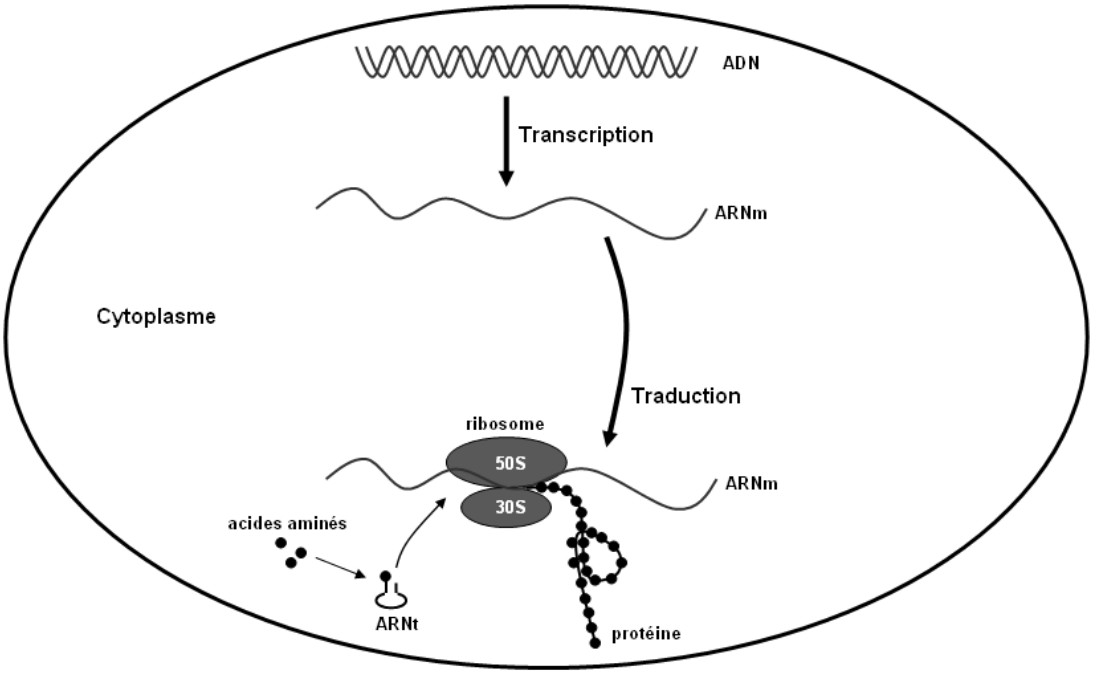


Figure 1.3 : Schéma de la synthèse des protéines dans une cellule procaryote. Durant la transcription, l'information génétique de l'ADN est codée en ARNm. Lors de la traduction, les acides aminés sont assemblés selon l'information provenant de l'ARNm créant la protéine. Remarquez que le ribosome d'une cellule procaryote est constitué d'une grande (50S) et d'une petite (30S) sous-unités.

Image adaptée de Korf et al., 2003

Les ARNt et les ARNr sont des ARN non-codants, car ils ne sont jamais traduits en protéine. Ils remplissent leur fonction en tant qu'ARN seulement.

1.2.2 Régulation de l'expression génétique

Il existe des ARN non-codants qui agissent en tant que régulateurs d'expression génétique.

Les riborégulateurs en sont des exemples. Ce sont des éléments de contrôle génétique localisés principalement dans la région 5' non-traduite (5'-UTR) de certains ARNm (Winkler et al., 2002a; Winkler & Breaker, 2003). Ils agissent comme des senseurs de métabolites cellulaires. La liaison entre un métabolite et un riborégulateur engendre un changement de conformation structurale de l'ARNm causant une interruption prématurée de la transcription ou une inhibition de la traduction de cet ARN (Winkler & Breaker, 2003; Vitreschak et al., 2004).

Les riborégulateurs modulent plusieurs voies métaboliques impliquées dans la biosynthèse des vitamines (Nahvi et al., 2002; Winkler et al., 2002b), des acides aminés (Epshtein et al., 2003; Vitreschak et al., 2004) et des purines (Mandal et al., 2003; Vitreschak et al., 2004). Par exemple, le riborégulateur de la flavine mononucléotide (FMN) contrôle l'expression génétique de la FMN, molécule dérivée de la vitamine B12 (Winkler et al., 2002b; Wickiser et al., 2005). Lorsque la concentration de la FMN est basse, le riborégulateur FMN forment une structure spécifique permettant la transcription de l'ARNm codant la FMN. Si la FMN est en concentration élevée, un complexe est formé entre le métabolite et le riborégulateur. Cet appariement entraîne une réorganisation structurale de l'ARNm et provoque une interruption prématurée de la transcription de l'ARNm, donc l'inhibition de l'expression de la FMN.

1.2.3 Conclusion

L'élucidation des fonctions biologiques des ARNr et des riborégulateurs a été possible grâce aux analyses biochimiques, biophysiques et cristallographiques (Ban et al., 2000; Wimberly et al., 2000; Marintchev & Wagner, 2004; Noller, 2005b; Edwards et al., 2007). Les expériences de cristallographie sur ces ARN, en particulier, ont permis de comprendre le repliement structural de l'ARN et les interactions entre les structures de l'ARN et des biomolécules dans la cellule. Les sections 1.3 et 1.4 de ce chapitre porte sur la description de la structure de l'ARN.

1.3 Structure de l'ARN

L'ADN double brin adopte dans l'espace une structure hélicoïdale, alors que l'ARN formé d'un seul brin adopte diverses structures pour accomplir ses différentes fonctions biologiques (*Figure 1.4*). Le brin d'ARN se replie sur lui-même pour former des segments double brin et des structures plus complexes telles que montrées à la *Figure 1.4b* (Saenger, 1984).

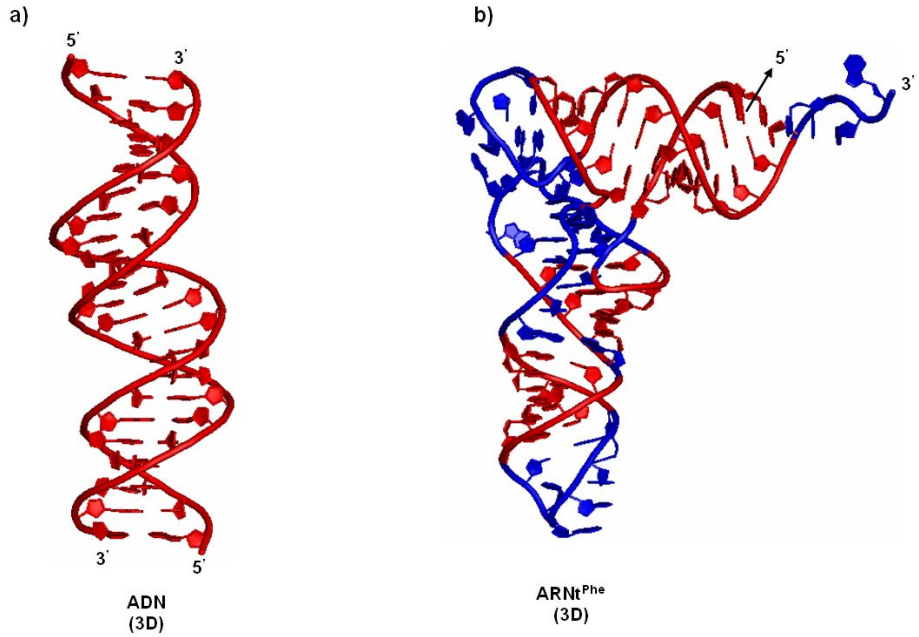


Figure 1.4 : Différences structurales entre une double hélice et un ARN de transfert. Les structures représentées sont obtenues par cristallographie à rayons-X de l'ADN clivé par une enzyme (2JYK.pdb) et de l'ARNt de la phénylalanine (Phe) de la levure *Saccharomyces cerevisiae* (1EVV.pdb). **(a)** L'ADN possède une structure hélicoïdale (rouge), un double brin dont l'un part de 5' → 3' et l'autre de 3' → 5', dans l'espace tridimensionnelle (3D). **(b)** L'ARN est formé d'un seul brin, 5' → 3', qui se replie sur lui-même pour former des molécules contenant des segments hélicoïdaux (rouge) comme la double hélice ADN. Les segments sont rattachés par les simples brins d'ARN (bleu).

1.4 Niveaux d'organisation structurale

La structure de l'ARN peut être divisée en trois niveaux fondamentaux d'organisation structurale : primaire, secondaire et tertiaire (Batey et al., 1999; Auffinger & Westhof, 2000).

1.4.1 Structure primaire

La structure primaire réfère à la séquence de nucléotides. Dans la cellule, elle provient de la transcription d'un des deux brins de l'ADN.

1.4.2 Structure secondaire

En présence des cofacteurs et en particulier des cations[1] en solution, les brins d'ARN se replient pour former différents éléments structuraux récurrents (*Figure 1.5*). Il y a les doubles hélices (*Figure 1.5a*) stabilisées par des ponts hydrogènes formés entre les bases complémentaires (A avec U et G avec C ou U). Les paires de bases A-U et G-C forment des appariements dits Watson-Crick, tandis que les paires de bases G-U sont dites « wobble » (Hermann & Westhof, 1999). Les paires A-U, G-C et G-U sont majoritaires. Les appariements entraînent la formation d'éléments structuraux non hélicoïdaux. Par exemple, il y a la tige-boucle (*Figure 1.5b*). Elle est formée lorsque le brin d'ARN se replie sur lui-même pour former une boucle en U. La boucle interne (*Figure 1.5c*) se forme lorsqu'il y a au moins une base non appariée sur chaque brin, séparant ainsi les hélices et formant une boucle. Il y a le bourgeon (*Figure 1.5d*) qui contient des bases non appariées sur seulement un brin. L'autre brin contient des appariements continus. Finalement, il y a la jonction d'hélices (*Figure 1.5e*) qui est créée lorsque les hélices distinctes se joignent ensemble (Auffinger & Westhof, 2000).

L'empilement stérique entre les bases dans la structure secondaire contribue à la stabilité de la structure (Petersheim & Turner, 1983).

---

[1] Des ions chargés positivement

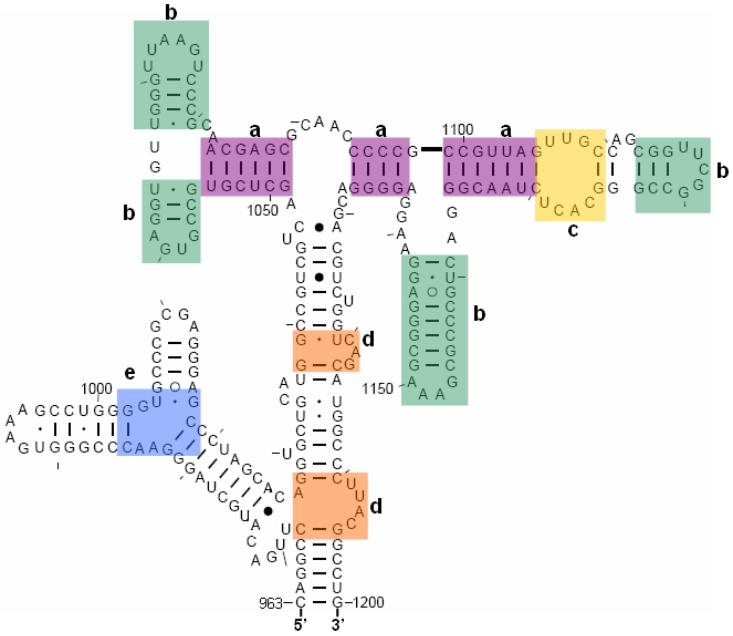


Figure 1.5: Éléments structuraux secondaires et classiques de l'ARN 16S de *Thermus thermophilus* (1J5E.pdb). Les points sont les appariements G-U dites « wobble » et les traits sont les appariements Watson-Crick (A-U et G-C). Les cercles noirs ou vides représentent des appariements non classiques (voir la section 1.5). Les éléments structuraux sont des sous-structures répétées. Certains de ces éléments sont **(a)** des hélices**, (b)** des tiges-boucles, **(c)** des boucles internes, **(d)** des bourgeons et **(e)** des jonction d'hélices.

1.4.3 Structure tertiaire

La structure tertiaire est définie par les coordonnées atomiques dans l'espace 3D. À ce niveau, les éléments structuraux secondaires peuvent s'associer pour former des interactions dites tertiaires. La présence d'interactions tertiaires mène à la formation d'éléments structuraux tertiaires (Leontis & Westhof, 2003).

Par exemple, il y a les pseudonœuds (*Figure 1.6a*). Les pseudonœuds sont formés à partir des interactions tertiaires entre deux tiges-boucles superposées. Il existe des liaisons entre deux tiges-boucles qui sont stabilisées par des interactions tertiaires (*Figure 1.6b*). Le même phénomène peut se produire entre les nucléotides d'un bourgeon et ceux d'une tige-boucle (*Figure 1.6c*).

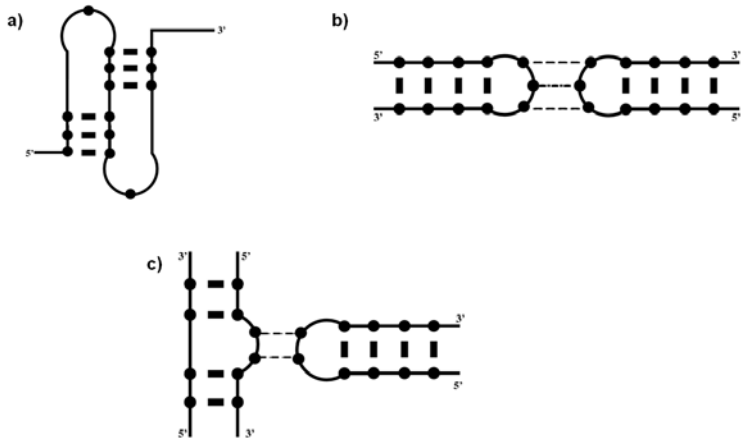


Figure 1.6: Exemples d'éléments structuraux tertiaires. Les points représentent les nucléotides, les traits sont des appariements et les lignes pointillées représentent les interactions tertiaires.
(a) Le pseudonoeud. **(b)** Liaisons entre deux tiges-boucles **(c)** Liaison entre un bourgeon et une tige-boucle.

La structure tertiaire confère à l'ARN sa fonction biologique.

1.4.4 Conclusion

La *Figure 1.7* montre les trois niveaux d'organisation structurale d'une molécule d'ARN : les structures primaire, secondaire et tertiaire. Le repliement suit cette hiérarchie. La séquence forme rapidement des éléments de structure secondaire. À plus long terme, les éléments secondaires interagissent et forment la structure tertiaire.

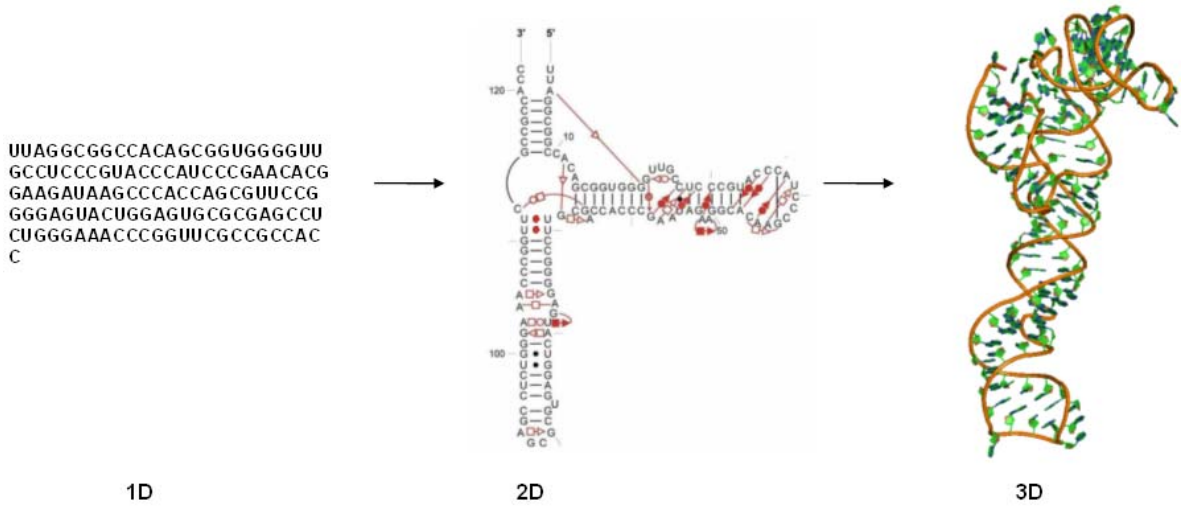


Figure 1.7. Repliement hiérarchique de l'ARNr 5S *Haloarcula marismortui*. La structure de l'ARN est divisée en trois niveaux d'organisation structurale. **(1D)** La structure primaire de l'ARN est la séquence de nucléotides de la molécule. **(2D)** La structure secondaire où la structure primaire de l'ARN se replie sur elle-même pour former des paires de bases**.** Les traits et les cercles noirs sont des appariements classiques et les liaisons rouges sont des appariements non classiques. Les symboles des appariements non classiques proviennent de la nomenclature Leontis-Westhof (voir la section 1.5.2). **(3D)** Les molécules structurées d'ARN transitent vers un arrangement en 3D pour former la structure tertiaire. La structure tertiaire représentée ici est celle obtenue par cristallographie à rayons-X de l'ARNr *5S Haloarcula marismortui* (1JJ2.pdb).

1.5 Paires de bases non classiques de l'ARN

1.5.1 Historique

À part les paires de bases G-U, on a remarqué la présence de différentes paires de bases non classiques dans les structures tertiaires de l'ARNt (Kim et al., 1974; Quigley & Rich, 1976), dans les ARNr 16S, 23S et 28S (Gutell et al., 1994; Gautheret et al., 1995; Gautheret & Gutell, 1997), dans les ribozymes[2] (Pley et al., 1994) et dans les introns de

[2] ARN possédant une activité catalytique tel que l'ARNr du ribosome.

groupe I[3] (Michel & Westhof, 1990; Cate et al., 1996). Des liaisons telles que G-A, A-C, A-A et U-U ont été souvent observées dans ces structures.

L'analyse des structures tertiaires d'ARN de plus grande taille, telles que celles des sous-unités ribosomales ARNr 30S et 50S (Ban et al., 2000; Carter et al., 2000; Gutell et al., 2002), révèle que mise à part les paires de bases Watson-Crick, les paires de bases non Watson-Crick dominent la structure tertiaire. On a observé qu'elles maintiennent la cohésion de l'architecture des ARN et elles permettent à la molécule d'ARN d'interagir avec des protéines ou d'autres molécules d'ARN (Leontis & Westhof, 2001; Lescoute et al., 2005). On a conclu que les appariements Watson-Crick ne constituent qu'une possibilité parmi plusieurs types d'appariement possibles entre les quatre bases (Westhof & Fritsch, 2000; Leontis & Westhof, 2001).

1.5.2 Introduction de la nomenclature Leontis-Westhof

Une nomenclature descriptive de toutes les interactions de bases observées dans les molécules d'ARN a été proposée par Leontis et Westhof (Leontis & Westhof 2001). Cette nomenclature permet d'identifier les paires de base qui ne forment nécessairement pas des appariements Watson-Crick et elle permet de décrire la structure tertiaire par un ensemble d'interactions spécifiques (Westhof & Fritsch, 2000; Leontis & Westhof, 2001). Ces paires de bases sont classées selon les faces de la base favorisant des ponts hydrogènes et selon l'orientation des liens glycosidiques (liens qui relient les riboses aux bases).

1.5.2.1 Les faces d'une base

Une base purine ou pyrimidine a trois faces : une face Watson-Crick (W), une face Hoogsteen (H) et une face Sucre (S) (*Figure 1.8*). Une face donnée d'une base peut

[3] Ribozymes catalysant eux-mêmes leur épissage (excision des parties non codantes [introns] de l'ARNm).

s'apparier avec l'une des trois faces de la seconde base. Alors, les bases peuvent créer six types d'appariements.

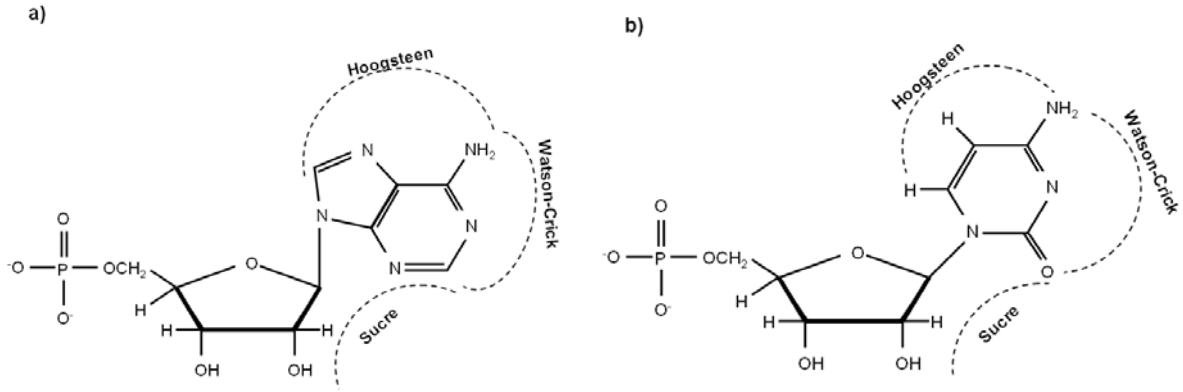


Figure 1.8. Appariements possibles d'une base d'ARN. Les bases purine **(a)** et pyrimidine **(b)** ont trois faces pour former des ponts hydrogènes: la face Watson-Crick ou W, la face Hoogsteen ou H et la face Sucre ou S (comprenant le groupe hydroxyle 2' du sucre). Une face donnée d'une base peut s'apparier avec l'une des trois faces de la seconde base.

1.5.2.2 L'orientation des liens glycosidiques

Les liaisons entre les bases peuvent se faire dans l'orientation *cis* ou *trans* selon les positions des liens glycosidiques : les liens sont dans l'orientation *cis*, s'ils se trouvent du même côté que le plan des ponts hydrogènes reliant les bases. Ils sont dans l'orientation *trans*, s'ils sont de chaque côté du plan des ponts hydrogènes reliant les bases (*Figure 1.9*).

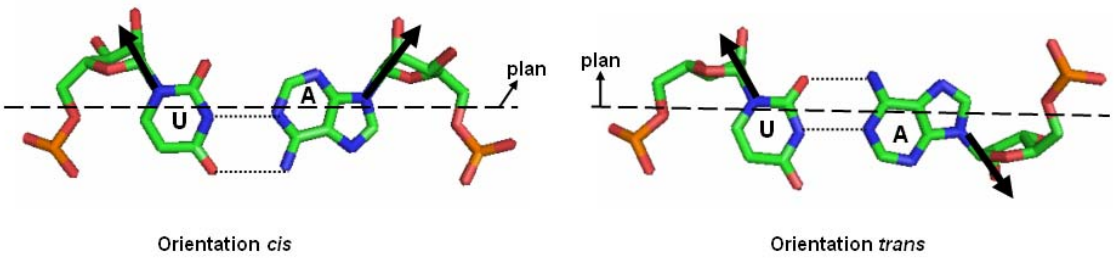


Figure 1.9. Orientation des liens glycosidiques. Les liens glycosidiques reliant la base à son ribose (les liens sont représentés par les flèches) peuvent être dans l'orientation *cis* ou *trans*. Si les liens glycosidiques se trouvent du même côté du plan des ponts hydrogènes (lignes pointillées), l'orientation est *cis*. S'ils se trouvent de chaque côté opposé, l'orientation est *trans*.

En combinant les appariements possibles entre deux bases avec les deux orientations des liens glycosidiques, il existe douze familles d'appariement (*Table 1.1*) (Leontis & Westhof, 2003).

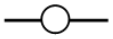
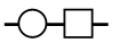
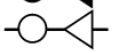
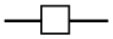
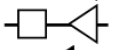
Symboles représentant les douze familles d'appariements		
Appariements	**Orientation du lien glycosidique**	**Symboles**
W/W	Trans	
W/W	Cis	
W/H	Trans	
W/H	Cis	
W/S	Trans	
W/S	Cis	
H/H	Trans	
H/H	Cis	
H/S	Trans	
H/S	Cis	
S/S	Trans	
S/S	Cis	

Table 1.1 : Douze familles d'appariement distinctes. Les familles sont caractérisées par les faces des bases impliquées dans l'appariement et par l'orientation des liens glycosidiques. Chaque face est représentée par un symbole: Watson-Crick (W) est représenté par un cercle, Hoogsteen (H) représenté par un carré et le côté Sucre (S) représenté par un triangle. Pour l'orientation des liens glycosidiques: *cis* (le symbole est plein) et *trans* (le symbole est vide).

1.5.3 Utilité de la nomenclature Leontis-Westhof

La nomenclature facilite la description des relations isostériques parmi les paires de bases appartenant à la même famille d'appariement. Elle facilite également la représentation des structures en des graphes d'interactions (le concept de graphe sera introduit à la prochaine section). La structure tertiaire de l'ARNr 5S d'*Haloarcula marismortui* de la *Figure 1.10* en est un exemple. La nomenclature est utilisée pour représenter les éléments structuraux contenant les appariements.

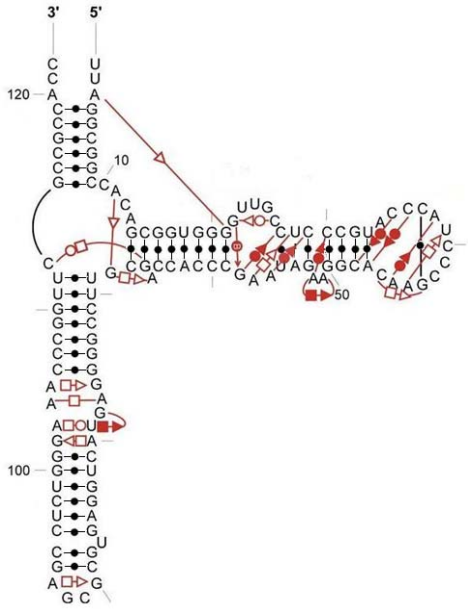


Figure 1.10. Graphe d'interactions du 5S de l'ARNr *Haloarcula marismortui*. Les interactions sont indiquées selon la nomenclature Leontis–Westhof. Les interactions en rouges sont les interactions tertiaires. Les appariements ont été identifiés dans la structure cristalline de l'ARNr *5S Haloarcula marismortui* (1JJ2.pdb).

1.6 Les graphes d'interactions

Comme la structure secondaire, la structure tertiaire de l'ARN peut être décrite sous forme de graphe (Major et al., 1991; Lemieux et al., 1998; Bermúdeza et al., 1999; Gan et al., 2003). Un graphe consiste en un ensemble de sommets et d'arêtes. Les sommets sont les nucléotides, alors que les arêtes représentent les interactions entre les nucléotides. Les graphes servent de structure de données en informatique, c'est-à-dire qu'ils permettent d'organiser des ensembles d'objets (dans le cas de l'ARN, des nucléotides) reliés par des relations abstraits (ici les interactions chimiques). Prenons l'exemple d'une section du graphe d'interactions de l'ARNr 5S (*Figure 1.11*). L'interprétation du graphe de cette section (76-80; 102-105) se fait comme suit: G76 est lié à A105 par une interaction S/H *trans*; A77 est lié à A104 par une interaction H *trans*; A77 est lié à G78 et celui-ci est lié à U79 par une interaction S/H *cis*; U79 est lié à A103 par une interaction W/H *trans* et A80

est lié G102 par une interaction H/S *trans*. La base G76 est empilée sur A77, la base U79 sur A80 et la base A104 sur A105. Alors, un graphe structurel peut contenir toutes les informations nécessaires sur une molécule d'ARN.

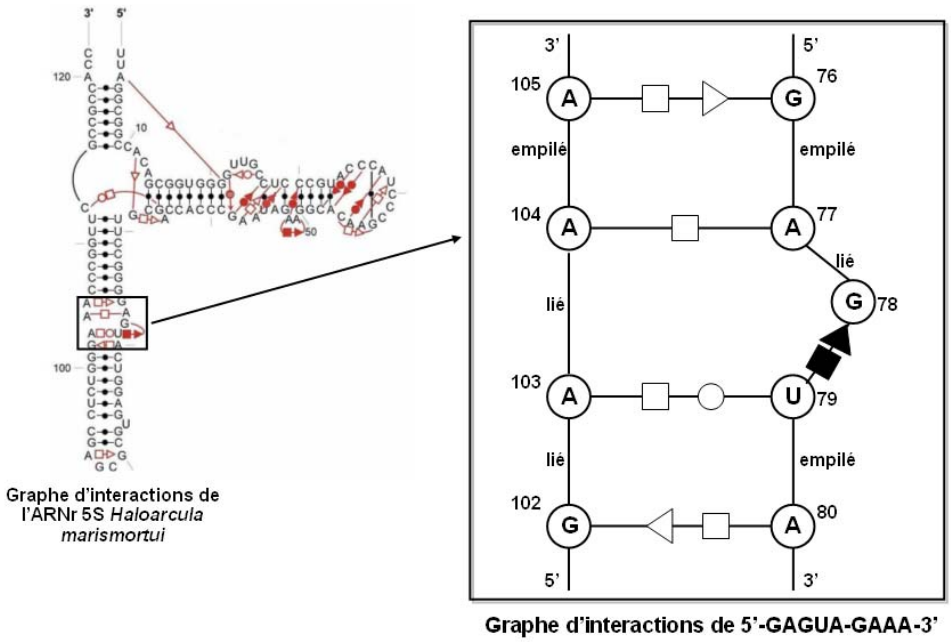


Figure 1.11 : Graphe structurel agrandi de 5'-GAGUA-GAAA-3'. Dans le graphe, les nucléotides sont les sommets. Les liaisons inter-nucléotidiques sont les arêtes du graphe.

Plusieurs algorithmes ont été développés avec des graphes d'interactions pour analyser les molécules d'ARN. Par exemple, notre laboratoire a créé un programme informatique nommé *MC-Annotate* (Gendron et al., 2001). L'outil prend en entrée les coordonnées atomiques de la structure tertiaire d'un ARN, et génère le graphe d'interactions tel que celui de la *Figure 1.11*.

Comme les éléments structuraux de l'ARN peuvent être caractérisés par des graphes d'interactions, ils peuvent être recherchés dans les structures tertiaires d'ARN selon un

algorithme d'isomorphisme[4] de sous-graphes (Aho et al., 1974; Lengauer, 2007; Major & Thibault, 2007). Notre laboratoire a implanté un algorithme d'isomorphisme de sous-graphes dans *MC-Search* (Gendron et al., 2001; Hoffmann et al., 2001). L'outil prend en entrée la description d'un élément structural (descripteur) et une base de données de structures d'ARN. La base de données est préalablement annotée avec *MC-Annotate*, puis le descripteur est transformé en un graphe cible qui sera recherché, comme sous-graphe, dans les graphes d'interactions de la base de données. *MC-Search* trouve tous les fragments d'ARN correspondant au graphe cible dans les structures de la base de données. Les fragments d'ARN sont retournés en format PDB (Berman et al., 2000). La *Figure 1.12* donne un exemple d'un descripteur décrivant le graphe d'interactions d'une tige-boucle.

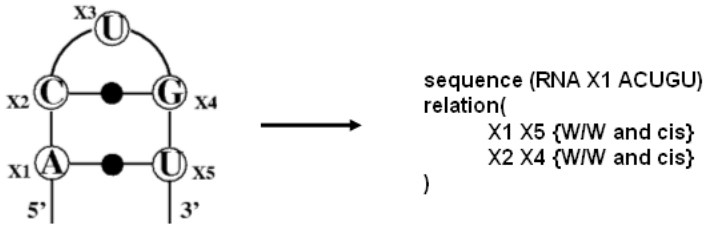


Figure 1.12 : Descripteur pour une tige-boucle 5'-ACUGU-3'. Le descripteur (droite) contient deux sections : une pour la séquence et l'autre pour les relations. La section « **séquence** » représente les sommets du graphe. Elle est formée du type de chaîne **RNA**, d'un identificateur de nucléotide (caractère suivi d'un entier) et de la séquence de nucléotides. Le premier nucléotide de la séquence recevra cet identificateur, le deuxième recevra l'identificateur où l'entier est incrémenté et ainsi de suite. Donc, dans le descripteur, X1 réfère à A, X2 à C, X3 à U, X4 à G et X5 à U.

La section « **relation** » définit les arêtes du graphe. Les expressions dans cette section sont composées de deux sommets suivis des symboles, entre {}, qui décrivent la relation. Les symboles sont placés de chaque cotés d'un /. Ils peuvent faire partie de la nomenclature Leontis-Westhof : H pour Hoogsteen, S pour sucre, W pour Watson-Crick, cis et trans pour préciser l'orientation glycosidique des relations. Ainsi, dans le descripteur, les nucléotides X1 et X5 (A et U) ont une relation Watson-Crick *cis*; les nucléotides X2 et X4 (C et G) ont également une relation Watson-Crick *cis*.

[4] Deux graphes $G = (S, A)$ et $G' = (S', A')$ sont dits isomorphes (identiques ou égaux) s'il existe une bijection $f : S \rightarrow S'$ telle que deux sommets s_i et s_j sont adjacents dans G si et seulement si leur images $f(s_i)$ et $f(s_j)$ sont deux sommets adjacents dans G'. En d'autres mots, on peut renommer les sommets de G avec les étiquettes des sommets de G' sans modifier les arêtes correspondant dans G et G'.

1.7 Les motifs cycliques de l'ARN

Un graphe d'interactions peut contenir des cycles. On décrit un cycle sous forme d'une liste de sommets dans laquelle le premier sommet et le dernier sommet sont reliés. Par exemple, sur le graphe d'interactions de la tige-boucle 5'-ACUGU-3' de la *Figure 1.12*, {A, C, U, G, U} est un cycle. Il est à noter que l'ordre des sommets dans un cycle n'est pas important. Donc, le cycle {A, C, U, G, U} équivaut au cycle {C, U, G, U, A}. Les cycles peuvent être séparés en cycles indivisibles. On considère un cycle indivisible un cycle qui ne contient pas d'autres cycles. Par exemple, le cycle {A, C, U, G, U} contient deux cycles indivisibles : {A, C, G, U} et {C, U, G}.

Lemieux et Major ont découvert que l'ARNr 23S est composé d'un nombre de petits fragments cycliques et indivisibles (Lemieux & Major, 2006). Pour arriver à cette découverte, les chercheurs ont annoté, avec *MC-Annotate*, la structure tertiaire de l'ARNr 23S. Par la suite, ils ont implanté l'algorithme de Horton (Horton, 1987) sur le graphe d'interactions de l'ARNr. L'algorithme de Horton affiche une base minimale de cycles indivisibles d'un graphe. Lemieux et Major ont extrait les structures 3D de la base des cycles et ils ont effectué une classification hiérarchique de ces cycles basée sur la dérivation RMSD[5]. Cela indique que les cycles sont répétés dans la structure d'ARN, donc ils sont considérés comme des motifs.

En assumant que les cycles sont répétés et qu'il existe une relation entre la séquence et la structure, on peut prédire les cycles qui apparaissent dans la structure d'une certaine séquence. Pour montrer l'aspect prédictif de la formation des cycles, Parisien et Major ont formé une base de dix-neuf sortes de motifs cycliques les plus représentatifs dans les structures 3D d'ARN (*Figure 1.13*) (Lemieux & Major, 2006; Parisien & Major, 2008). Ces motifs cycliques forment, pour le moment, l'alphabet structural secondaire de l'ARN : cinq types de motifs cycliques à simple brin représentant des boucles d'ARN (*Figure*

1.13a), et quinze types de motifs cycliques à double brin. Certains de ces motifs à double brin représentent deux paires de bases (*Figure 1.13b*), d'autres des bourgeons (*Figure 1.13c*) et des boucles internes (*Figure 1.13d*). Les motifs cycliques sont décrits par les appariements, les interactions d'empilement entre les bases et les liens phosphodiesters.

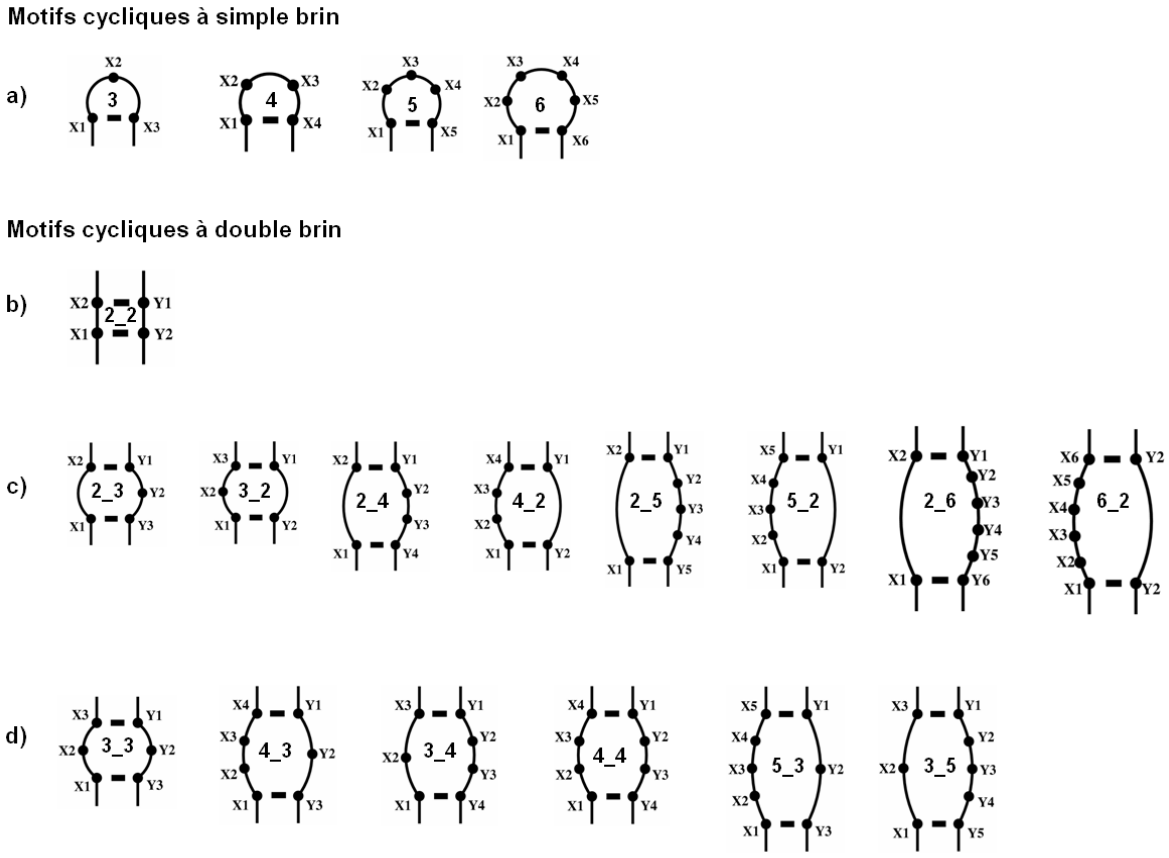


Figure 1.13 : L'alphabet structural. Base de motifs cycliques observés dans les structures d'ARN résolues expérimentalement. Les points noirs et les variables (Xn et Yn) représentent des nucléotides et les traits des appariements dans les motifs cycliques à double brin. Les nucléotides Xn appartiennent au brin d'ARN allant de 5' → 3' et les nucléotides Yn sur le brin 3' → 5'. Pour les motifs cycliques à simple brin, la notation « m » signifie m nucléotides sur la boucle. Pour les motifs cycliques à double brin, la notation « m_n » signifie m nucléotides sur le brin 5' → 3' et n nucléotides sur le brin 3' → 5'.
(a) Motifs cycliques représentant les boucles de nucléotides. **(b)** Motif cyclique représentant deux paires de bases empilées l'une sur l'autre. **(c)** Motifs cycliques représentant les bourgeons. **(d)** Motifs cycliques représentant les boucles internes.

---

[5] **R**oot **M**ean **S**quare **D**eviation

Parisien et Major ont développé un algorithme, *MC-Fold*, qui utilise les motifs cycliques comme élément de base pour prédire la structure secondaire d'un ARN (Parisien & Major, 2008). Le programme génère plusieurs structures secondaires selon leurs probabilités d'occurrence à partir d'une séquence donnée (*Figure 1.14*).

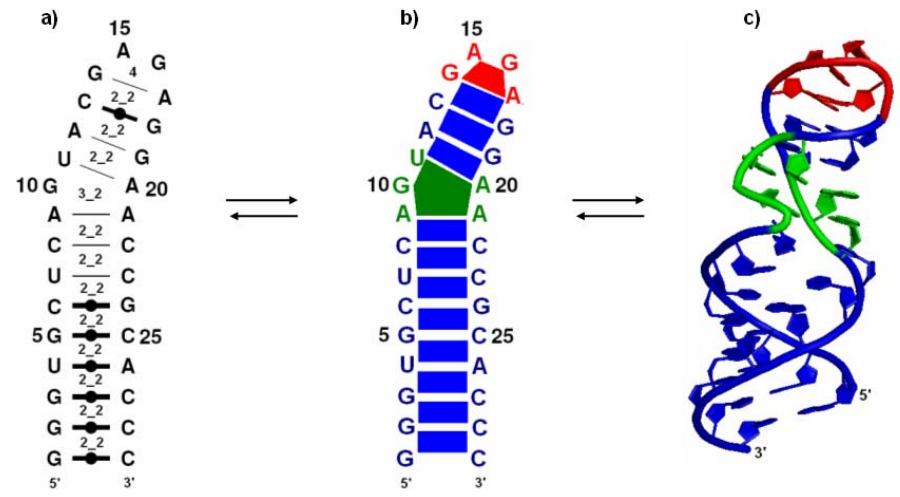


Figure 1.14 : Structure secondaire de la boucle E prédite par *MC-Fold*. La séquence de la boucle E provient de la structure cristallographique 430d.pdb **(a)** Exemple d'une structure secondaire de la boucle E prédite par *MC-Fold*. Les traits foncés accompagnés d'un cercle correspondent aux appariements Watson-Crick et les traits fins correspondent aux appariements non-classiques. **(b)** Structure composée d'une suite de treize motifs cycliques : huit motifs cycliques 2_2 (bleu), suivi d'un motif cyclique 3_2 (vert), trois motifs cycliques 2_2 et un motif cyclique 4 (rouge). **(c)** Structure tertiaire de la boucle E où les couleurs correspondent aux motifs cycliques.

Par ce nouveau formalisme, la structure secondaire d'ARN est une suite de motifs cycliques.

1.8 But du mémoire

Les motifs structuraux sont définis par leurs séquences, leurs interactions nucléotidiques et leurs conformations structurales (tige-boucle, bourgeons, jonctions d'hélices, etc.). Il existe des programmes informatiques, tel que *RNAMotif* (Macke et al.,

2001), qui utilisent ces informations pour rechercher des motifs dans les structures secondaires de l'ARN.

Dans notre laboratoire, Lemieux et Major ont remarqué que les nucléotides dans les structures 3D d'ARN formaient localement des motifs cycliques (Lemieux & Major, 2006). Par la suite, Parisien et Major ont généralisé en proposant dix-neuf motifs cycliques fondamentaux à partir desquels on peut générer toute structure secondaire (Parisien & Major, 2008). Ainsi, peut-on utiliser les motifs cycliques pour représenter des motifs fonctionnels d'ARN? Si oui, peut-on chercher ces motifs d'ARN définis en terme de motifs cycliques dans des structures secondaires définies également par leurs motifs cycliques? Dans ce mémoire, je vous propose une approche qui répond à ces questions.

Au prochain chapitre, je présente des motifs structuraux fonctionnels de l'ARN en utilisant les formalismes de *MC-Search* et des motifs cycliques. J'introduis également ma base de données de motifs qui sera utilisé par mon algorithme de recherche, *MC-Motifs*, pour indiquer la présence possible de motifs fonctionnels dans une structure secondaire.

CHAPITRE 2

BASE DE DONNÉES DE MOTIFS DE L'ARN

Dans ce chapitre, je présente quelques exemples de motifs structuraux fonctionnels classiques : la tétraboucle GNRA, la boucle T, le motif sarcine-ricine, le kink-turn, le motif C et le motif « UA_handle ». Ces motifs ont tous été observés dans des structures cristallines d'ARN. Ils ont été utilisés pour établir ma base de données de motifs (base de données *SQL*) appelée *bdMotifs* qui, rappelons-le, sera employée par mon algorithme de recherche de motifs, *MC-Motifs*.

bdMotifs contient les noms des motifs fonctionnels, leurs représentations structurales en termes de motifs cycliques et les séquences des occurrences de ces motifs tels que déterminé par *MC-Search*. Le chapitre est divisé en six sections, où chaque section contient la description du motif, le graphe d'interactions de celui-ci et les résultats de recherche de *MC-Search* dans les ARN dont la structure a été déterminée expérimentalement (ARNm, ARNr, ARNt, ribozymes, etc.).

2.1. La tétraboucle GNRA

Dans les tiges-boucles, les boucles sont souvent des tétraboucles (boucle composée de 4 nucléotides) dont la plupart des occurrences montrent des séquences compatibles au GNRA (N symbolise l'une des quatre bases; R symbolise A ou G) (Leontis & Westhof, 2003). Le motif GNRA a été observé dans les ARNr, les introns du groupe I et II, et la ribonucléase P, un ribozyme (Correll & Swinger, 2003). Il crée des interactions tertiaires avec les éléments structuraux éloignés et forme des sites de liaisons avec des protéines et des ARN (*Figure 2.1*).

Ce motif possède un appariement Sucre/Hoogsteen (S/H) *trans* entre les bases G et A et une liaison hydrogène entre le groupement 2'-hydroxyle (OH) de la base G et la face Hoogsteen (H) de la base R. (Heus & Pardi, 1991). Il est représenté par un seul motif cyclique composé de quatre nucléotides* (*Figure 2.1b*).

Figure 2.1 : Exemple de la tétraboucle GNRA dans l'ARNr 23S d'*Haloarcula marismortui*. (a) Motif GNRA. Dans la figure, les cercles noirs représentent les nucléotides. L'appariement S/H *trans* a lieu entre G2696 et A2699. Le cercle contenant le symbole O2' signifie que la liaison a lieu entre l'oxygène du groupement 2'-hydroxyle du G2696 et la face H de la base G2698. **(b)** Motif GNRA représenté par un seul motif cyclique.

2.1.1 Génération du graphe représentant la tétraboucle GNRA

La *Figure 2.2* montre le graphe d'interactions et le descripteur du motif GNRA. Dans la section « relation » du descripteur, je n'ai pas spécifié l'appariement O2'/H *trans*, car, dans certaines structures GNRA provenant de la littérature, *MC-Annotate* donne l'annotation O2'/Bh *trans* au lieu de O2'/H *trans*.

Le symbole « Bh » signifie *bifurcated Hoogsteen*. L'expression « O2'/Bh » signifie que l'appariement a lieu entre le groupement 2'-hydroxyle d'une base et un atome situé entre la face Watson-Crick et la face Hoogsteen de l'autre base. J'ai alors remplacé l'appariement O2'/H *trans* par l'expression « O2'/(H or Bh) ».

* J'ai écrit un script en langage Perl, *pdb2NCM.pl*, qui permet de représenter la structure 3D d'un motif en termes de motifs cycliques. L'algorithme de ce script se trouve dans l'annexe A.

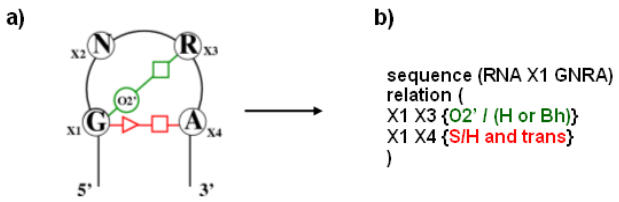


Figure 2.2 : Descripteur de la tétraboucle GNRA. (a) Graphe d'interactions de la tétraboucle GNRA. **(b)** Descripteur de la tétraboucle GNRA.

2.1.2 Résultats de recherche de la tétraboucle GNRA par *MC-Search*

Ce motif est en abondance dans les molécules d'ARN. *MC-Search* les a détecté dans les ARNr 5S, 16S et 23S, les ribozymes et les introns, tels que mentionné dans la littérature. La *Table 2.1* affiche quelques exemples de séquences de motif GNRA trouvées.

Séquences (5' → 3')	**Nombre d'occurrences**	**Positions des nucléotides composant le motif dans la structure PDB**	**Molécules**
GAAA	1	1KXK: 34-37	Domaine du 5 et 6 de l'intron Groupe II
GAAA	1	1HMH: 21-24	Structure cristallographique du ribozyme Hammerhead
GCGA	1	1JJ2: 90-93	ARNr 5S
GAAA	3	1J5E: 159-162 1013-1016 1166-169	ARNr 16S
GCAA	3	1J5E: 380-383 898-901 1266-1269	ARNr 16S
GAAA	2	1JJ2: 691-694 2412-2415	ARNr 23S
GUGA	1	1JJ2: 469-472	ARNr 23S

GCAA	2	1JJ2: 196-199 1863-1866	ARNr 23S
GUAA	2	1JJ2: 1055-1058 2877-2880	ARNr 23S
GGAA	1	1JJ2: 1794-1797	ARNr 23S
GAGA	1	2AWB: 2659-2662	ARNr 23S
GCGA	1	2AWB: 2857-2860	ARNr 23S

Table 2.1: Exemples de séquences des motifs GNRA détectés par *MC-Search*.

2.2 La boucle T

La boucle T est une boucle que l'on retrouve dans la tige-boucle TΨC des ARNt et dans les tiges-boucles des ARNr 30S et 50S (Nagaswamy & Fox, 2002; Krasilnikov & Mondragon, 2003; Zhuang et al., 2007). La boucle T contient cinq nucléotides (*Figure 2.3a*). Elle est caractérisée par la présence d'une paire de base Watson-Crick/Hoogsteen (W/H) *trans* entre les bases N1 et N5. Il est impliqué dans la formation de liaisons tertiaires avec d'autres éléments structuraux qui stabilisent la structure tertiaire. La boucle T est représentée par un seul motif cyclique composé de cinq nucléotides (*Figure 2.3b*).

Figure 2.3 : La boucle T. (a) Boucle T. Dans la figure, les cercles noirs représentent les nucléotides. Le motif structural boucle T est une pentaboucle (boucle contenant cinq nucléotides). L'interaction W/H *trans* entre les bases N1 et N5 caractérise le motif. **(b)** La boucle T est représentée par un seul motif cyclique.

2.2.1 Génération du graphe représentant la boucle T

La *Figure 2.4* montre le graphe d'interactions et le descripteur de la boucle T. Dans la section « sequence » du descripteur (*Figure 2.4b*), la séquence n'est pas spécifiée (la valeur N symbolise l'une des quatre bases : A, C, G, U), de sorte que *MC-Search* cherchera dans l'ARN tous les exemples possibles du graphe d'interactions correspondant à la boucle T. L'appariement W/H *trans* qui définit le motif est décrit dans la section « relation ».

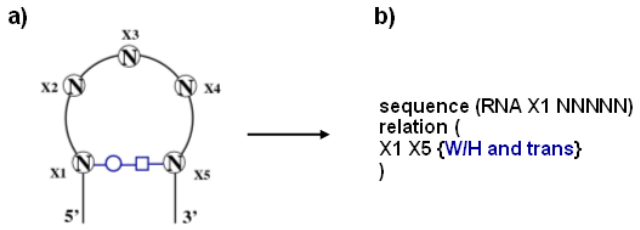


Figure 2.4 : Graphe d'interactions et descripteur de la boucle T. (a) Graphe d'interactions de la boucle T. **(b)** Descripteur de la boucle T. La valeur « N » dans la section « sequence » du descripteur symbolise l'une des quatres bases : A, C, G et U.

2.2.2 Résultats de recherche de la boucle T par *MC-Search*

MC-Search a détecté des occurrences de la boucle T dans les ARNr 16S, 23S et dans les ribonucléases P. Certaines sont les mêmes que celles qui sont mentionnées dans la littérature (Krasilnikov & Mondragon, 2003) (*Table 2.2*). Les séquences étoilées appartiennent aux séquences des motifs observés et analysés dans la littérature.

Séquences (5' → 3')	**Nombre d'occurrences**	**Positions des nucléotides composant le motif dans la structure PDB**	**Molécules**
UGAGA*	1	1J5E: 323-327 1315-1319	ARNr 16S
UGCAA*	1	1J5E: 1315-1319	ARNr 16S
UGGAA*	1	1JJ2 : 313-317	ARNr 23S
UGCAA*	1	1JJ2 : 481-485	ARNr 23S

CGAAA*	1	1JJ2: 505-509	ARNr 23S
UUUGA*	1	1JJ2: 624-628	ARNr 23S
UUAAA*	1	1JJ2: 2597-2601	ARNr 23S
UUUGU	1	2AWB: 567-571	ARNr 23S
AGUGA*	1	1NBS: 187-191	Ribonucléase P
UGGAA*	1	1NBS: 218-222	Ribonucléase P

Table 2.2: Séquences des boucles T détectées par *MC-Search*.

2.3 Le motif sarcine-ricine

Le motif sarcine-ricine est un bourgeon. Il porte ce nom car il fait partie d'une tige-boucle nommée la boucle ribosomale E. Cette tige-boucle interagit avec des protéines nommées EF-1 et EF-2. La liaison entre ces protéines et la boucle ribosomale E est inhibée par des enzymes α-sarcine et ricine bloquant la synthèse protéique (Szewczak et al., 1993; Noller, 2005a; Spackova & Sponer, 2006). Cependant, le motif sarcine-ricine ne se trouve pas seulement dans les tiges-boucles, mais aussi dans les jonctions d'hélices (Leontis & Westhof, 1998). Le motif est conservé, entre autres, dans les ARNr 23S et 28S (Leontis & Westhof 1998; Spackova & Sponer 2006).

Le sarcine-ricine comporte un ensemble d'appariements non Watson-Crick : (1) une paire Hoogsteen (H/H) *trans* entre A et A; (2) un G expulsé hors de l'hélice créant ainsi un bourgeon; une paire Sucre/Hoogsteen (S/H) *cis* avec le nucléotide voisin U; (3) une paire Watson-Crick/Hoogsteen (W/H) *trans* entre U et A; (4) une paire Hoogsteen/Sucre (H/S) *trans* entre A et G (*Figure 2.5a*). Le motif est représenté par un ensemble de deux motifs cycliques à double brin : 3_2 et 2_2 (*Figure 2.5b*).

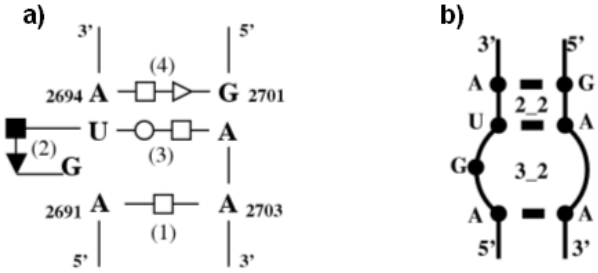


Figure 2.5 : Motif sarcine-ricine de l'ARNr 23S d'*Haloarcula marismortui*. **(a)** Le motif correspondant aux nucléotides 2691-2694 et 2701-2703 se trouve dans le domaine VI de l'ARNr 23S d'*Haloarcula marismortui* (Leontis & Westhof, 2002). Les appariements définissent le motif sarcine-ricine : (1) H/H *trans*; (2) S/H *cis*; (3) W/H *trans* et (4) H/S *trans*. **(b)** Le motif est composé de deux motifs cycliques : 3_2 et 2_2.

2.3.1 Génération du graphe représentant le motif sarcine-ricine

Comme pour la boucle T, je cherche avec *MC-Search* tous les exemples possibles du graphe correspondant au motif sarcine-ricine. Donc, dans la section «séquence» du descripteur (*Figure 2.6a* et *2.6b*), je n'ai spécifié aucune séquence particulière. Les appariements qui définissent le motif sont décrits dans la section «relation» du descripteur.

Le descripteur permet à *MC-Search* de trouver des motifs sarcine-ricine qui se forment au niveau de la structure secondaire seulement, car il existe des motifs sarcine-ricine qui se forment dans la structure tertiaire (Leontis et al., 2002). Ces motifs sont composés de nucléotides appartenant à trois brins différents d'ARN ou plus. Le motif sarcine-ricine aux positions 953-955 et 1012-1014 dans l'ARNr 23S d'*Haloarcula marismortui* est un exemple (*Figure 2.6c*). Il est composé de deux brins d'ARN : l'A1014 forme l'appariement Hoogsteen *trans* avec une base A2302 appartenant à un brin d'ARN éloigné. Pour ma recherche, je ne tiens pas compte de ce type de motif.

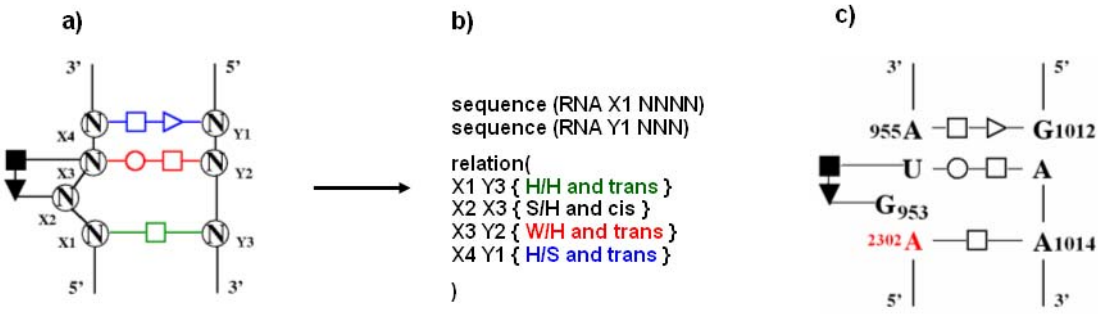


Figure 2.6: Descripteur du motif sarcine-ricine. a) Graphe d'interactions du motif sarcine-ricine. Les brins d'ARN dans le graphe sont arbitrairement nommés X et Y. **b)** Descripteur utilisé pour la recherche des occurrences du motif dans les structures d'ARN avec *MC-Search*. **c)** Ce motif est formé dans la structure tertiaire où la base A2302 (rouge) provenant d'un autre brin d'ARN forme l'appariement Hoogsteen *trans* avec A1014.

2.3.2 Résultats de recherche de sarcine-ricine par *MC-Search*

MC-Search a trouvé des structures sarcine-ricine ayant la séquence 5'-AGUA-GAA-3' dans les ARNr 5S, 16S et 23S. Une structure ayant la séquence 5'-AGUA-AAA-3' a été détectée dans l'ARNr 16S d'*Escherchia coli*. La *Table 2.3* affiche les séquences de sarcine-ricine. Les séquences étoilées appartiennent aux séquences des motifs sarcine-ricine observés et analysés dans la littérature.

Séquence (5' → 3')	**Nombre d'occurrences**	**Positions des nucléotides composant le motif dans la structure PDB**	**Molécules**
AGUA-GAA*	1	1JJ2: 77-80; 102-104	ARNr 5S
AGUA-GAA*	2	1J5E: 889-892; 906-908 1346-1349; 1373-1375	ARNr 16S
AGUA-AAA	1	2AVY: 889-892; 906-908	ARNr 16S
AGUA-GAA*	6	1JJ2: 174-177; 159-161 212-215; 225-227	ARNr 23S

		380-383; 406-408 463-466; 475-477 1369-1372; 2053-2055 2691-2694; 2701-2703	ARNr 23S

Table 2.3 : Séquences des motifs sarcine-ricine détectés par *MC-Search*.

2.4 Le kink-turn

Le motif kink-turn est une boucle interne asymétrique. Il a été identifié dans les ARNr du 16S *Thermus thermophilus* et du *23S Haloarcula marismortui* (Klein et al., 2001; Lescoute et al., 2005), l'ARN nucléaire U4 (ARNsn U4), les introns de l'ARNm et les riborégulateurs (Turner & Lilley, 2008). Il a été identifié car la boucle interne cause une courbure « kink » de 120 degrés dans le squelette phosphodiester de l'ARN, d'où son nom.

Le kink-turn est impliqué dans plusieurs processus métaboliques comme la traduction des ARNm et le contrôle d'expressions génétiques (Turner & Lilley 2008). La *Figure 2.7a* affiche les appariements qui caractérisent le motif (Lescoute et al. 2005). Le motif peut être représenté par deux ensembles de deux motifs cycliques à double brin : 2_2 et 6_2 (*Figure 2.7b*), et, 6_2 et 2_2 (*Figure 2.7c*).

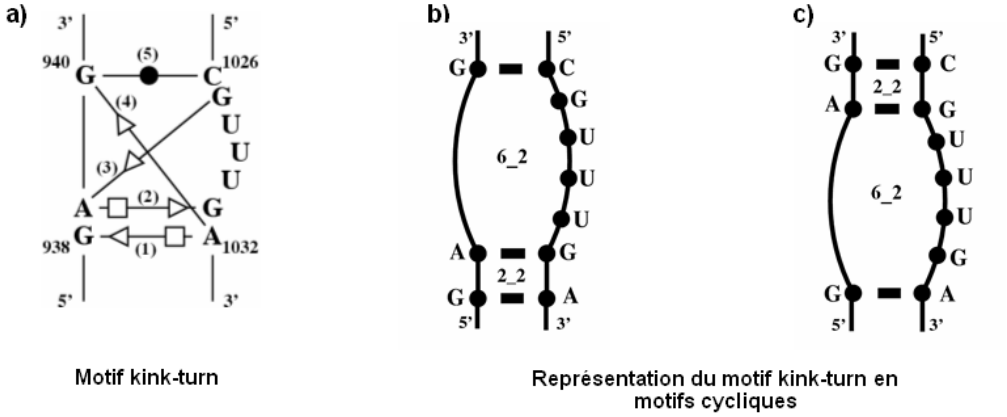


Figure 2.7 : Motif structural de kink-turn de l'ARNr 23S d'*Haloarcula marismortui*. (a) Le motif correspondant aux nucléotides 938-940 et 1026-1032 se trouve dans l'hélice 38 de l'ARNr 23S d' *H. marismortui* (Leontis & Westhof, 2002). Les cinq paires de bases caractérisant le motif sont:

- (1) et (2) → deux paires Sucre/Hoogsteen (S/H) *trans*
- (3) et (4) → deux paires de bases Sucre/Sucre (S/S) *trans*
- (5) → une paire canonique Watson-Crick (W/W) *cis*

(b) Le motif peut être composé de deux motifs cycliques 2_2 et 6_2 ou **(c)** 6_2 et 2_2.

2.4.1 Génération du graphe représentant le kink-turn

Dans la littérature, il existe différents variants du kink-turn. Un membre de notre laboratoire a étudié le motif. Il a remarqué que les séquences et les appariements intra- et extra-motif ne sont pas conservés parmi les structures du kink-turn. Par conséquent, il a découvert un ensemble d'interactions nécessaires et suffisantes pour garantir l'existence du motif au sein de la molécule d'ARN (communication personnelle). La figure ci-dessous montre deux descripteurs *MC-Search* pour localiser les kink-turn (*Figure 2.8b* et *2.8d*). Les descripteurs spécifient onze nucléotides sur deux brins nommés arbitrairement X et Y. Il y a un appariement S/H *trans* ou W/H *trans* entre X1 et Y6, plus une interaction d'empilement «stacking» entre X2 et Y6. Leur ressemblance s'arrête là. Dans le premier descripteur (*Figure 2.8b*), il y a un appariement S/S *trans* entre X3 et Y6. Au deuxième descripteur (*Figure 2.8d*), une interaction d'empilement entre X2 et Y3.

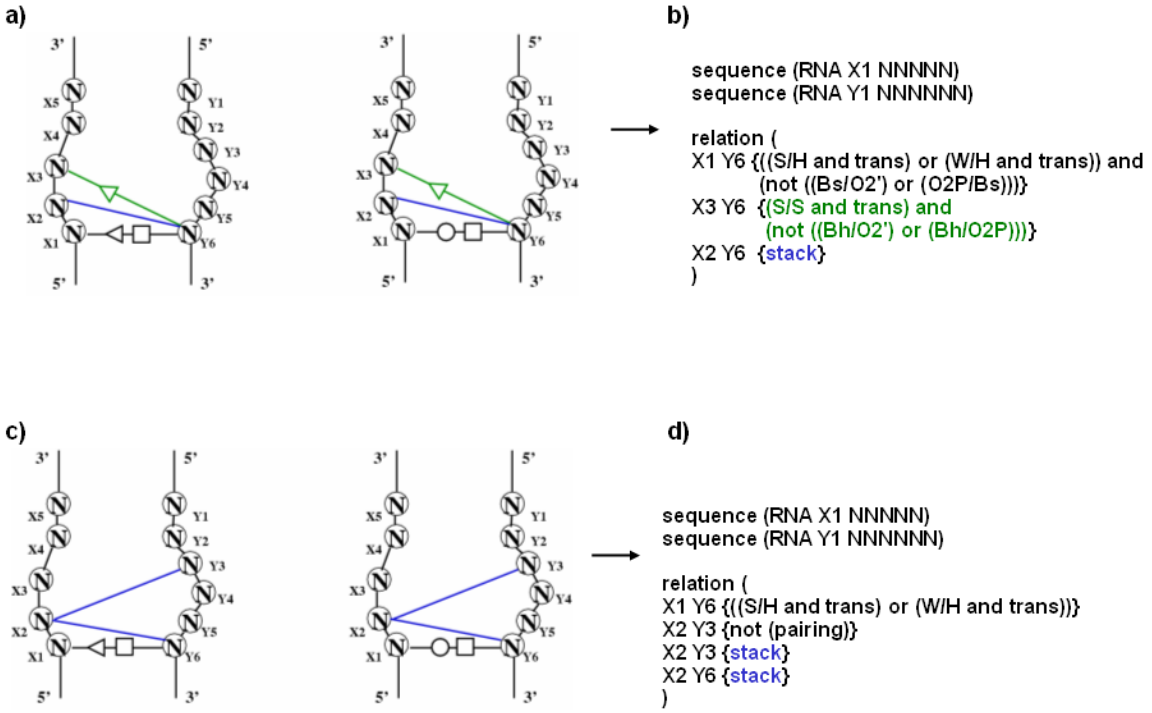


Figure 2.8 : Descripteurs pour localiser des kink-turn. L'utilisation des deux descripteurs permet de localiser des kink-turn dans les ARN. **(a)** Graphes d'interactions PPS (Pairing Pairing Stacking) car il contient deux appariements entre les nucléotides X1 et Y6, et, X3 et Y6 respectivement, et un empilement entre les nucléotides X2 et Y6. **(b)** Descripteur des graphes d'interactions PPS. Dans la section « relation », les expressions « not (((Bs/O2') or (O2P)))» et « not (((Bh/O2') or (Bh/O2P))) » signifient que les appariements doivent avoir lieu spécifiquement entre les faces des bases. **(c)** Graphes d'interactions PSS (Pairing Stacking Stacking) car il contient un appariement entre les nucléotides X1 et Y6, et deux empilements entre les nucléotides X2 et Y3, et, X2 et Y6. **(d)** Descripteur des graphes d'interactions PSS. Dans la section « relation », l'expression « X2 Y3 {not (pairing)} » signifie qu'il ne doit pas y avoir une paire de base entre les nucléotides X2 et Y3.

2.4.2 Résultats de recherche de kink-turn par *MC-Search*

Avec les descripteurs PPS et PSS, *MC-Search* a détecté sept kink-turn dans les ARNr 16S et 23S (*Table 2.4*). Ces kink-turn ont tous été observés et mentionnés dans la littérature.

Séquence (5' → 3')	Nombre d'occurrences	Positions des nucléotides composant le motif dans la structure PDB	Molécules
AUG-CGCAGA* (Kt-23)	1	1J5E: 686-688; 699-704	ARNr 16S
GAG-CGAAGA* (Kt-7)	1	1JJ2: 79-81; 93-98	ARNr 23S
GAAG-CAAUGU* (Kt-15)	1	1JJ2: 246-249; 260-265	ARNr 23S
GAG-CGUUUGA* (Kt-38)	1	1JJ2: 938-940; 1026-1032	ARNr 23S
CCUAGA-GAG* (Kt-42)	1	1JJ2: 1147-1152; 1214-1216	ARNr 23S
GAUGGA-GAC* (Kt-46)	1	1JJ2: 1312-1317; 1340-1342	ARNr 23S
GAAGC-GCAGGA* (Kt-58)	1	1JJ2: 1589-1593; 1601-1606	ARNr 23S

Table 2.4 : Séquences des motifs kink-turn détectés par *MC-Search*. Les séquences étoilées appartiennent aux séquences des motifs observés et analysés dans la littérature. Le symbole Kt suivi d'un chiffre réfère au kink-turn qui se trouve dans une hélice donnée. Par exemple, le Kt-23 signifie que le kink-turn se trouve dans l'hélice 23 de l'ARNr 16S.

Avec le descripteur PPS, *MC-Search* a identifié Kt-38. Avec le descripteur PSS, *MC-Search* a identifié Kt-58. Les autres kink-turn (Kt-7, Kt-15, Kt-23, Kt-42 et Kt-46) sont identifiés avec les deux descripteurs.

2.5 Le motif C

Le motif C est une boucle interne asymétrique (Leontis & Westhof 2003; Lescoute et al. 2005). Il a été découvert dans l'ARNm codant pour une enzyme, la thréonyl-ARNt synthétase (ThrRS) (Torres-Larios et al., 2002). Cette enzyme a la particularité de réguler sa propre expression : elle reconnaît son ARNm et le séquestre pour réprimer sa propre synthèse. La présence du motif C dans l'ARNm de ThrRS permet à l'enzyme de mieux adhérer à son ARNm. Le motif a également été observé dans les ARNr 16S et 23S (Leontis

& Westhof 2003; Lescoute et coll. 2005). Le motif C dans ces structures augmente la torsion hélicoïdale de l'ARN. Le nom du motif provient du fait que le premier nucléotide de la boucle est habituellement une cytosine et que celle-ci forme un appariement W/H *trans* avec une base impliquée dans un appariement W/W. Il possède un deuxième appariement, la paire W/S *cis*, formé avec une base impliquée dans un appariement W/W (*Figure 2.9*).

Comme le kink-turn, il existe des variants de motifs C dans l'ARN. Dans le motif, la longueur des deux brins peut varier dû à la présence de bourgeons. Le nombre maximal de nucléotides pouvant être ajoutés dans ces bourgeons n'est pas connu. Les Figures *2.9a* et *2.9c* affichent un exemple de deux motif structural C de type 3x5 et 6x4 respectivement. La notation m x n désigne le nombre de nucléotides de chaque côté de la boucle. Ainsi, la boucle de type 3 x 5 contient trois nucléotides sur le brin allant du 5' à 3' et cinq nucléotides sur le brin allant de 3' à 5'. Le motif C de type 3x5 est représenté par un seul motif cyclique 3_5 (*Figure 2.9b*) et celui de type 6x4 par un motif cyclique 6_4 (*Figure 2.9d*).

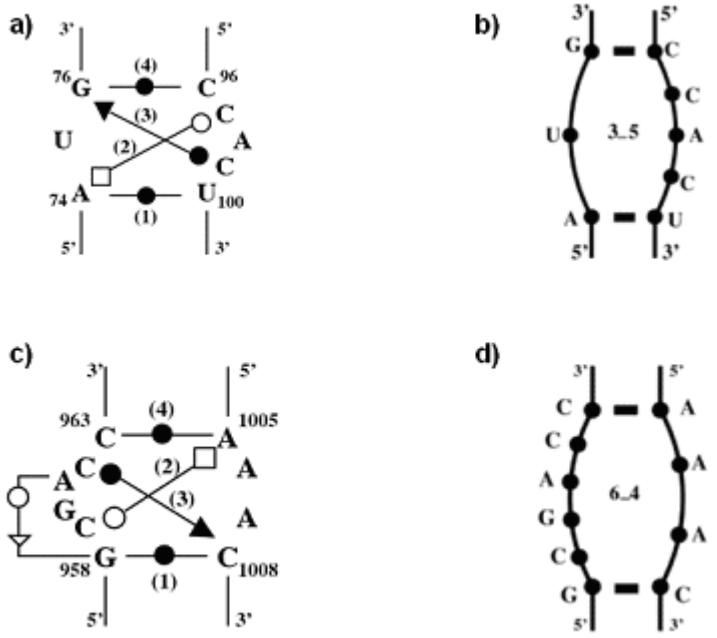


Figure 2.9 : Exemples des variants du motif C. (a) Motif C observé dans l'ARNm du thréonyl-ARNt synthétase (ThrRS). Il contient les appariements qui définissent le motif : (1) Watson-Crick (W/W) *cis*; (2) Watson-Crick/Hoogsteen (W/H) *trans*; (3) Watson-Crick/Sucre (W/S) *cis* et (4) Watson-Crick (W/W) *cis*. **(b)** Le motif est représenté par un motif cyclique 3_5. **(c)** Motif C observé dans le domaine II du 23S de l'ARNr d'*Haloarcula marismortui*. Il contient les mêmes appariements définissant le motif C et une paire de bases supplémentaire W/S *trans* dans le brin 5' → 3'. **(d)** Il est représenté par un motif cyclique 6_4.

2.5.1 Génération du graphe représentant le motif C

J'ai analysé la structure de quelques motifs C dans la littérature, avec *MC-Search* et *MC-Annotate*, afin de décrire le graphe d'interactions permettant de détecter tous les variants du motif. C'est-à-dire les motifs C de type 5xn et 6xn où n correspond à un nombre x de nucléotides. La *Figure 2.10c* montre les descripteurs de motif C.

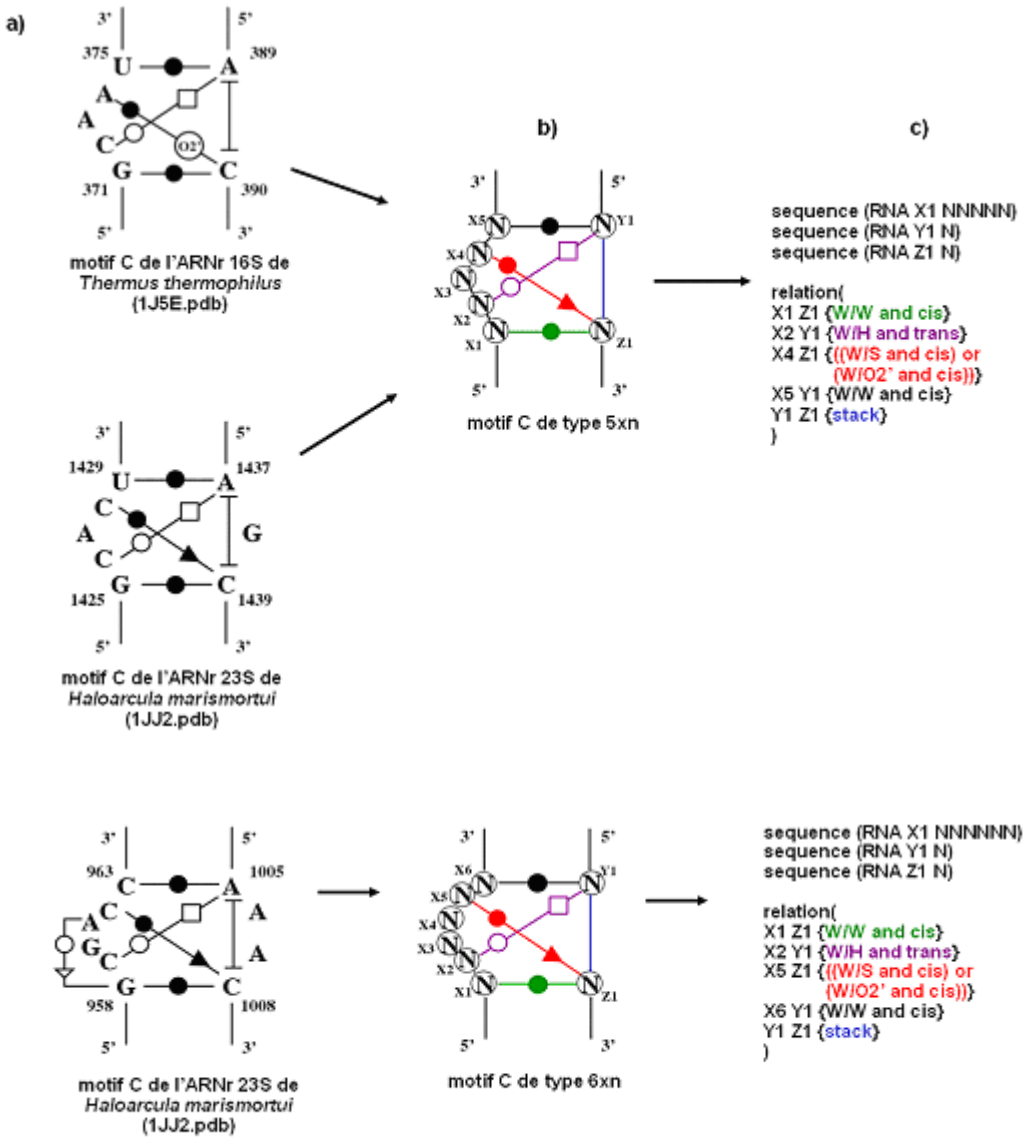


Figure 2.10 : Descripteurs de motif C. **(a)** D'autres exemples de variants de motif C. Les structures secondaires sont annotées par *MC-Annotate*. Le symbole (I) dans les motifs annotés signifie une interaction d'empilement entre les bases. **(b)** Les graphes d'interactions de chaque type de motif C. **(c)** Dans chaque descripteur, un brin d'ARN est nommé arbitrairement X. Un deuxième, contenant qu'un seul nucléotide, est nommé Y et un troisième, contenant également qu'un seul nucléotide, est nommé Z. J'ai décrit le motif de cette façon afin de considérer les insertions potentielles de nucléotides entre Y1 et Z1.

2.5.2 Résultats de recherche de motif C par *MC-Search*

MC-Search a détecté les motif C dans les ARNr 16S, 23S et dans l'ARNm codant pour l'enzyme thréonyl-ARNt synthétase (ThrRS), tel que mentionné dans la littérature (*Table 2.5* à 2*.7*).

Séquences (5' → 3')	**Nombre d'occurrences**	**Positions des nucléotides composant le motif dans la structure pdb**	**Molécules**
GCAAU-AC*	1	1J5E: 371-375; 389,390	ARNr 16S

Table 2.5: Séquence de motif C de type 5x2 trouvé par *MC-Search*. Il n'y a qu'un motif C de type 5x2 trouvé dans l'ARNr 16S du *Thermus thermophilus*. Ce motif a été observé et analysé dans la littérature.

Séquences (5' → 3')	**Nombre d'occurrences**	**Positions des nucléotides composant le motif dans la structure pdb**	**Molécules**
GCACU-AGC*	1	1JJ2: 1425-1429; 1437-1439	ARNr 23S
CCACU-ACG*	1	1JJ2: 2717-2721; 2761-2763	ARNr 23S
UCACU-AAA	1	2AWB: 2680-2684; 2725-2727	ARNr 23S
GCAAU-AGC	1	2HGJ: 1319-1323; 1331-1333	ARNr 23S
CCUCU-AUG	1	2HGJ: 2680-2684; 2725-2727	ARNr 23S
AUG-CCACU*	1	1KOG: 74-76; 96-100	ARNm codant pour ThrRS

Table 2.6: Séquences de motif C de type 5x3 trouvés par *MC-Search*. Les séquences étoilées appartiennent aux séquences des motifs observés et analysés dans la littérature

Séquences (5' → 3')	Nombre d'occurrences	Positions des nucléotides composant le motif dans la structure pdb	Molécules
GCGACC-AAAC*	1	1JJ2: 958-963; 1005-1008	ARNr 23S

Table 2.7: Séquence de motif C de type 6x4 trouvés par *MC-Search*. Il n'y a qu'un motif C de type 6x4 dans l'ARNr 23S d'*Haloarcula marismortui*. Cette structure est la même que celle observée et analysée dans la littérature.

2.6 Le motif « UA_handle »

Les motifs présentés aux sections précédentes sont des motifs qui se forment dans la structure secondaire, mais il existe des motifs qui se forment dans la structure tertiaire. Le motif « UA_handle » en est un exemple (Geary et al., 2008; Jaeger et al., 2008). Ce motif se forme lorsque les nucléotides qui le composent créent différentes interactions tertiaires avec d'autres nucléotides dans la structure d'ARN. Il est le motif le plus répandu dans les molécules d'ARN. Il est aussi considéré comme un sous-motif, car il entre dans la composition d'autres motifs plus complexes tels que la boucle T. Le « UA_handle » est présent dans les jonctions d'hélices, les pseudonœuds des ribosomes, et les sites de liaisons des riborégulateurs.

Le motif contient un appariement Watson-Crick/Hoogsteen (W/H) *trans*. Cette paire de base est empilée sur une paire de base Watson-Crick (W/W). Le motif possède un bourgeon composé d'un ou de plusieurs nucléotides. L'équipe de Jaeger a classifié les structures « UA_handle » selon le nombre de nucléotides présents dans le bourgeon (Jaeger et al., 2008). Il existe trois classes de « UA_handle » : type I pour un nucléotide dans le bourgeon, type II pour deux nucléotides et type III pour trois nucléotides (*Figure 2.11*).

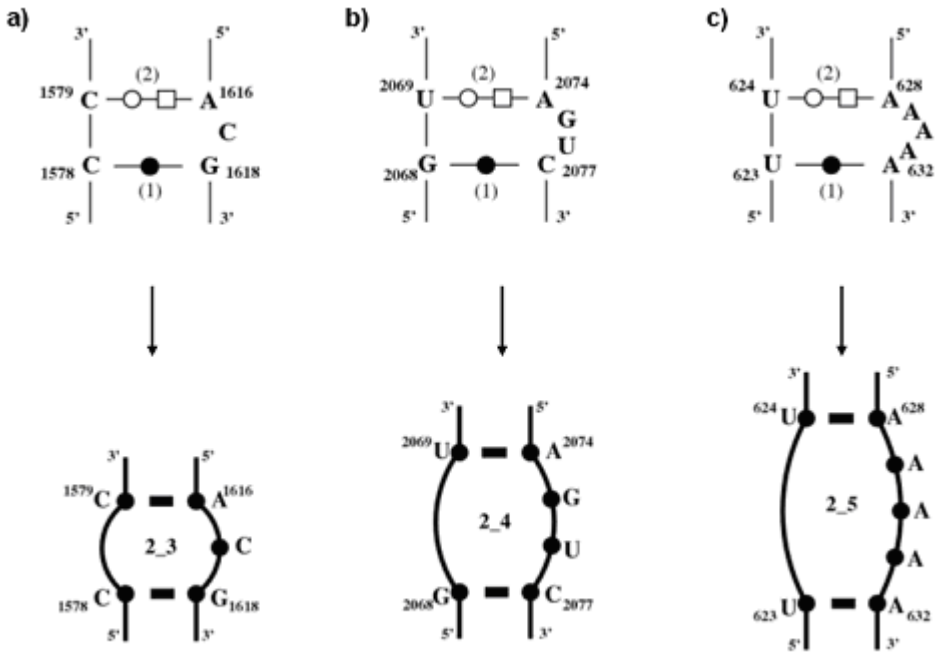


Figure 2.11 : Exemples de motif « UA_handle ». Ces motifs « UA_handle » ont été observés dans l'ARNr 23S d'*Haloarcula marismortui*. Ils possèdent un appariement W/W *cis* (1) et un appariement W/H *trans* (2). Il existe trois classes de motifs « UA_handle ». **(a)** Le « UA_handle » de type I qui est représenté par un seul motif cyclique 2_3. **(b)** Le « UA_handle » de type II qui est représenté par un seul motif cyclique 2_4. **(c)** Le « UA_handle » de type III qui est représenté par un seul motif cyclique 2_5.

2.6.1 Génération du graphe représentant le « UA_handle »

Je me suis référée à l'article de Jaeger pour décrire le motif « UA_handle » (Jaeger et al., 2008). Chaque classe de « UA_handle » a une ou plusieurs séquences consensus différentes. J'ai écrit des descripteurs pour chacune de ces classes (*Figure 2.12a-c*).

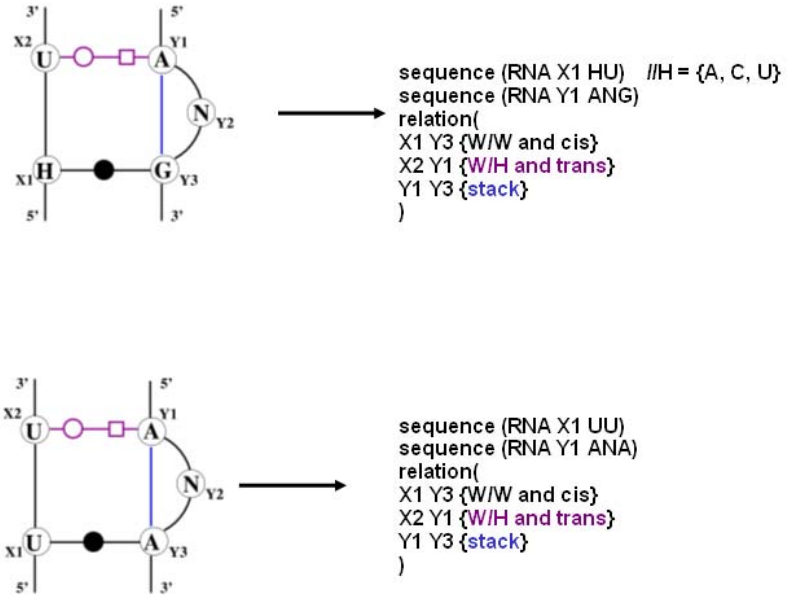


Figure 2.12a : Graphes d'interactions et descripteurs du motif « UA_handle » de type I.

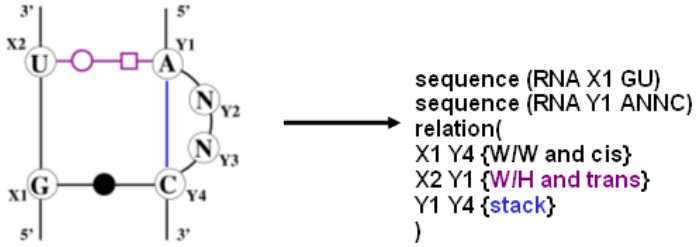


Figure 2.12b : Graphes d'interactions et descripteur du motif « UA_handle » de type II.

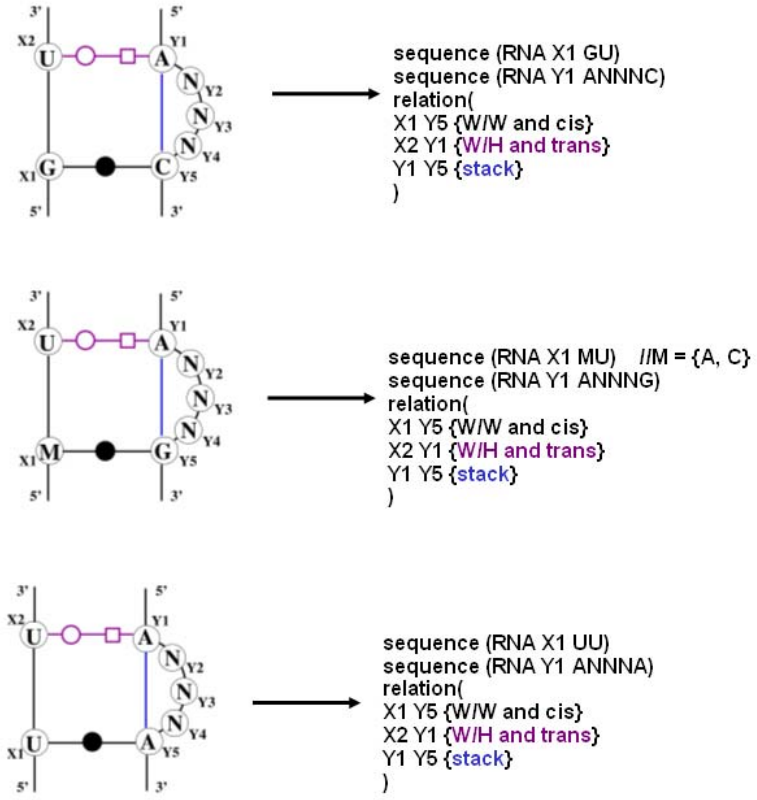


Figure 2.12c: Graphes d'interactions et descripteurs du motif « UA_handle » de type III.

2.6.2 Résultats de recherche de « UA_handle » par *MC-Search*

Le motif « UA_handle » est en abondance dans les molécules d'ARN. *MC-Search* a trouvé les motifs « UA_handle » dans les ARNr 16S et 23S. Je n'ai trouvé que des « UA_handle » de type I dans le ribozyme, l'intron et la ribonucléase P. Les Tables *2.8a-c* affichent quelques exemples de séquences de motifs « UA_handle » détectés selon chaque type.

Séquences (5' → 3')	Nombre d'occurrences	Positions des nucléotides composant le motif dans la structure pdb	Molécules
CU-AUG	1	1HR2 : 223-224; 248-250	Domaine du P4 et P6 du ribozyme (Groupe I)
CU-ACG	1	1J5E : 990-991; 1213-1215	ARNr 16S
UU-ACA	1	1JJ2 : 312-313; 317-319	ARNr 23S
UU-AAA	2	2AWB : 1018-1019; 1142-1143 1777-1778; 1785-1787	ARNr 23S
CU-AUG	1	1KXK : 15-16; 55-57	Domaine 5 et 6 de l'intron (Groupe I)
CU-AUG	1	1NBS : 146-147; 160-162	Ribonucléase P

Table 2.8a: Exemples de séquences des motifs « UA_handle » de type I trouvés par *MC-Search*

Séquences (5' → 3')	Nombre d'occurrences	Positions des nucléotides composant le motif dans la structure pdb	Molécules
GU-AUAC	1	1J5E : 515-516; 534-536	ARN 16S
GU-AACC	2	1JJ2 : 334-339; 340-342 1387-1388; 1392-1395	ARN 23S
GU-AUUC	1	1JJ2 : 1498-1499; 1504-1507	ARN 23S
GU-AUGC	1	2AWB : 2027-2028; 2033-2036	ARN 23S
AAAC-GU	1	2AWB : 1608-1611; 1620-1621	ARN 23S

Table 2.8b: Exemples de séquences des motifs « UA_handle » de type II trouvés par *MC-Search*

Séquences (5' → 3')	**Nombre d'occurrences**	**Positions des nucléotides composant le motif dans la structure pdb**	**Molécules**
CU-ACCCG	1	1J5E : 1314-1315; 1319-1323	ARN 16S
CU-AAAUG	2	1JJ2 : 1456-1457; 1485-1489	ARN 23S
AGAAC-GU	1	1JJ2 : 1682-1686; 1695-1696	ARN 23S
UU-AAAAA	1	1JJ2 : 623-624; 628-632	ARN 23S

Table 2.8c: Exemples de séquences des motifs « UA_handle » de type III trouvés par *MC-Search*

2.7 Conclusion

La *Table 2.9* affiche un exemple du contenu de ma base de données *bdMotifs*. Je vous rappelle que *bdMotifs* contient des informations sur les motifs fonctionnels recherchés par *MC-Search* ainsi que leurs représentations structurales en terme de motifs cycliques.

MOTIFS FONCTIONNELS	SÉQUENCES (5' → 3')	MOTIF(S) CYCLIQUE(S)
GNRA	GAGA	4
GNRA	GCGA	4
GNRA	GCAA	4
GNRA	GUGA	4
GNRA	GGAA	4
GNRA	GAAA	4
kink-turn (Kt-7)	GAG-CGAAGA	2_5 + 2_2
kink-turn (Kt-7)	GAG-CGAAGA	2_2 + 2_5
kink-turn (Kt-15)	GAAG-CAAUGU	2_5 + 3_2
kink-turn (Kt-15)	GAAG-CAAUGU	2_2 + 3_5
kink-turn (Kt-23)	UAG-CGCAGA	2_2 + 2_5
kink-turn (Kt-23)	UAG-CGCAGA	2_5 + 2_2
kink-turn (Kt-38)	GAG-CGUUUGA	2_6 + 2_2

kink-turn (Kt-38)	GAG-CGUUUGA	2_2 + 2_6
kink-turn (Kt-42)	CCUAGA-GAG	2_2 + 5_2
kink-turn (Kt-42)	CCUAGA-GAG	5_2 + 2_2
kink-turn (Kt-46)	GAUGGA-GAC	2_2 + 5_2
kink-turn (Kt-46)	GAUGGA-GAC	5_2 + 2_2
kink-turn (Kt-58)	GAAGC-GCAGGA	2_5 + 4_2
kink-turn (Kt-58)	GAAGC-GCAGGA	4_5 + 2_2
motif C type 5x2	GCAAU-AC	5_2
motif C type 5x3	GCACU-AGC	5_3
motif C type 5x3	CCACU-ACG	5_3
motif C type 5x4	GCACU-AAAC	5_4
motif C type 6x4	GCGACC-AAAC	6_4
sarcine-ricine	AGUA-GAA	3_2 + 2_2
sarcine-ricine	AGUA-AAA	3_2 + 2_2
boucle T	UGAGA	5
boucle T	UGCAA	5
boucle T	AGAGA	5
boucle T	CGAAA	5
UA_handle (type I)	CU-AUG	2_3
UA_handle (type I)	UU-AAA	2_3
UA_handle (type II)	GU-AUAC	2_4
UA_handle (type II)	AAAC-GU	4_2
UA_handle (type III)	UU-AAAAA	2_5
UA_handle (type III)	AGAAC-GU	5_2

Table 2.9 : Exemple du contenu de la base de données *bdMotifs*. Quelques résultats sur les représentations structurales des motifs fonctionnels, recherchés par *MC-Search*, en terme de motifs cycliques.

Au lieu de définir les motifs fonctionnels par leurs séquences et leurs interactions nucléotidiques, je les ai définis en gardant leurs séquences et en substituant leurs appariements par un ensemble de motifs cycliques. Les cycles nucléotidiques me permettent aussi de représenter la conformation structurale d'un motif donné. Donc, si on mentionne que le motif GNRA, 5'-GAAA-3', est composé d'un motif cyclique « 4 », alors on comprend que le motif GNRA est une tétraboucle. Si on mentionne, par exemple, que le kink-turn, 5'- CCUAGA-GAG-3' (Kt-46), est composé des motifs cycliques « 2_2 + 5_2 » ou « 5_2 + 2_2 », alors on comprend que ce motif est composé d'un bourgeon « 5_2 » et de

deux paires de bases empilée l'une sur l'autre « 2_2 » (voir la *Figure 1.13*). Ainsi, peut-on utiliser les motifs cycliques pour représenter des motifs fonctionnels d'ARN? Oui, on peut. L'étape suivante sera de développer un algorithme de recherche qui me permettra de détecter ces motifs dans des structures d'ARN.

Au chapitre suivant, je vous présente cet algorithme, *MC-Motifs*, et j'explique comment *MC-Motifs* emploie la base de données, *bdMotifs*, pour rechercher l'existence possible des motifs fonctionnels dans la structure secondaire d'une molécule d'ARN donnée.

CHAPITRE 3

RECHERCHE DE MOTIFS STRUCTURAUX ET FONCTIONNELS DANS UNE STRUCTURE SECONDAIRE

La recherche de motifs est un problème classique en informatique notamment en bio-informatique (Gotoh, 1987; Altschul et al., 1990; Gautheret et al., 1990). Il s'agit de repérer des sous-structures (des motifs) dans une structure. En informatique, les structures peuvent être, par exemple, des graphes ou des séquences finies de caractères. Je m'intéresse ici à la recherche d'un ou plusieurs motifs fonctionnels traduits en termes de motifs cycliques dans une structure secondaire d'ARN définie également par des motifs cycliques.

La structure secondaire d'ARN est habituellement représentée par une chaîne de points et parenthèses (Hogeweg & Hesper, 1984; Hofacker et al., 1994). Dans la chaîne, un point représente une base non appariée et une parenthèse, qu'elle soit gauche ou droite, représente une base appariée. La recherche des motifs structuraux et fonctionnels s'effectue au niveau de la structure secondaire donc au niveau de la chaîne de points et parenthèses où celle-ci est transformée en chaîne de motifs cycliques.

J'ai implanté un algorithme de recherche nommé *MC-Motifs*, ce programme a été développé en Java. Il prend en entrée la séquence nucléotidique de l'ARN en format FASTA (séquence commençant avec une ligne simple de description de la séquence), la chaîne de points et parenthèses représentant la structure secondaire et la position du premier nucléotide (*Figure 3.1*). Ces informations peuvent provenir des sorties de *MC-Fold*, de tout autre programme de prédiction de structure secondaire ou directement de la littérature.

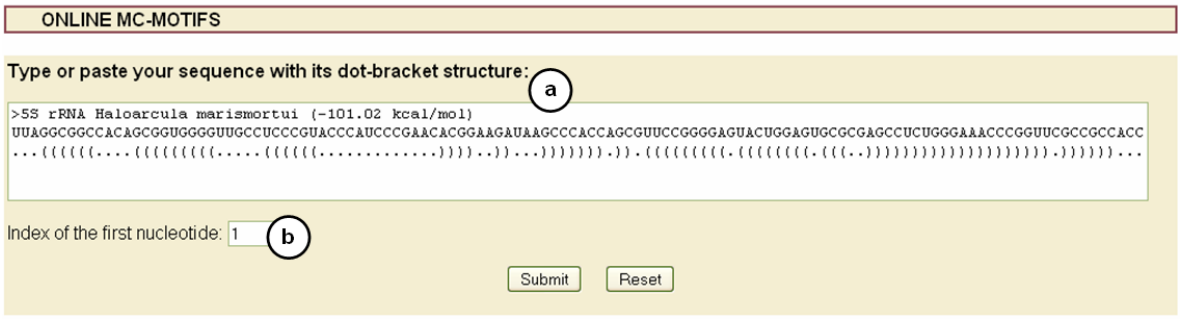


Figure 3.1 : Interface de saisie de données de *MC-Motifs*. (a) Dans ce champ, l'utilisateur tape ou copie, la séquence d'ARN en format FASTA et la chaîne de points et parenthèses représentant la structure secondaire. La structure secondaire dans la figure (ARNr 5S d'*Haloarcula marismortui*) a été prédite par *MC-Fold*. **(b)** Dans le deuxième champ, l'utilisateur doit indiquer la position du premier nucléotide.

Ce chapitre est divisé en trois sections. Les deux premières sections portent sur les étapes de recherche de motifs structuraux et fonctionnels par *MC-Motifs*. La dernière section décrit comment le programme recherche des motifs dans une structure secondaire contentant des pseudonœuds.

3.1 Prétraitement de la chaîne de points et parenthèses

MC-Motifs vérifie d'abord que la chaîne de points et parenthèses est balancée, c'est-à-dire que le nombre de parenthèses ouvrantes « (» équivaut au nombre de parenthèses fermantes «) ». Par la suite, il divise la structure secondaire en tiges et tiges-boucles. Chaque élément est ensuite traduit en motifs cycliques (*Figure 3.2*). Par exemple, dans la *Figure 3.2d*, la tige-boucle II est composée des motifs cycliques suivants : 2_2 + 2_2 + 2_2 + 2_2 + 2_2 + 3_2 + 2_2 + 2_2 + 2_2 + 2_2 + 2_2 + 2_2 + 2_2 + 3_2 + 2_2 + 2_2 + 4.

La prochaine étape consiste à appareiller les motifs cycliques de chaque élément structural à ceux des motifs de la base de données *bdMotifs*.

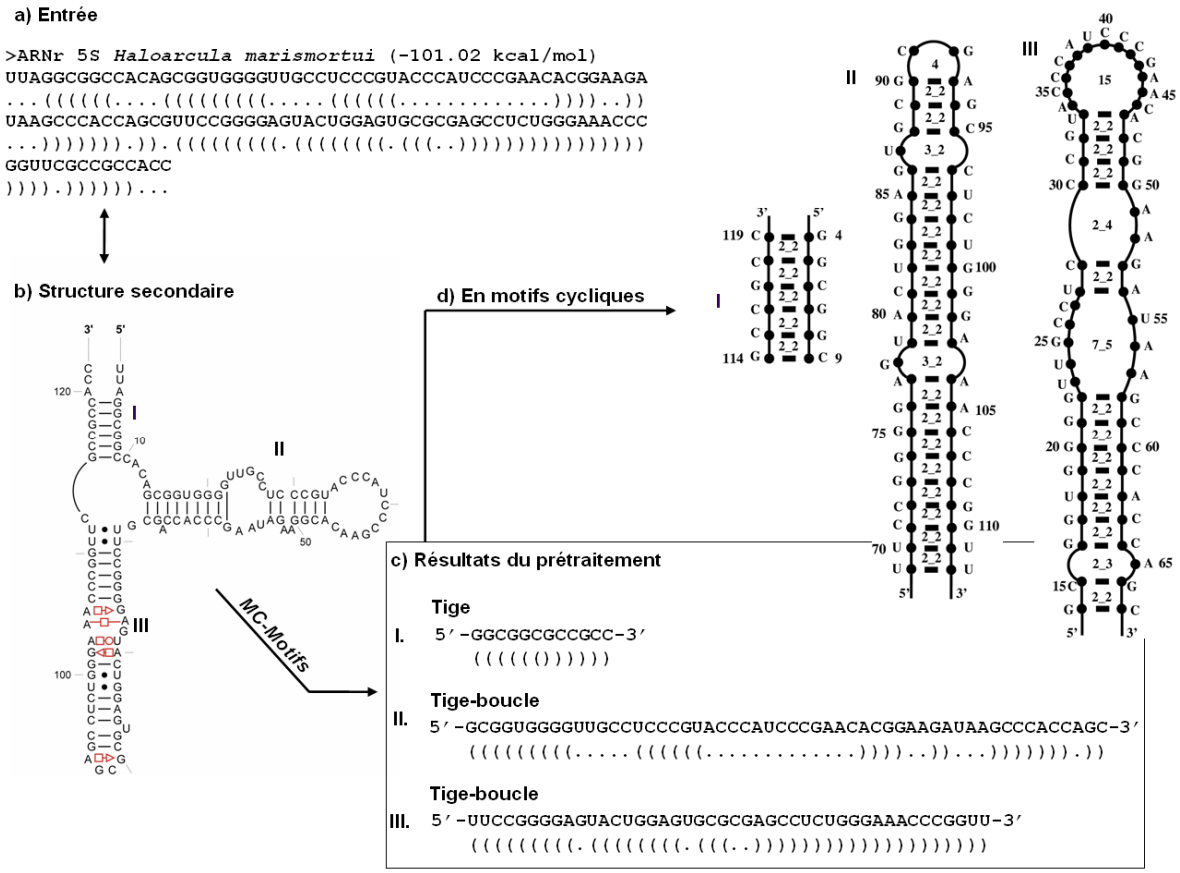


Figure 3.2 : Prétraitement de la chaîne de points et parenthèses. (a) *MC-Motifs* reçoit en entrée la structure secondaire de l'ARNr 5S d'*Haloarcula marismortui* prédite par *MC-Fold*. **(b et c)** *MC-Motifs* divise la structure secondaire en une tige (**I**) et deux tiges-boucles (**II** et **III**). **(d)** Chaque élément structural est traduit en motifs cycliques sous forme de notation algébrique.

3.2 Algorithme de recherche de motifs structuraux et fonctionnels

L'algorithme utilisé cherche un ensemble fini de mots (dans ce cas-ci, un motif structural défini en un ensemble de motifs cycliques) dans une phrase donnée (chaîne de motifs cycliques d'une tige ou tige-boucle). Je reprends l'exemple de la tige-boucle II provenant de l'ARNr 5S d'*Haloarcula marismortui* (*Figure 3.3a*). *MC-Motifs* considère la

chaîne de motifs cycliques de cette tige-boucle comme un tableau dynamique[6] où chaque élément du tableau contient un code d'identification du motif cyclique (*Figure 3.3.b*), et chaque code d'identification réfère à des informations concernant le motif cyclique. C'est-à-dire les nucléotides composant le motif et leurs positions dans la structure (*Figure 3.3c*).

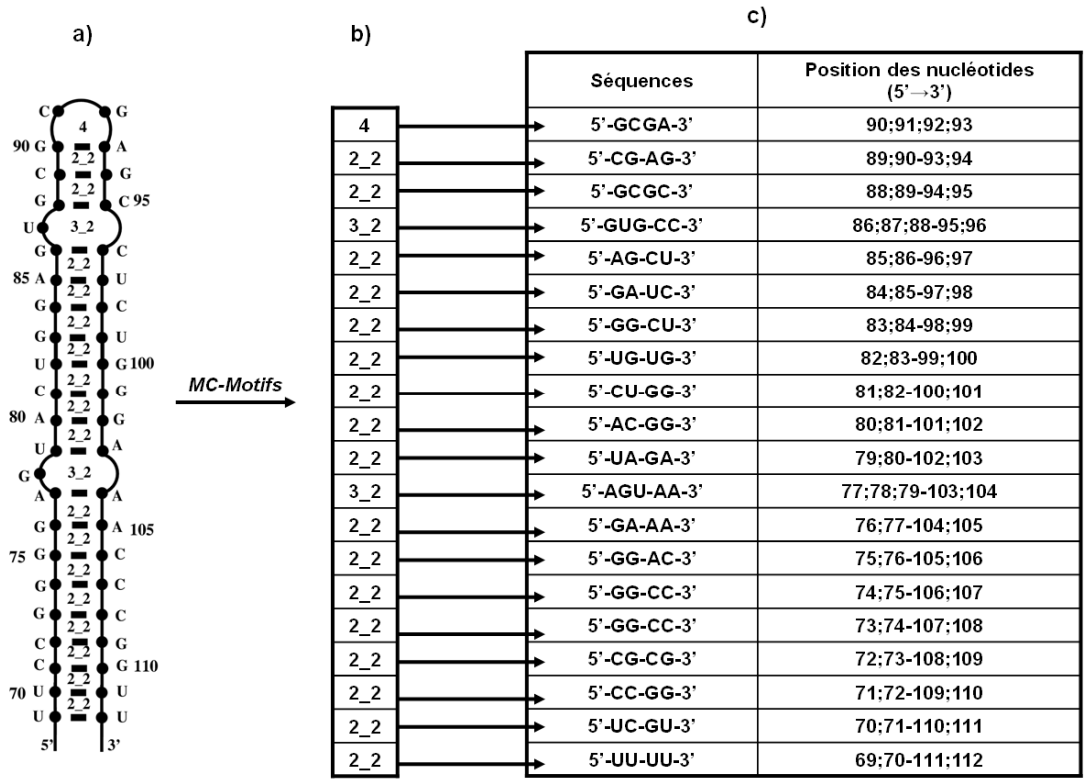


Figure 3.3 : Traitement de la chaîne de motifs cycliques par *MC-Motifs*. (a) Les motifs cycliques qui composent la structure secondaire de la tige-boucle II de l'ARNr 5S d'*Haloarcula marismortui*. **(b)** *MC-Motifs* considère la chaîne de motifs cycliques comme un tableau contenant les codes d'identifications des motifs cycliques. **(c)** Chaque code d'identification réfère aux informations concernant le motif cyclique : les nucléotides composant le motif et leurs positions dans la structure.

Les Figures *3.4* et 3.5 montrent l'algorithme de *MC-Motifs*.

[6] Un tableau dynamique est un tableau dont la taille peut grandir ou rétrécir.

Selon la chaîne de motifs cycliques définissant un motif fonctionnel dans *bdMotifs*, *MC-Motifs* cherche si celle-ci fait partie de la chaîne de motifs cycliques en parcourant le tableau (*Figures 3.4a* et *3.4b*). Si le programme trouve une sous-chaîne, il vérifie si la séquence correspondant à cette sous-chaîne équivaut à celle du motif défini dans *bdMotifs* (*Figure 3.4c*). Si la séquence de la sous-chaîne équivaut à celle du motif, *MC-Motifs* a détecté la présence possible du motif et il poursuit sa recherche (*Figures 3.4c* et *3.4d*).

À la fin du parcours de tableau, *MC-Motifs* passe à la définition du motif suivant dans *bdMotifs*. Le programme cherche la présence de ce motif dans la chaîne de motifs cycliques (*Figure 3.5*).

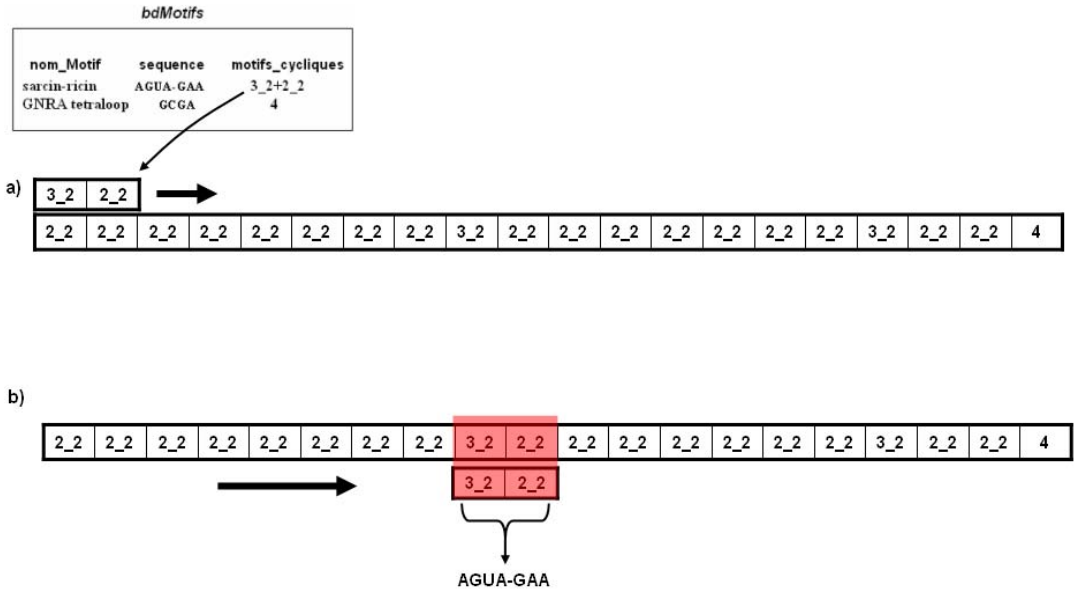


Figure 3.4a-b : Algorithme de recherche de sarcine-ricine par *MC-Motifs*. (a) *MC-Motifs* débute la recherche avec le motif sarcine-ricine dont sa structure est définie en termes de motifs cycliques « 3_2 + 2_2 ». **(b)** *MC-Motifs* parcoure le tableau afin de trouver la sous-chaîne « 3_2 + 2_2 » dans la chaîne de motifs cycliques.

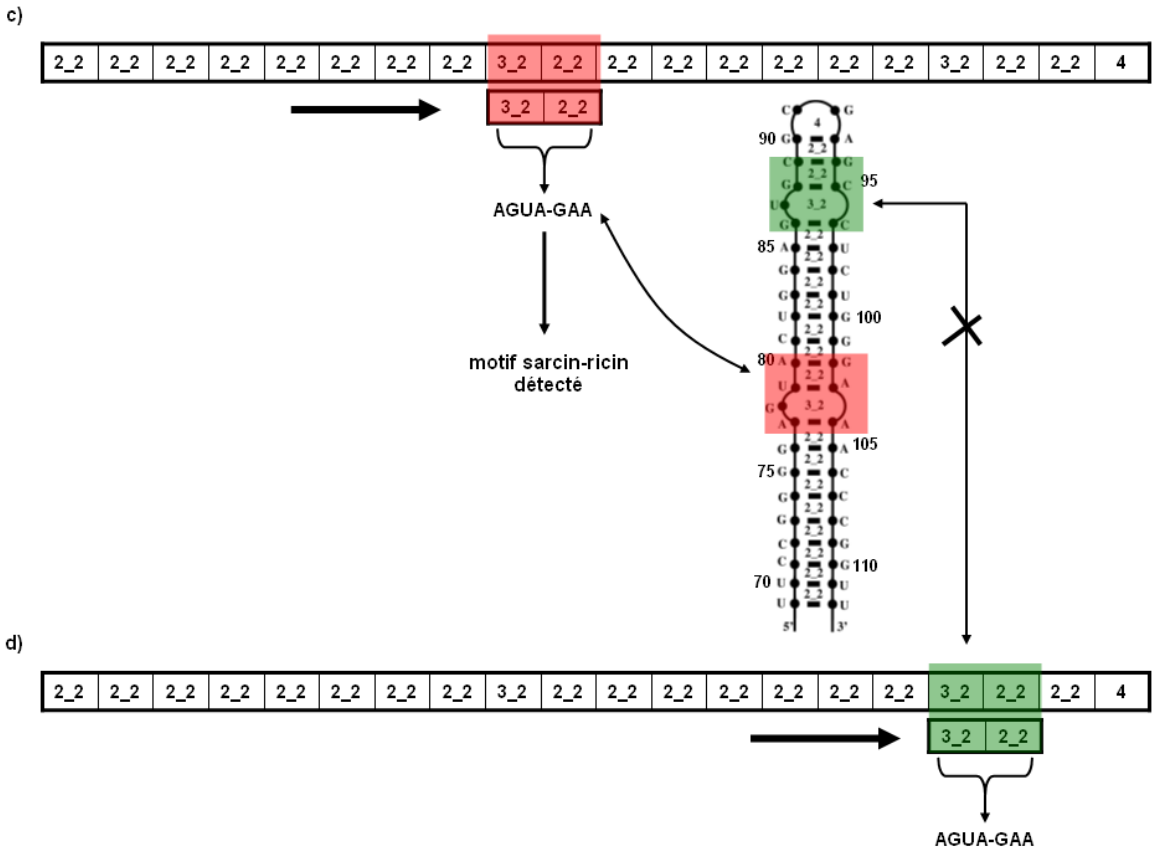


Figure 3.4c-d : Algorithme de recherche de sarcine-ricine par *MC-Motifs*. **(c)** Le programme trouve une sous-chaîne « 3_2 + 2_2 ». Dans *bdMotifs*, « 3_2 + 2_2 » correspond à la séquence 5'-AGUA-GAA-3'. Dans la chaîne de motifs cycliques, les motifs cycliques « 3_2 + 2_2 » correspond également à la séquence 5'-AGUA-GAA-3'. *MC-Motifs* a détecté la présence possible d'un motif sarcine-ricine. **(d)** *MC-Motifs* poursuit la recherche de d'autres sous-chaînes « 3_2 + 2_2 ». Dans l'exemple ci-dessous, le programme détecte une seconde sous-chaîne « 3_2+2_2 ». Elle représente 5'-GUGC-GCC-3' et ne correspond pas à la séquence de sarcine-ricine 5'-AGUA-GAA-3'.

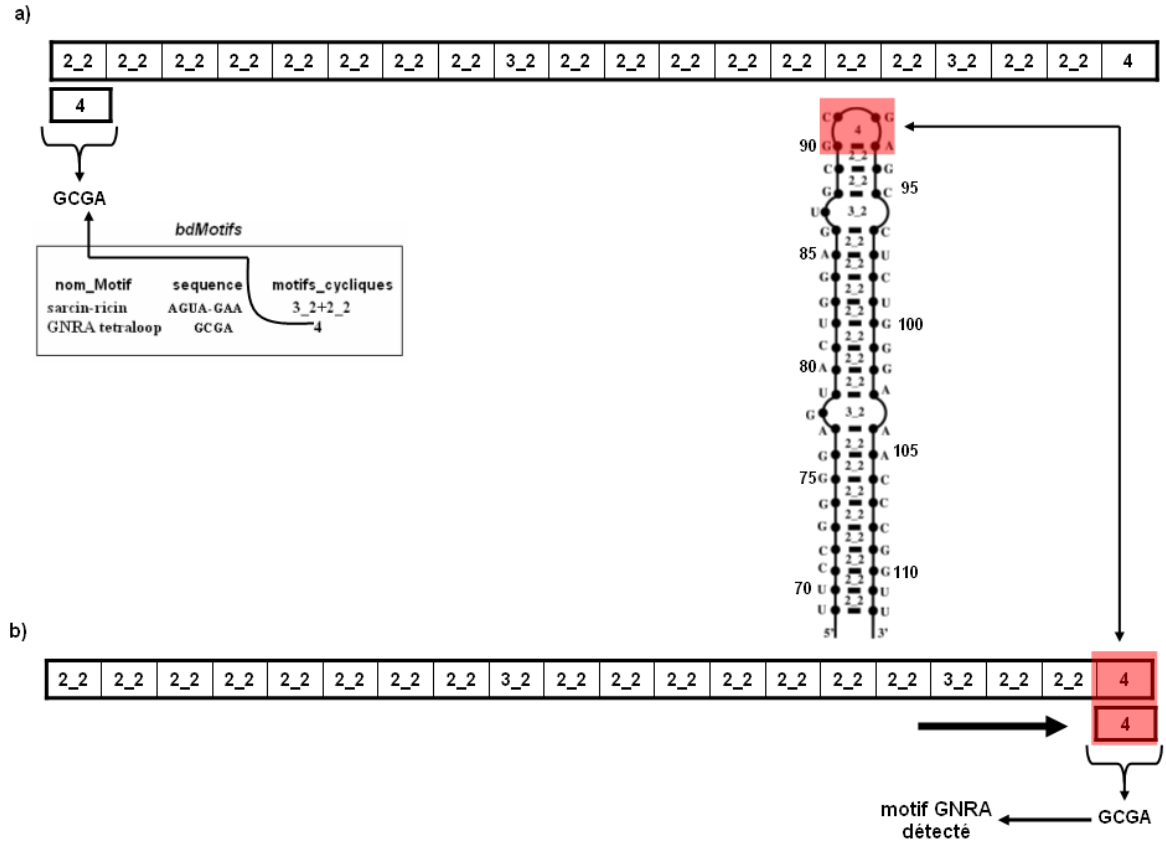


Figure 3.5 : Algorithme de recherche de la tétraboucle GNRA *MC-Motifs*. (a) *MC-Motifs* débute la recherche avec le motif GNRA composé d'un seul motif cyclique « 4 ». **(b)** Le programme trouve une sous-chaîne « 4 » dans la chaîne de motifs cycliques. Dans *bdMotifs*, le motif cyclique « 4 » correspond à la séquence 5'-GCGA-3'. Cette séquence correspond également au motif cyclique « 4 » dans la chaîne de motifs cycliques. *MC-Motifs* a alors détecte la présence possible d'une tétraboucle GNRA.

Ainsi, selon les figures précédentes, *MC-Motifs* a détecté la présence des motifs sarcine-ricine et GNRA dans la tige-boucle II, en assignant les motifs cycliques « 3_2+2_2 » à la séquence 5'-AGUA-GAA -3' et le motif cyclique « 4 » à la séquence 5'-GCGA-3' (*Figure 3.6*).

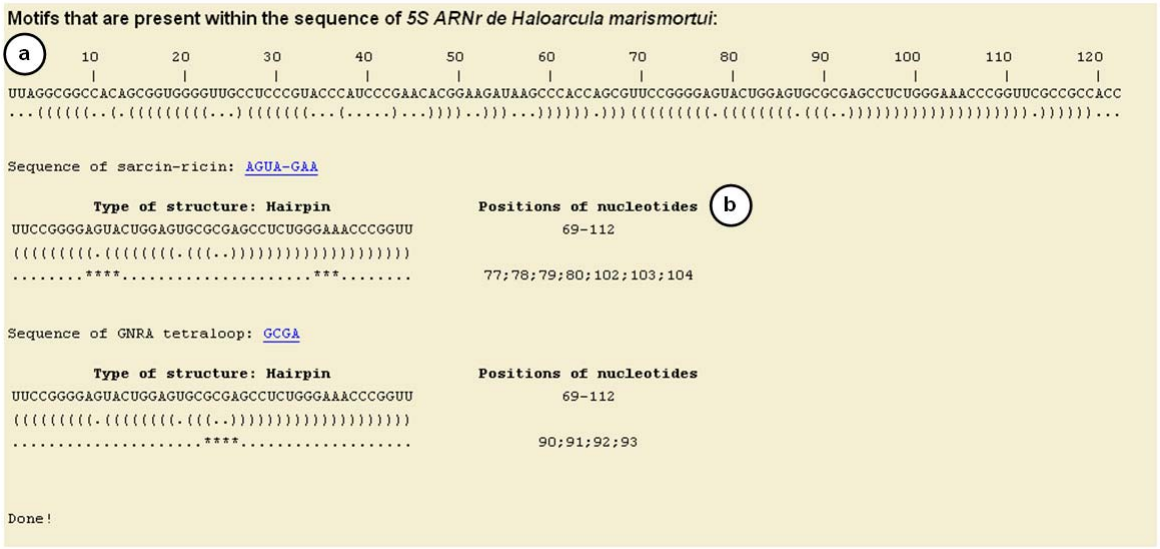


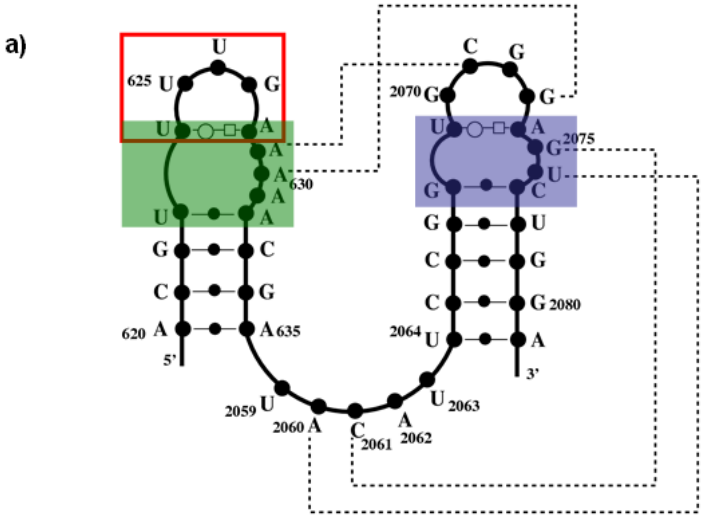
Figure 3.6 : Motifs sarcine-ricine et GNRA trouvés dans la tige-boucle II par *MC-Motifs*. La figure affiche l'interface graphique des résultats de recherche par *MC-Motifs*. **(a)** Le programme affiche la séquence ainsi que sa structure en chaîne de points et parenthèses. **(b)** Dans cette section, *MC-Motifs* affiche les positions nucléotidiques (les positions sont représentées par des étoiles) de chaque motif détecté dans la séquence, ainsi que l'endroit (69-112) où ces motifs se trouvent. Cet endroit correspond à la tige-boucle II.

Le temps d'exécution de l'algorithme de recherche est de O(n) où n correspond au nombre de motifs cycliques dans la chaîne.

3.3 Les pseudonœuds

MC-Motifs recherche également des motifs faisant partie des pseudonœuds ou d'autres éléments structuraux complexes. Ces éléments structuraux ont des liaisons tertiaires qui sont représentées par des paires de crochets, « [» et «] », dans la chaîne de points et parenthèses. La Figure *3.7* est un exemple d'une structure d'ARN où un motif « UA_handle » de type III forme des liaisons tertiaires avec des nucléotides d'une tige-boucle et un second motif « UA_handle » de type II forme des liaisons tertiaires avec un brin d'ARN. Cette structure se nomme le domaine « T-loop PK » (Jaeger et al., 2008), car il est composé d'un motif « UA_handle » de type II faisant partie d'un pseudonoeud (PK),

d'une boucle T (T-loop) et d'autres sous-motifs. Le domaine provient des ARNr 23S des bactéries et archaebactéries. Dans la figure ci-dessous, il est évident qu'il existe des éléments structuraux entre A635 et U2059. J'ai omis ces éléments afin de mieux illustrer la manière dont *MC-Motifs* cherche les motifs dans les pseudonœuds.

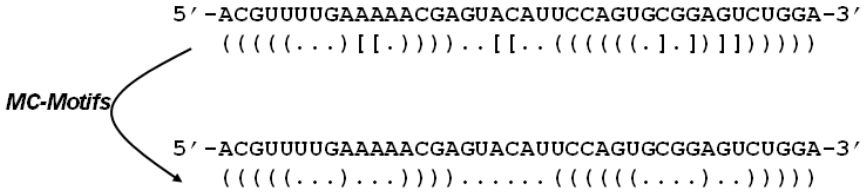


b) 5′-ACGUUUUGAAAAACGAGUACAUUCCAGUGCGGAGUCUGGA-3′
((((((...)[[.)))))..[[..((((((.].])]]))))))

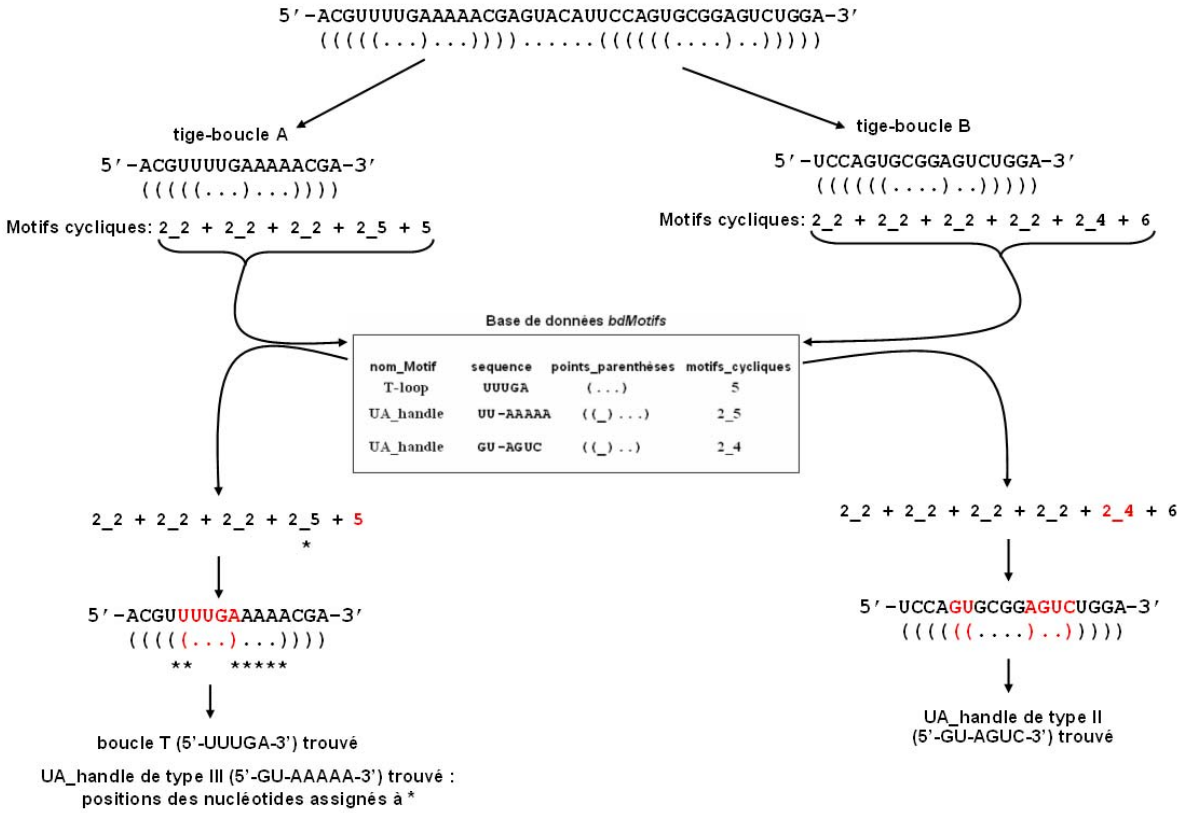
Figure 3.7 : Domaine « T-loop PK ». **(a)** Ce domaine provient de l'ARNr 23S de *Haloarcula marismortui*. Il est composé d'une boucle T (motif encadré en rouge), d'un motif « UA_handle » de type III (vert) formant des liaisons tertiaires (lignes pointillées) avec C2070 et G2073 de la tige-boucle, et d'un motif « UA_handle » de type II (bleu) formant des liaisons tertiaires avec A2060 et C2061 du brin d'ARN. **(b)** La chaîne de points et parenthèses représente le domaine « T-loop PK » de la *Figure 3.7a*. Les paires de crochets, « [» et «] », représentent les liaisons tertiaires.

La présence des crochets dans une chaîne de points et parenthèses cache l'existence possible d'un motif. J'ai donc développé une stratégie afin que *MC-Motifs* traite ces types de chaîne contenant les paires de crochets. La chaîne de points et parenthèses de la *Figure 3.7b* est choisie comme exemple :

1) Avant de débuter la recherche de motifs, *MC-Motifs* converti les paires de crochets en points. Cela permet au programme de chercher l'existence d'un motif en absence des crochets :



2) Par la suite, *MC-Motifs* effectue l'algorithme de recherche. Dans le schéma ci-dessous, *MC-Motifs* divise la structure deux tiges-boucles (**A** et **B**). Mon programme a trouvé une boucle T (5'-UUUGA-3') et un motif « UA_handle » de type III (5'-UU-AAAAA-3') dans la tige-boucle **A**. Dans la tige-boucle **B**, *MC-Motifs* a détecté un motif « UA_handle » de type II (5'-GU-AGUC-3'). Ces motifs ont été identifiés car leurs données les concernant se trouvent dans *bdMotifs* :



3.4 Conclusion

J'ai établi une base de données, *bdMotifs*, qui contient les séquences des motifs fonctionnels et leurs représentations structurales en termes de motifs cycliques. Comme on peut définir une structure secondaire par une chaîne de motifs cycliques, j'ai réussi à développer un algorithme de recherche, *MC-Motifs*, qui, à partir de *bdMotifs*, peut détecter des motifs fonctionnels (représentés en sous-chaînes de motifs cycliques) dans une structure secondaire d'ARN (représentée en chaîne de motifs cycliques).

Au prochain chapitre, je valide mon algorithme de recherche sur des structures d'ARN connues et je l'utilise pour prédire des motifs fonctionnels dans les riborégulateurs.

CHAPITRE 4

VALIDATION ET UTILISATION DE *MC-Motifs*

4.1 Validation de *MC-Motifs*

Pour valider *MC-Motifs*, j'ai cherché les motifs définis au chapitre précédent dans la structure secondaire du domaine I de l'ARNr 16S de *Thermus thermophilus*. Cette structure provient du laboratoire de Gutell (www.rna.ccbb.utexas.edu/SAE/2A/xtal_Info/). Elle a été déterminée par l'analyse des structures cristallines et l'analyse de covariances des séquences d'ARN ribosomales. Le domaine de 533 nucléotides contient des tétraboucles GNRA, un motif C et un kink-turn. J'ai soumis à *MC-Motifs* la séquence FASTA du domaine I et sa structure secondaire en chaîne de points et de parenthèses. Mon programme a retrouvé tous les motifs (*Figure 4.1*). Il a même détecté une boucle T dans l'hélice 13 du domaine I, un motif UA_handle de type I dans l'hélice 11 et un motif UA_handle de type II dans l'hélice 18.

Par la suite, j'ai choisi la séquence de l'ARNr 5S d'*Haloarcula marismortui* provenant également du laboratoire de Gutell. Il y a trois motifs dans cet ARNr : le motif GNRA, le motif sarcine-ricine et un motif nommé « lonepair triloop » (Lee et al., 2003; Lisi & Major, 2007). J'ai soumis à *MC-Motifs*, la séquence FASTA de l'ARNr 5S et sa structure secondaire en chaîne de points et de parenthèses. Les motifs GNRA et sarcine-ricine ont été détectés, ainsi que le motif « lonepair triloop », car j'ai ajouté les informations concernant ce motif dans *bdMotifs* (*Figure 4.2*).

a)

```
> 16s_seq(26:558; domaine I Thermus thermophilus)
AGGGUGAACGCUGGCGGCGUGCCUAAGACAUGCAAGUCGUGCGGGCCGCGGGGUUUUACUCCGUGGUCAGCGGCGGACGGGUGAGUAACGCGUGGGUACCUACCC
(((((.((((((.(((((((((.....(((.(((.(((.((((..(((((((((((((((...)))))))))))))))))))....(((.......(((((((.(((..((
GGAAGAGGGGGACAACCCGGGGAAACUCGGGCUAAUCCCCCAUGUGGACCCGCGUCCAAAGGGCUUUGCCCGCUUCCGGAUGGGCCCGCGUCCCAUCAGCUAGUU
((((((((((((((((((((((..))))))))))).)))))((..(((((((.)))))))).((((...))))))))))).))))))))))).((((..((((..((
GGUGGGGUAAUGGCCCACCAAGGCGACGACGGGUAGCCGGUCUGAGAGGAUGGCCGGCCACAGGGGCACUGAGACACGGGCCCCACUCCUACGGGAGGCAGCAGU
((((((.......))))))))))...)).)))..((((((((((..)))))...)))))))).((((((((...).)..))))))((((((..)))))).)))))
UAGGAAUCUUCCGCAAUGGGCGCAAGCCUGACGGAGCGACGCCGCUUGGAGGAAGAAGCCCUUCGGGGUGUAAACUCCUGAACCCGGGACGAAACCCCCGACGAG
).)(..)((.(((...((((((..))))))..)))))))).))))))..((((((....((((((..))))))...)))))).((((((((((((.((((((((...))
GGGACUGACGGUACCGGGGUAAUAGCGCCGGCCAACUCCGUGCCAGCAGCCGCGGUAAUACGGAGGGCGCGAGCGUUACCCGG
))))))..)))..)))))))))).).)..((((((.....((((((...(((....)))...)..))))))))))).))))))))))).)
```

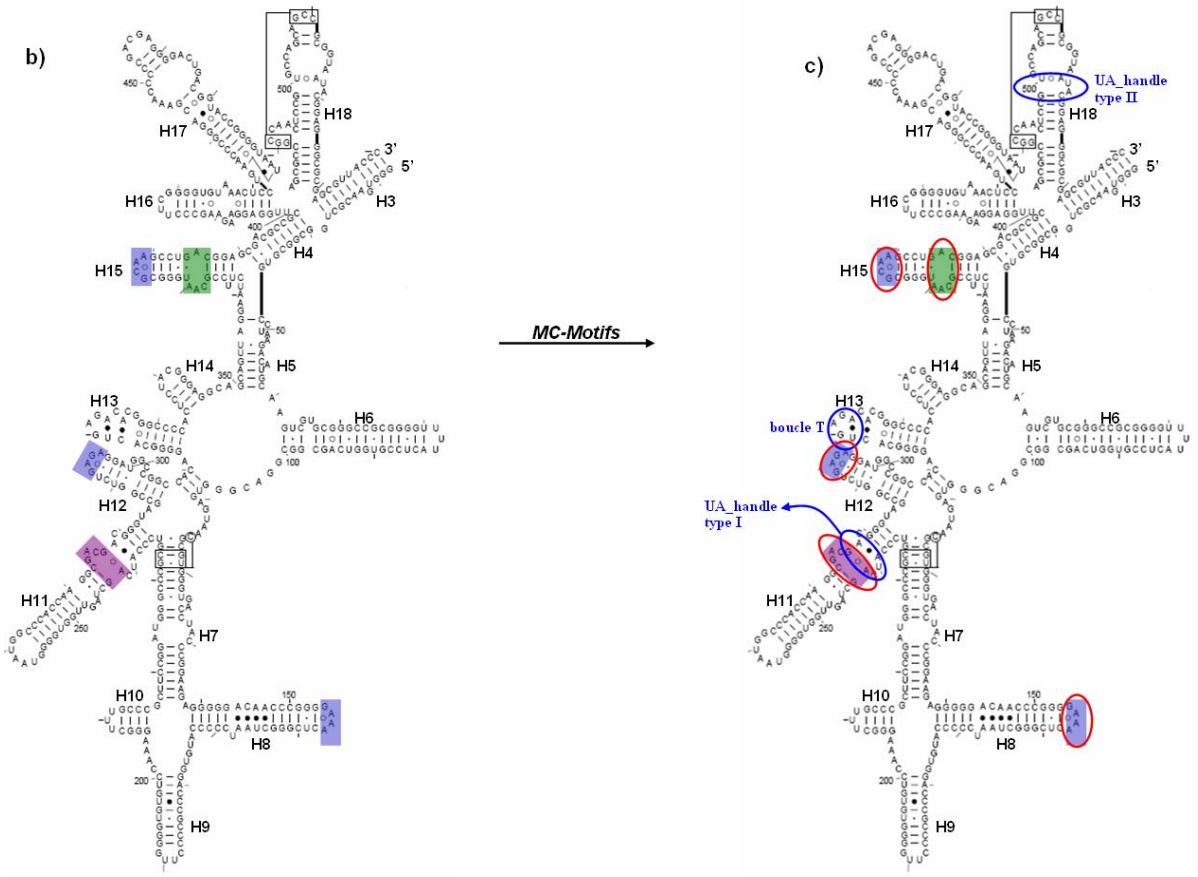


Figure 4.1 : Motifs trouvés dans le domaine I de l'ARNr 16S de *Thermus thermophilus* par *MC-Motifs*. (a) *MC-Motifs* reçoit en entrée la séquence de l'ARNr 16S de *Thermus thermophilus* et sa structure secondaire en chaîne de points et de parenthèses. **(b)** Structure secondaire du domaine I. Le symbole **H** suivi d'un chiffre signifie le numéro d'une hélice. Les nucléotides encadrés en bleu forment le motif GNRA, ceux encadrés en vert le motif C et ceux encadrés en mauve le kink-turn. **(c)** *MC-Motifs* a trouvé ces motifs (encerclés en rouge) indiqués par l'équipe de Gutell. Les motifs encerclés en bleu correspondent à la boucle T et les motifs UA_handle. Ils ont été détectés par *MC-Motifs* seulement. L'équipe de Gutell n'a pas identifié ces motifs.

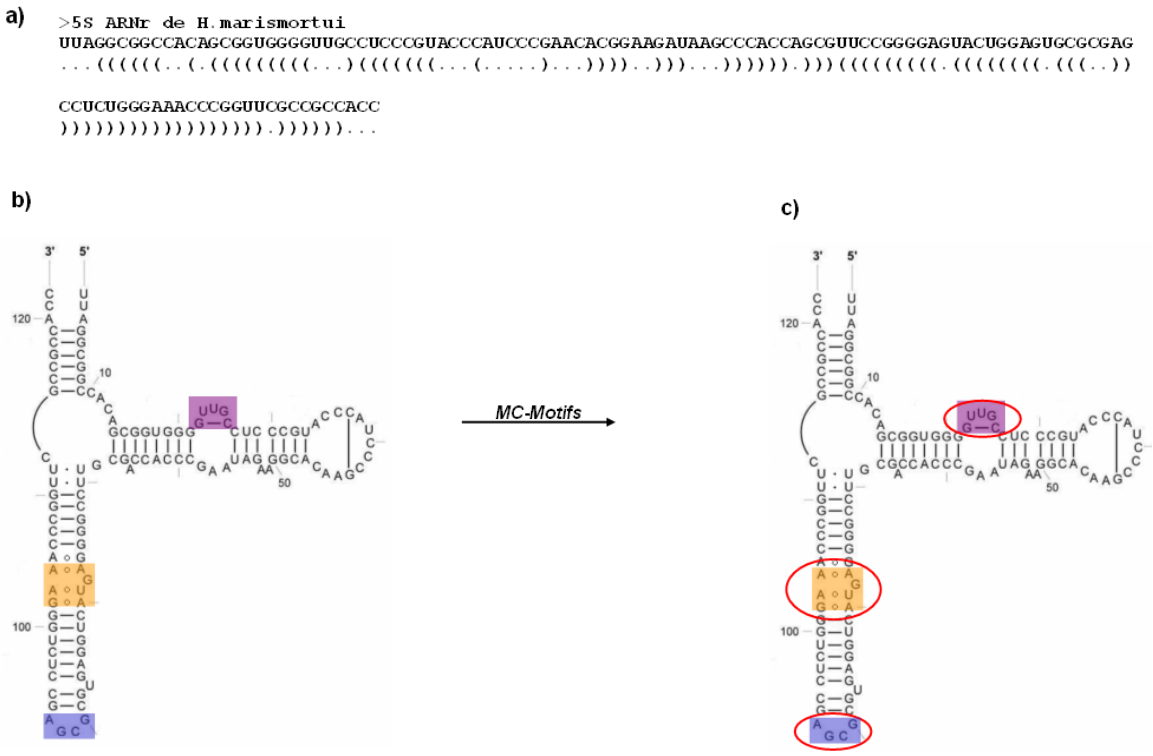


Figure 4.2 : Motifs trouvés dans l'ARNr 5S d'*Haloarcula marismortui* par *MC-Motifs*. (a) *MC-Motifs* reçoit en entrée la séquence d'ARNr 5S de *Haloarcula marismortui* et sa structure secondaire en chaîne de points et de parenthèses. **(b)** Structure secondaire de l'ARNr 5S contenant les trois motifs: le motif GNRA (bleu) et le motif sarcine-ricine (jaune) ainsi que le motif « lonepair triloop » (mauve). **(c)** *MC-Motifs* a détecté les trois motifs (encerclés en rouge).

Ces résultats montrent bien le bon fonctionnement de mon programme. Dans la section suivante, j'utilise *MC-Motifs* pour prédire des motifs fonctionnels dans une famille d'ARN.

4.2 Prédiction de motifs dans une famille d'ARN

J'ai appliqué *MC-Motifs* pour chercher, cette fois-ci, les motifs dans des éléments de séquences d'ARN de riborégulateurs de la flavine mononucléotide (FMN), car récemment la structure consensus du noyau de cet ARN a été prédite (Barrick & Breaker, 2007; Jaeger et al., 2008). Dans la structure, les boucles T ont été identifiées dans les régions P2, P3 et P5 par Barrick et Breaker (*Figure 4.3*). L'équipe de Jaeger a identifié des

motifs « UA_handle » de type II (UA_h_II) reliés aux boucles T dans les régions P2, P3 et P5. Le UA_h_II du P2 interagit (I.t) avec des nucléotides du brin d'ARN de la région P5. Le « UA_handle » est un motif instable, il est formé lorsque ses nucléotides créent des interactions avec d'autres motifs tels que les boucles T et les tétraboucles GNRA (Jaeger et al., 2008). Dans la Figure *4.3*, l'équipe de Jaeger a alors prédit de nouvelles interactions tertiaires (I.t *) entre le motif UA_h_II du P5 et les nucléotides d'une boucle T ou d'autre motif du P3. L'équipe n'est pas certaine de la présence d'interactions tertiaires (I.t ?) entre la boucle T du P2 et les nucléotides de la boucle du P6.

J'ai sélectionné, aléatoirement, deux séquences de riborégulateur FMN provenant de l'alignement de séquences : l'une de l'espèce *Vibrio cholerae* et l'autre de *Mesorhizobium loti* (Barrick & Breaker, 2007). J'ai voulu vérifié si *MC-Motifs* peut détecter les boucles T et les motifs UA_h_II dans ces séquences aux endroits indiqués de la *Figure 4.3*, s'il y a des motifs autres que les boucles T et les UA_h_II, et si, effectivement, il peut y avoir des interactions tertiaires entre la boucle T du P2 et les nucléotides de la boucle du P6.

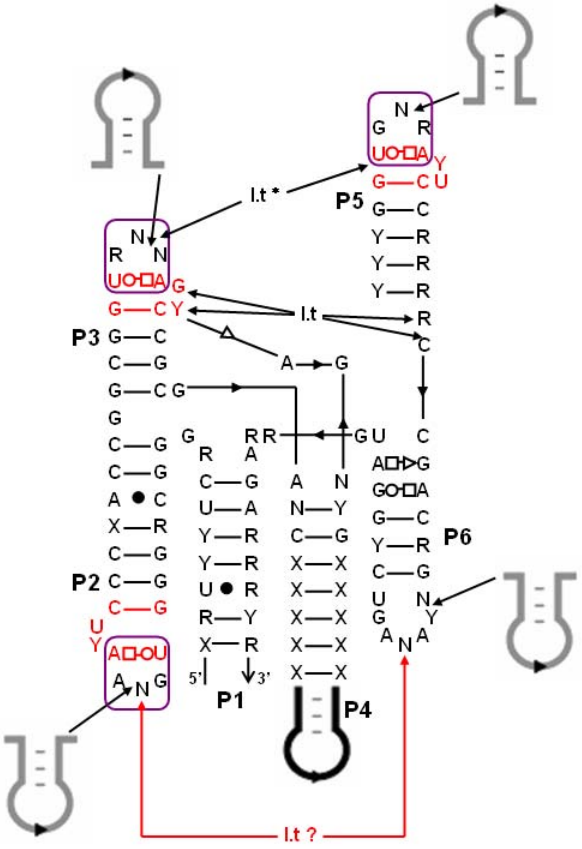


Figure 4.3: Prédiction du noyau structural du riborégulateur FMN. Les boucles dans un cadre mauve sont des boucles T. Les nucléotides formant les motifs UA_h_II sont en rouges. Le code « Y » représente les bases C ou U, le code « R » les bases A ou G, le code « N » les quatre bases (A, C, U, G), et le code « X » représente n'importe quelle base (A, C, U, G) appariée à une autre via l'interaction W/W *cis*.
La tige-boucle en noire signifie qu'elle est toujours présente. Les tiges-boucles en gris signifient qu'elles sont parfois insérées dans la séquence.

L'équipe de Jaeger prédit de nouvelles interactions tertiaires (I.t *) entre le motif UA_h_II du P5 et les nucléotides de la boucle du P3. Il peut y avoir des interactions tertiaires (I.t ?) entre la boucle T du P2 et les nucléotides de la boucle du P6. Remarquez que le UA_h_II du P3 forme des liaisons tertiaires (lignes pointillées) avec des nucléotides dans la région du P5.

Image modifiée de Jaeger et al., *N.A.R*, 1, 1-16 (2008)

4.2.1 Riborégulateur FMN de *Vibrio cholerae*

J'ai soumis la séquence FASTA de *Vibrio cholerae* et sa structure secondaire, en chaîne de points et de parenthèses, à *MC-Motifs* (*Figure 4.4*). *MC-Motifs* a trouvé deux boucles T (5'-CGAAA-3' et 5'-UGAGA-3') dans les régions P2 et P5. Les mêmes régions que celles affichées dans la *Figure 4.3*. Le programme a également détecté des motifs UA_h_II (5'-GC-AUUC-3' et 5'-GU-AAGC-3') reliés aux boucles T dans les régions P2 et P5 respectivement, et la présence d'un motif UA_h_II (5'-GU-AGCC-3') dans la région P3 comme indiqué dans la *Figure 4.3*.

Figure 4.4 : Motifs structuraux détectés par *MC-Motifs* dans l'ARN du riborégulateur FMN de *Vibrio cholerae*. Les bases formant la boucle T sont en jaune, les bases formant le motif UA_h de type II (UA_h_II) sont assignées aux symboles *.

Dans la séquence du *Vibrio cholerae*, j'ai aperçu qu'il y a peu d'appariements dans les régions P3, P4 et P6. Les bases appartenant à ces régions sont en rouge (*Figure 4.5*).

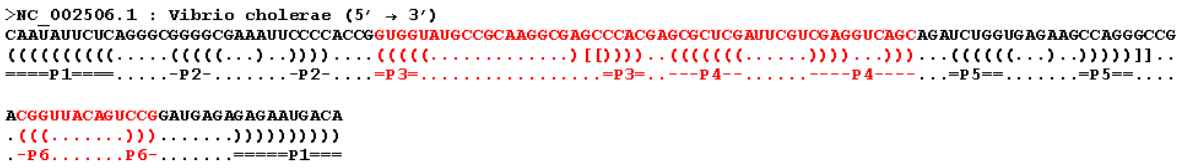


Figure 4.5 : Régions ayant peu d'appariements dans l'ARN de *Vibrio cholerae*.

Alors, j'ai soumis ces régions à *MC-Fold* pour prédire leurs structures secondaires. Rappelons que *MC-Fold* génère, à partir d'une séquence donnée, une liste triée de structures sous-optimales possibles (des structures les plus stables aux structures les plus faibles). J'ai limité à cinquante le nombre de structures secondaires prédites par *MC-Fold* pour les régions P3, P4 et P6. *MC-Fold* a prédit quatre structures secondaires pour les régions P3 et P4 (*Figure 4.6a*), et cinq structures pour la région P6 (*Figure 4.6b*).

a) Avant *MC-Fold* (régions P3 et P4):

```
>Vibrio cholerae
GUGGUAUGCCGCAAGGCGAGCCCACGAGCGCUCGAUUCGUCGAGGUCAGC
(((((.............)[])))..(((((((.....))))))...)))
=P3=.................=P3=..---P4--......----P4----
```

Après *MC-Fold* (régions P3 et P4):

```
>Vibrio cholerae
GUGGUAUGCCGCAAGGCGAGCCCACGAGCGCUCGAUUCGUCGAGGUCAGC
(((((.((((((.)))))).))))).(((((((((..)))))))...))) -43.75 kcal/mol
(((((.((((....))))))..))))..(((((((((..)))))))...))) -42.61 kcal/mol
(((((.((((((.)))))).))))).(((((((((....)))))))...))) -42.47 kcal/mol
(((((.((((....))))))..))))..(((((((((....)))))))...))) -41.33 kcal/mol
```

b) Avant *MC-Fold* (région P6) :

```
>Vibrio cholerae
CGGUUACAGUCCG
(((.......)))
-P6.......P6-
```

Après *MC-Fold* (région P6) :

```
>Vibrio cholerae
CGGUUACAGUCCG
((((((...))))))  -9.74 kcal/mol
((((((..)).)))  -8.55 kcal/mol
```

```
>Vibrio cholerae
CGGUUACAGUCCG
((((.....))))  -8.25 kcal/mol
((((....).)))  -7.70 kcal/mol
(((.(....))))  -7.11 kcal/mol
```

Figure 4.6 : Structures secondaires des régions P3, P4 et P6 de *Vibrio cholerae* générées par *MC-Fold*. (a) Structures secondaires prédites des régions P3 et P4. **(b)** Structures secondaires prédites de la région P6.

J'ai traité les structures prédites des régions P3, P4 et P6 de *Vibrio cholerae* ainsi que leurs séquences à *MC-Motifs* (*Figure 4.7*).

a)

```
>Vibrio cholerae (régions P3 et P4)
GUGGUAUGCCGCAAGGCGAGCCCACGAGCGCUCGAUUCGUCGAGGUCAGC
(((((.((((..)))))..)))))..(((((((((..)))))))...))) -43.75 kcal/mol
...**............****.........................

GUGGUAUGCCGCAAGGCGAGCCCACGAGCGCUCGAUUCGUCGAGGUCAGC
(((((.((((....)))))..)))))..(((((((((..)))))))...))) -42.61 kcal/mol
...**............****.........................

GUGGUAUGCCGCAAGGCGAGCCCACGAGCGCUCGAUUCGUCGAGGUCAGC
(((((.((((..)))))..)))))..(((((((....)))))...))) -42.47 kcal/mol
...**............****.........................

GUGGUAUGCCGCAAGGCGAGCCCACGAGCGCUCGAUUCGUCGAGGUCAGC
(((((.((((....)))))..)))))..(((((((....)))))...))) -41.33 kcal/mol
...**............****.........................
```

UA_handle type II : positions des nucléotides assignés à *
GNRA : nucléotides en couleur bleue

b)

```
>Vibrio cholerae (région P6)
CGGUUACAGUCCG
((((((...))))))  -9.74 kcal/mol
.............

CGGUUACAGUCCG
((((((..)).)))  -8.55 kcal/mol
.............

CGGUUACAGUCCG
((((.....))))  -8.25 kcal/mol
.............

CGGUUACAGUCCG
((((....).)))  -7.70 kcal/mol
.............
```

```
CGGUUACAGUCCG
(((.(....)))  -7.11 kcal/mol
.............
```

Figure 4.7 : Motifs structuraux détectés par *MC-Motifs* dans les régions P3, P4 et P6 de *Vibrio cholerae*. **(a)** Les bases formant le motif GNRA sont en couleur bleu, les bases formant le motif (UA_h_II) sont assigné aux symboles *. Parmi les quatre structures secondaires prédites par *MC-Fold*, deux contient les motifs GNRA (5'-GCAA-3') et UA_h_II (5'-GU-AAGC-3') dans la région P3. **(b)** Aucun motif détecté parmi les cinq structures prédites de la région P6.

Dans la région P3, *MC-Motifs* a identifié un motif GNRA (5'-GCAA-3') dans deux structures secondaires prédites par *MC-Fold* (*Figure 4.7a*). Je m'attendais à une boucle T dans cette même région comme celle dans la *Figure 4.3*. Cependant, en examinant la figure, j'ai remarqué qu'une tige-boucle peut être insérée dans la région P3. Donc, au lieu d'une boucle T, la boucle de la région P3 est un motif GNRA. Comme dans la *Figure 4.3*, il peut y avoir des interactions tertiaires entre le motif GNRA et le motif UA_h_II (5'-GU-AAGC-3') du P5, car ces deux motifs forment des interactions tertiaires avec des éléments structuraux éloignés (*Figure 4.8*). Dans la région du P3, *MC-Motifs* a également détecté un sous-motif UA_h_II (5'-GU-AGCC-3') tel qu'affiché dans la *Figure 4.3*. Aucun motif n'a été détecté par le programme dans les structures secondaires prédites des régions P6 (*Figure 4.7b*). Cela ne signifie pas qu'il n'y a pas la présence possible de motifs dans cette région. La base de données *bdMotifs* ne contient que six motifs définis (sarcine-ricine, motif C, boucle T, kink-turn, tétraboucle GNRA et « UA_handle »), il pourrait y en avoir d'autres. Compte tenu du peu de données dans *bdMotifs*, je ne peux pas prédire la possibilité d'interactions tertiaires entre la boucle T du P2 et la boucle du P6.

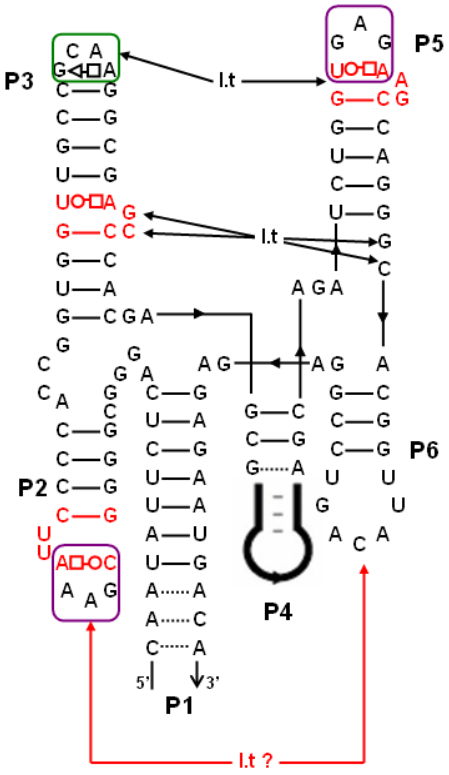


Figure 4.8 : Structure secondaire du noyau structural de l'ARN FMN de *Vibrio cholerae*. Cette image est une modification de la *Figure 4.3*. *MC-Motifs* a détecté les boucles T (nucléotides dans un cadre mauve) dans les régions P2 et P5, les motifs UA_h_II (nucléotides rouges) dans les régions P2, P3 et P5, ainsi qu'une tétraboucle GNRA (nucléotides dans un cadre vert) dans la région P3.

Les traits pointillés représentent les interactions non classiques, alors que les autres traits représentent des interactions classiques W/W *cis*.

Il peut y avoir des interactions tertiaires (I.t) entre le motif GNRA (5'-GCAA-3') du P3 et le motif UA_h_II (5'-GU-AAGC-3') du P5. Je ne peut pas prédire la présence possible d'interactions tertiaires entre la boucle T (5'-CGAAA-3') du P2 et la boucle du P6, car *MC-Motifs* n'a pas pu détecté des motifs.

La section suivante porte sur les résultats de recherche de motifs fonctionnels dans le riborégulateur FMN de *Mesorhizobium loti*.

4.2.2 Riborégulateur FMN de *Mesorhizobium loti*

La partie supérieure de la *Figure 4.9* affiche la séquence de l'espèce et sa structure secondaire en chaîne de points et de parenthèses. J'ai soumis la séquence et la chaîne de points et de parenthèses à *MC-Motifs*. Le programme a identifié les boucles T (5'-UGAA-3' et 5'-UGUGA-3') et les motifs UA_h_II (5'-GU-AGUC-3' et 5'-GU-AUUC-3') dans les régions P2 et P5 respectivement, telles qu'affichées dans la *Figure 4.3*. *MC-Motifs* a aussi détecté un motif UA_h_II (5'-GU-AGCC-3') dans la région P3.

Figure 4.9 : Motifs structuraux détectés par *MC-Motifs* dans l'ARN du riborégulateur FMN de *Mesorhizobium loti*. Les bases formant la boucle T sont en jaune et les bases formant le motif UA_h de type II (UA_h_II) sont assignées aux symboles *.

Comme pour la structure secondaire de *Vibrio cholerae*, il y avait peu d'appariements dans les régions P3, P4 et P6 de l'ARN de *Mesorhizobium loti* :

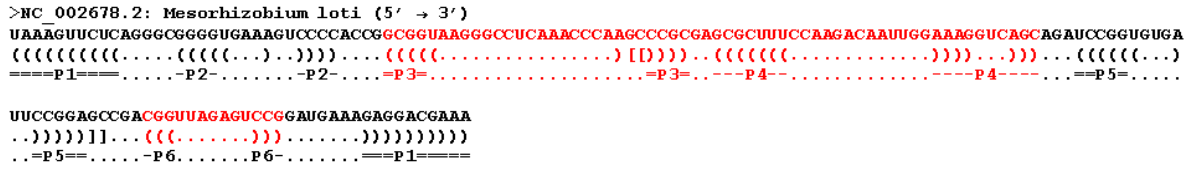


Figure 4.10 : Régions ayant peu d'appariements dans l'ARN de *Mesorhizobium loti*.

J'ai soumis ces régions à *MC-Fold* pour obtenir leurs structures secondaires. J'ai également limité à cinquante le nombre de structures secondaires prédites par *MC-Fold* pour les régions P3, P4 et P6. *MC-Fold* a prédit cinquante structures pour les régions P3 et

P4 (quelques structures prédites sont montrées à la *Figure 4.11a*), et six structures pour la région P6 (*Figure 4.11b*):

a) Avant *MC-Fold* (régions P3 et P4):

```
>NC_002678.2: Mesorhizobium loti
GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
(((((.................)[])))..(((((((.............))))...)))
=P3=.....................=P3=..---P4--.................----P4----
```

Après *MC-Fold* (régions P3 et P4) :

```
>Mesorhizobium loti MAFF303099
GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC (kcal/mol)
((((((((((....))))))).))))..((((((((((...)))))))))...)))  -49.86
(((((.((((((...))))))).))))..((((((((((...)))))))))...)))  -49.33
((((((((((..))))))))..))))..((((((((((...)))))))))...)))  -49.11
((((((((((....))))))).))))..((((((((((..).))))))))...)))  -48.81
(((((.(((((.(..))))))).))))..((((((((((...)))))))))...)))  -48.71
((((((((((.....))))))).))))..((((((((((...)))))))))...)))  -48.39
(((((.(((((.....))))))).))))..((((((((((...)))))))))...)))  -48.38
(((((.((((((...))))))).))))..((((((((((..).))))))))...)))  -48.28
((((((((((..))))))))..))))..((((((((((..).))))))))...)))  -48.05
(((((.((((......))))).))))..((((((((((...)))))))))...)))  -47.87
(((((.(((((.(..))))))).))))..((((((((((..).))))))))...)))  -47.65
((((((((((....))))))).))))..((((((((((.....))))))))...)))  -47.63
((((((((((.....))))))).))))..((((((((((..).))))))))...)))  -47.34
(((((.(((((.....))))))).))))..((((((((((..).))))))))...)))  -47.33
(((((.((((((...))))))).))))..((((((((((.....))))))))...)))  -47.11
((((((((((..))))))))..))))..((((((((((.....))))))))...)))  -46.88
(((((.((((((...))))))).))))..(((((((((.((....))))))))...)))  -44.61
((((((((((...).))))))).))))..((((((((((.....))))))))...)))  -44.52
(((((.(((((.....))))))).))))..((((((((((.(..)))))))))...)))  -44.51
((((((((((.....))))))).))))..((((((((((.(..)))))))))...)))  -44.51
(((((.((((((...))))))).))))..((((((((((.......))))))))...)))  -44.50
```

b) Avant *MC-Fold* (section P6) :

```
>Mesorhizobium loti
CGGUUAGAGUCCG
(((.......)))
-P6.......P6-
```

Après *MC-Fold* (section P6) :

```
>Mesorhizobium loti
CGGUUAGAGUCCG
(((((...)))))  -9.46 kcal/mol
```

```
>Mesorhizobium loti
CGGUUAGAGUCCG
(((((..)).)))   -8.77 kcal/mol
((((.....))))   -8.25 kcal/mol
((((....).)))   -7.93 kcal/mol
(((.......)))   -6.99 kcal/mol
(((.(....))))   -6.96 kcal/mol
```

Figure 4.11 : Structures secondaires des régions P3, P4 et P6 de *Mesorhizobium loti* générées par *MC-Fold*. (a) Exemple de quelques structures secondaires prédites des régions P3 et P4. **(b)** Structures secondaires prédites de la région P6.

J'ai traité les structures prédites des régions P3, P4 et P6 de *Mesorhizobium loti* ainsi que leurs séquences à *MC-Motifs* (*Figure 4.12*).

a)

```
>Mesorhizobium loti MAFF303099
GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC (kcal/mol)
(((((.(((((...)))))).))))..((((((((((...)))))))))...)))   -49.33
...**..............****..................................

GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
((((((((((....))))))).))))..((((((((((..).))))))))...)))   -48.81
...**..............****..................................

GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
(((((.(((((...)))))).))))..((((((((((..).))))))))...)))   -48.28
...**..............****..................................

GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
((((((((((..))))))))).))))..((((((((((..).))))))))...)))   -48.05
...**..............****..................................

GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
(((((.(((((.(..)))))))..))))..((((((((((..).))))))))...)))   -47.65
...**..............****..................................

GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
((((((((((.....))))))..))))..((((((((((..).))))))))...)))   -47.34
...**..............****..................................

GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
(((((.(((((.....))))))..))))..((((((((((..).))))))))...)))   -47.33
...**..............****..................................

GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
(((((.(((((...)))))).))))..((((((((((.....))))))))...)))   -47.11
...**..............****..................................
```

```
GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
(((((.((((.......)))))..))))..((((((((((..).))))))))...)))   -46.81
...**.................****.................................

GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
((((((((((...).))))))..))))..((((((((((..).))))))))...)))   -45.69
...**.................****.................................

GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
((((((((((.....).))))))..))))..((((((((((..).))))))))...)))   -45.52
...**.................****.................................

GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
(((((((((........)))))..))))..((((((((((..).))))))))...)))   -45.47
...**.................****.................................

GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
(((((.(((((...)))))).))))..((((((((((.(..).))))))))...)))   -45.46
...**.................****.................................

GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
(((((.(((((...)))))).))))..(((((((((.((....)))))))))...)))   -44.61
...**.................****.................................

GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
(((((.(((((...)))))).))))..((((((((((.......))))))))...)))   -44.50
...**.................****.................................

GCGGUAAGGGCCUCAAACCCAAGCCCGCGAGCGCUUUCCAAGACAAUUGGAAAGGUCAGC
(((((.(((((...)))))).))))..((((((((((....).))))))))...)))   -44.48
...**.................****.................................
```

Boucle T : nucléotides en couleur jaune
UA_handle type II : positions des nucléotides assignés à *
GNRA : nucléotides en couleur bleu

b)

```
>Mesorhizobium loti
CGGUUAGAGUCCG
((((..)).)))   -8.77 kcal/mol
............
```

GNRA : nucléotides en couleur bleu

Figure 4.12: Motifs structuraux détectés par *MC-Motifs* dans les régions P3, P4 et P6 de *Mesorhizobium loti*. Les bases formant le motif GNRA sont en couleur bleu, les bases formant la boucle T sont en jaune et les bases formant le motif (UA_h_II) sont assigné aux symboles *. **(a)** Parmi les cinquante structures secondaires prédites par *MC-Fold*, six contient les boucles T (5'-CUCAA-3') et les motifs UA_h_II (5'-GU-AGCC-3') dans la région P3. Il y a neuf structures qui

contiennent les motifs GNRA (5'-GACA-3') dans la région P4 et les motifs UA_h_II (5'-GU-AGCC-3') dans la région P3. Seulement une structure dont l'énergie est -48.28 kcal/mol contient la boucle T (5'-CUCAA-3') et le motif UA_h_II (5'-GU-AGCC-3') dans la région P3, et un motif GNRA (5'-GACA-3') dans la région P4. **(b)** Une structure parmi les six structures secondaires prédites par *MC-Fold* contient un motif GNRA (5'-UAGA-3') dans la région P6.

MC-Motifs a identifié une boucle T (5'-CUCAA-3') et un motif UA_h_II (5'-GU-AGCC-3') dans la région P3 (*Figures 4.12a* et *4.13*). Cependant, le motif UA_h_II est loin de la boucle T. Il n'est pas lié à la boucle telle qu'affichée dans la *Figure 4.3*. Cette boucle T appartient, alors, à une tige-boucle qui est insérée dans la région P3. Dans la région P4, *MC-Motifs* a détecté la présence d'un motif GNRA (5'-GACA-3'). Le programme a aussi détecté la présence possible d'un motif GNRA (5'-UAGA-3') dans la région P6. Donc, il pourrait y avoir des interactions tertiaires entre la boucle T du P2 et le motif GNRA du P6, comme soupçonné par l'équipe de Jaeger (Jaeger et al., 2008). Je prédis aussi des interactions tertiaires entre la boucle T du P2 et le motif GNRA du P4.

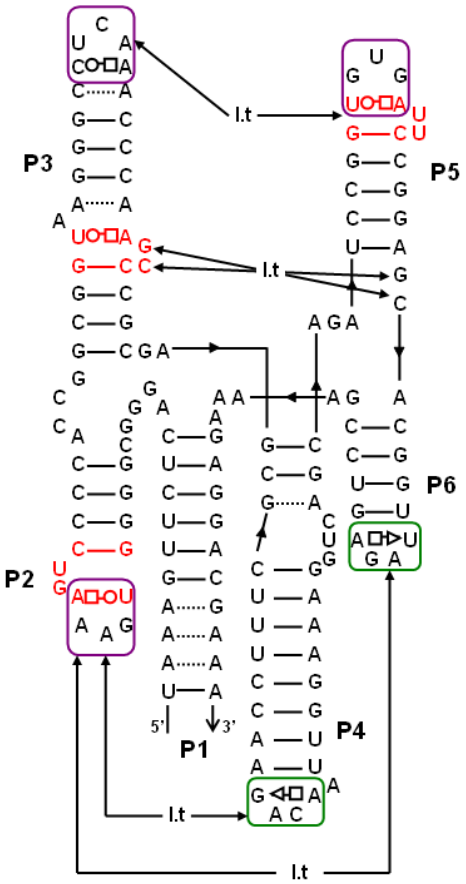


Figure 4.13 : Structure secondaire du noyau structural de l'ARN FMN de *Mesorhizobium loti*. Cette image est une modification de la *Figure 4.3*. *MC-Motifs* a détecté les boucles T (nucléotides dans un cadre mauve) dans les régions P2, P3 et P5, les motifs UA_h_II (nucléotides rouges) dans ces mêmes régions, ainsi que deux tétraboucles GNRA (nucléotides dans un cadre vert) dans les régions P3 et P4.

Les traits pointillés représentent les interactions non classiques, alors que les autres traits représentent des interactions classiques W/W *cis*.

Il peut y avoir des interactions tertiaires (I.t) entre la boucle T (5'-CUCAA-3') du P3 et le motif UA_h_II (5'-GU-AUUC-3') du P5, des interactions tertiaires entre la boucle T (5'-UGAAA-3') du P2 et le motif GNRA (5'-GACA-3') du P4 et le motif GNRA (5'-UAGA-3') du P6.

4.3 Conclusion

En utilisant le formalisme de motifs cycliques, je peux chercher des motifs structuraux et fonctionnels dans les structures secondaires de l'ARN. Mes résultats de recherche démontrent que *MC-Motifs* peut détecter des motifs fonctionnels (sarcine-ricine, motif C, boucle T, kink-turn, « lonepair triloop », tétraboucle GNRA et « UA_handle ») dans les structures secondaires d'ARNr de *Thermus thermophilus* et d'*Haloarcula marismortui* déterminées par l'équipe de Gutell. Le programme a repéré des tétraboucles GNRA, un motif C et un kink-turn dans la structure secondaire du domaine I de l'ARNr 16S de *Thermus thermophilus*. Il a repéré des motifs « UA_handle » et une boucle T qui n'ont pas été découverts par l'équipe de Gutell. Dans la structure secondaire de l'ARNr 5S d'*Haloarcula marismortui*, *MC-Motifs* a détecté un motif sarcine-ricine, une tétraboucle GNRA et un motif nommé « lonepair triloop ».

Mon programme a réussi à identifier des boucles T et des motifs « UA_handle » dans les structures secondaires prédites du noyau structural de l'ARN riborégulateur FMN. *MC-Motifs* les a détectés dans les mêmes régions que celles indiquées dans la *Figure 4.3* et mentionnées dans la littérature (Barrick & Breaker, 2007; Jaeger et al., 2008). Il a même détecté la présence possible des motifs GNRA dans les régions P3 et P4. Avec *MC-Motifs*, je peux également suggérer la présence possible des interactions tertiaires entre la boucle T du P2 et le motif GNRA du P4.

CHAPITRE 5

DISCUSSION ET CONCLUSION

5.1 Utilité des motifs cycliques

Au chapitre deux de mon mémoire, j'ai démontré que l'on peut représenter les conformations structurales des motifs fonctionnels en utilisant les formalismes de *MC-Search* (Gendron et al., 2001; Hoffmann et al., 2001) et des motifs cycliques. Avec *MC-Search*, je peux décrire la structure tertiaire d'un motif donné en un graphe d'interactions. Comme un graphe peut contenir des cycles, je peux diviser le graphe d'interactions d'un motif en plusieurs motifs cycliques. Le motif GNRA et la boucle T, par exemple, sont des tiges-boucles d'ARN. Ils sont représentés par un motif cyclique « 4 » et « 5 » respectivement (voir les *Figures 2.1* et *2.3*). Je peux également représenter des motifs fonctionnels formés au sein des tiges-boucles (voir les *Figures 2.5*, *2.7* et *2.9*) ou dans la structure tertiaire de l'ARN (*Figure 2.11*), tels que le motif sarcine-ricine, qui est composé d'un bourgeon représenté par le motif cyclique « 3_2 » et de deux paires de bases empilée l'une sur l'autre « 2_2 » (*Figure 2.5*), et les variants du motif C qui sont composées d'un seul motif cyclique représentant une boucle interne : le motif C de type 3x5, par exemple, est représenté par un motif cyclique « 3_5 », le motif C de type 6x4 est représenté par un motif cyclique « 6_4 » (*Figure 2.9*).

J'ai pu établir une base de donnée de motifs fonctionnels, *bdMotifs*, dans laquelle sept motifs (GNRA, boucle T, sarcine-ricine, kink-turn, motif C, « UA_handle » et « lonepair triloop ») sont définis selon leurs séquences et leurs compositions structurales en terme de motifs cycliques. Je les ai choisis car ce sont des motifs qui ont été observés dans plusieurs structures cristallines d'ARN de différents types (ARNm, ARNt, ARNr, ribozyme, etc.).

5.2 Avantages de *MC-Motifs*

J'ai implanté un algorithme de recherche, *MC-Motifs*, qui, à partir de *bdMotifs*, peut détecter la présence possible des motifs structuraux et fonctionnels dans les structures secondaires. En représentant les structures secondaires des ARN et les structures tertiaires des motifs en chaînes de motifs cycliques, *MC-Motifs* a pu identifié le motif C, le GNRA, le kink-turn, le « lonepair triloop » et la sarcine-ricine dans les ARNr. Il les a repéré aux mêmes positions que celles indiquées par l'équipe de Gutell (*Figures 4.1* et *4.2*). Le programme a pu également détecté la boucle T et des motifs tertiaires tels que les motifs « UA_handle » dans la structure secondaire de l'ARNr 16S de *Thermus thermophilus* (*Figure 4.1*). Ces motifs n'ont pas été repérés par l'équipe de Gutell mais leur présence a été confirmée par d'autres (Krasilnikov & Mondragon, 2003; Jaeger et al., 2008).

MC-Motifs permet de prédire les interactions tertiaires entre les éléments structuraux, car il permet de repérer la présence possible des motifs fonctionnels faisant partie des motifs structuraux plus complexes comme les pseudonœuds. Par exemple, dans la structure secondaire du noyau structural de l'ARN du riborégulateur FMN (*Figures 4.8* et *4.13*), *MC-Motifs* a détecté un motif « UA_handle » de type II (UA_h_II) dans la région P3 où les nucléotides du bourgeon de UA_h_II s'apparient avec des nucléotides dans la région P5 formant ainsi un pseudonœud. L'équipe de Jaeger prédit que le noyau structural du riborégulateur FMN possède un domaine « T-loop PK » (Jaeger et al., 2008). Rappelons que ce domaine est composé d'un motif UA_h_II faisant partie d'un pseudonoeud (PK), d'une boucle T (T-loop) et d'autres motifs. Le domaine est situé dans les régions P3 et P5. Il y a possiblement la présence d'un domaine « T-loop PK » dans l'ARN du riborégulateur FMN de *Mesorhizobium loti* et de *Vibrio cholerae* (*Figures 5.1* et *5.2*).

Dans le domaine « T-loop PK », les boucles T sont impliquées dans la formation d'éléments structuraux tertiaires (Jaeger et al., 2008). Ces éléments structuraux sont formés

des liaisons tertiaires qui peuvent être de type « A-mineur » (Nissen et al., 2001; Jaeger et al., 2008). Les interactions « A-mineur » implique une adénine provenant du sillon mineur d'une hélice et un appariement de bases Watson-Crick, préférablement C-G, dans le sillon mineur d'une hélice voisine. Ces types d'interactions stabilisent les liaisons entre les hélices, les hélices et les boucles, et les jonctions d'hélices. L'équipe de Jaeger prédit des interactions tertiaires entre le motif UA_h_II du P5 et les nucléotides de la boucle du P3 dans le riborégulateur FMN (*Figure 4.3*). Similairement, je prédis que des interactions « A-mineur » peuvent avoir lieu entre une adénine de la boucle T (5'-CUC**A**A-3') dans la région P3 et l'appariement G-C de UA_h_II dans la région P5 du riborégulateur FMN de *Mesorhizobium loti* (*Figure 5.1*).

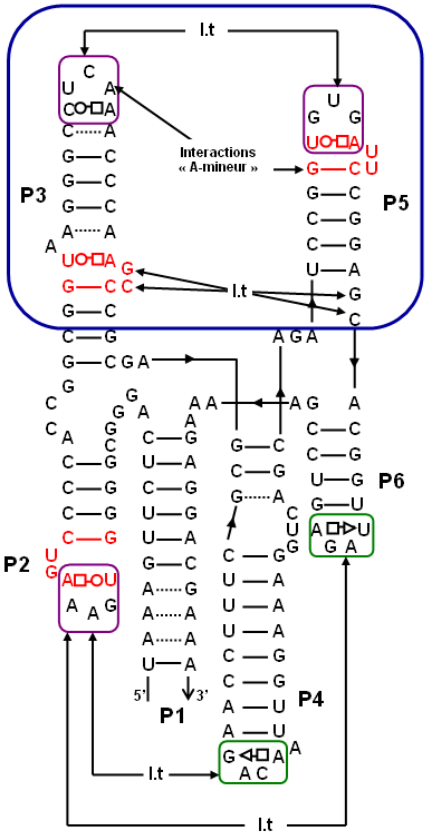


Figure 5.1 : Interactions tertiaires prédites dans la structure secondaire du noyau structural de l'ARN FMN de *Mesorhizobium loti*. Cette image est une modification de la *Figure 4.13*. Les boucles T sont encadrées en mauve, les motifs GNRA encadrés en vert et les nucléotides formant le « UA_handle » sont en rouge. Le domaine « T-loop PK » est encadré en bleu. Il est composé d'une boucle T 5'-CUCAA-3' du P3; de deux motifs UA_h_II : l'un, 5'-GU-AGCC-3' dans la région P3, faisant partie d'un pseudonoeud et l'autre, 5'-GU-AUUC-3', d'une tige-boucle du P5; et d'une deuxième boucle T 5'-UGUGA-3' dans la même tige-boucle du P5.

Je prédis la présence d'interactions tertiaires (I.t) entre les boucles T (5'-CUCAA-3' et 5'-UGUGA-3') du P3 et P5 respectivement. Il peut y avoir des interactions de type « A-mineur » entre l'adénine de la boucle T (5'-CUC**A**A-3') du P3 et l'appariement Watson-Crick, G-C, du UA_h_II du P5. Je prédis également des interactions tertiaires entre la boucle T (5'-UGAAA-3') du P2 et le motif GNRA (5'-GACA-3') du P4 ou le motif GNRA (5'-UAGA-3') du P6.

Les interactions « A-mineur » et tertiaires entre les nucléotides des régions P3 et P5 causeraient le rapprochement de la tige-boucle P3 vers la tige-boucle P5. Même s'il

n'existe pas des interactions « A-mineur », il pourrait y avoir des interactions tertiaires entre les deux boucles T des régions P3 et P5 car, quoique qu'elle se forme dans la structure secondaire, la boucle T est plus stable lorsqu'elle rentre en contact avec des nucléotides voisins dans la structure tertiaire (Zhuang et al., 2007). Des interactions tertiaires entre deux boucles T ont été observées dans le domaine I de l'ARNr 23S et dans le ribozyme ARNase P (Krasilnikov et al., 2003; Zhuang et al., 2007). Je prédis des interactions tertiaires entre la boucle T (5'-UGAAA-3') du P2 et le motif GNRA (5'-UAGA-3') du P6, tel que soupçonné par l'équipe de Jaeger (Jaeger et al., 2008). Le GNRA crée des interactions tertiaires avec les éléments structuraux éloignés et forme des sites de liaisons avec des protéines et des ARN (Correll & Swinger, 2003). Des interactions entre la boucle T du P2 et le GNRA du P6 créent un motif tertiaire composé de deux tiges-boucles. Je prédis aussi des interactions tertiaires entre la boucle T du P2 et le motif GNRA (5'-GACA-3') du P4, car *MC-Motifs* a détecté le motif GNRA dans la région P4.

Dans le domaine « T-loop PK » du riborégulateur FMN de *Vibrio cholerae*, des interactions « A-mineur » peuvent avoir lieu entre l'adénine du motif GNRA du P3 (5'-GC**A**A-3') et l'appariement G-C de UA_h_II dans la région P5 (*Figure 5.2*). La présence des interactions « A-mineur » et tertiaires entre les nucléotides des régions P3 et P5 pourrait également engendrer la formation des interactions tertiaires entre le motif GNRA du P3 et la boucle T (5'-UGAGA-3') du P5. Il pourrait aussi y avoir des interactions tertiaires entre la boucle T (5'-CGAAA-3') du P2 et la boucle du P6.

La présence des interactions tertiaires dans les riborégulateurs FMN peut être confirmée en utilisant l'outil *MC-Sym* développé dans notre laboratoire (Major et al., 1991; Parisien & Major, 2008). *MC-Sym* est un programme de modélisation structurale 3D d'ARN. Comme le programme génère des structures en format PDB, l'outil de visualisation *Rasmol* (http://rasmol.org/) et *MC-Annotate* sont employés pour vérifier la présence de ces interactions dans les modèles produits.

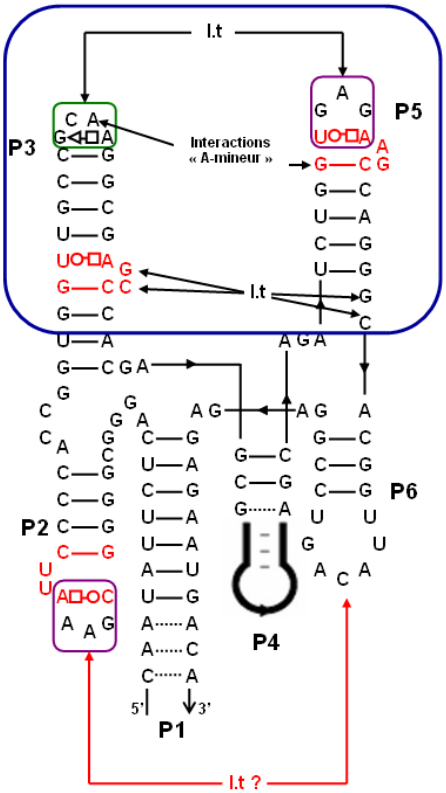


Figure 5.2 : Interactions tertiaires prédites dans la structure secondaire du noyau structural de l'ARN FMN de *Vibrio cholerae*. Cette image est une modification de la *Figure 4.8*. Les boucles T sont encadrées en mauve, le motif GNRA encadré en vert et les nucléotides formant le « UA_handle » sont en rouge. Le domaine « T-loop PK » est encadré en bleu. Il est composé d'une boucle T (5'-UGAGA-3') du P5; de deux motifs UA_h_II : l'un, 5'-GU-AGCC-3' dans la région P3, faisant partie d'un pseudonoeud et l'autre, 5'-GU-AAGC-3', d'une tige-boucle du P5; et d'un motif GNRA (5'-GCAA-3') dans la tige-boucle du P3.

Je prédis la présence d'interactions tertiaires (I.t) entre le motif GNRA (5'-GCAA-3') du P3 et la boucle T (5'-UGAGA-3') du P5. Je prédis également des interactions de type « A-mineur » entre l'adénine de GNRA (5'-GC**A**A-3') dans la région P3 et l'appariement Watson-Crick, G-C, du UA_h_II du P5. Je ne peut pas prédire la présence possible d'interactions tertiaires entre la boucle T (5'-CGAAA-3') du P2 et la boucle du P6, car *MC-Motifs* n'a pas pu détecté des motifs.

Un dernier avantage à propos de *MC-Motifs* est que l'utilisateur n'a pas à spécifier le motif à rechercher. D'autres programmes qui effectue une fonction similaire à *MC-Motifs*, par exemple, *RNAMOT* (Gautheret et al., 1990; Laferriere et al., 1994), *PatSearch* (Pesole et al., 2000; Grillo et al., 2003) et *RNAMotif* (Macke et al., 2001) nécessitent une description détaillée du motif recherché (un descripteur). L'utilisateur doit avoir une très bonne connaissance de la structure secondaire, des contraintes structurales et de la séquence du motif. Cela n'est pas nécessaire pour l'utilisateur de *MC-Motifs*, car le programme se sert de la base de données *bdMotifs* qui contient les informations structurales des motifs fonctionnels recherchés par *MC-Search* dans les structures cristallines d'ARN.

5.3 Inconvénients de *MC-Motifs*

MC-Motifs est limité à des motifs secondaires. Quoique les motifs cycliques puissent représenter la structure tertiaire des motifs structuraux, les recherches effectuées par *MC-Motifs* se font essentiellement au niveau des motifs 2D, c'est-à-dire des motifs secondaires formés d'un ou deux brins d'ARN. *MC-Motifs* ne permet pas de chercher l'existence de certaines occurrences de motifs fonctionnels formés par des liaisons tertiaires. La Figure *5.3a* illustre un exemple d'un motif tertiaire « lonepair triloop » (Lee et al., 2003; Lisi & Major, 2007) observé dans l'hélice six du domaine I de l'ARNr 16S de *Thermus thermophilus*. Ce motif, une boucle de cinq nucléotides, est formé d'une liaison tertiaire W/H *cis* entre G65 et G69. G69 forme en plus un appariement Watson-Crick avec A95. On ne peut pas représenter ce motif avec une chaîne de points et de parenthèses, c'est-à-dire un élément structural dans lequel un nucléotide s'apparie à plus d'une base (*Figure 5.3b*). Un nucléotide apparié à deux bases est nommé un triplet (Klosterman et al., 2004). Alors, en soumettant, par exemple, la chaîne de points et de parenthèses de l'hélice six à *MC-Motifs*, celui-ci ne détectera pas l'existence de « lonepair triloop ». Par contre, *MC-Motifs* a repéré un « lonepair triloop » dans l'ARNr 5S d'*Haloarcula marismortui* (*Figure 4.2*). Ce motif n'est pas formé d'une liaison tertiaire.

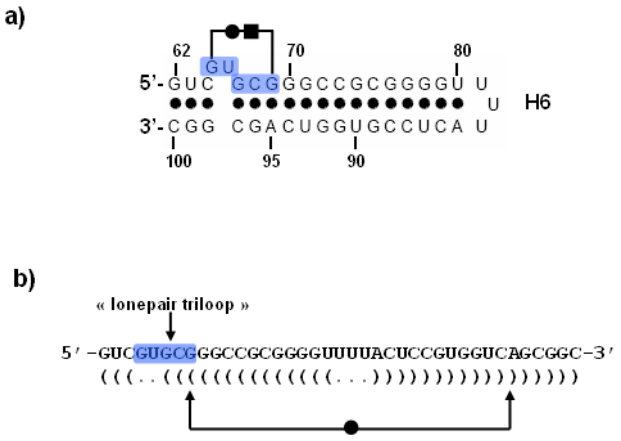


Figure 5.3 : Motif « lonepair triloop » dans l'hélice six du domaine I de l'ARNr 16S *Thermus thermophilus*. (a) Le « lonepair triloop », 5'-GUGCG-3' (nucléotides encadrés en bleu), est formé d'une liaison tertiaire W/H *cis* entre G65 et G69. G69 forme aussi un appariement W/W *cis* avec A95. **(b)** Représentation structurale de l'hélice six en chaîne de points et parenthèses. Remarquez qu'on ne peut pas assigner une paire de parenthèses à la paire de bases G65-G69 dû à la présence d'une paire de parenthèses pour l'appariement Watson-Crick G69 – A95. Dans cette exemple, *MC-Motifs* ne pourra pas détecter la présence de « lonepair triloop » dans l'hélice six.

Pour détecter la présence de « lonepair triloop » dans l'hélice six de l'ARNr 16S, il faut modéliser le motif avec une chaîne de points et de parenthèses en ignorant le triplet (voir la *Figure 5.3b*) et le projeter en 3D avec *MC-Sym*. Ensuite, il faut déterminé si le motif en 3D forme le triplet avec *MC-Search*. Cette approche fonctionnera également pour rechercher certaines occurrences de motifs fonctionnels contenant des triplets tels que celles de sarcine-ricine (Leontis et al., 2002) et de kink-turn (Lescoute et al., 2005).

L'habileté de *MC-Motifs* à repérer l'existence possible des motifs fonctionnels dépend aussi des informations disponibles sur ceux-ci dans la base de données *bdMotifs*. C'est-à-dire que les séquences exactes des motifs ainsi que leurs représentations structurales en termes de motifs cycliques doivent être connues. Par exemple, il se peut qu'un motif C existe mais que sa séquence spécifique n'existe pas dans *bdMotifs*. Heureusement, ce problème peut être remédié en utilisant l'outil *ERPIN* (Gautheret & Lambert, 2001; Lambert et al., 2004). On soumet au programme l'alignement multiple de séquences d'ARN contenant le motif recherché accompagnée d'une annotation de structure secondaire du motif. Employant une approche par profil structural, *ERPIN* repère la présence du motif au sein des séquences homologues d'ARN. Cela permet d'obtenir tous les variants du motif et d'ajouter leurs informations structurales (séquences et structure définie en termes de motifs cycliques) dans *bdMotifs*.

5.4 Conclusion

Les motifs cycliques peuvent représenter des conformations structurales spécifiques : boucles, boucles internes, bourgeons et hélices. Considérant la structure 3D d'ARN comme un graphe d'interactions, je peux utiliser ces motifs cycliques comme un vocabulaire pour décrire les structures tertiaires des motifs structuraux et fonctionnels tels que les motifs GNRA, boucle T, sarcine-ricine, « lonepair triloop », kink-turn, motif C et

« UA_handle ». J'ai alors établi une base de données, *bdMotifs*, pour rassembler ces motifs fonctionnels identifiés par *MC-Search* et définis en termes des motifs cycliques. Je peux aussi représenter les structures secondaires des ARNr et des riborégulateurs comme une chaîne de motifs cycliques. En utilisant ce formalisme, j'ai développé un algorithme, *MC-Motifs*, qui recherche l'existence possible de motifs dans une structure secondaire d'ARN donnée à partir de *bdMotifs*. *MC-Motifs* a pu détecter avec succès les motifs fonctionnels dans les ARNr d'*Haloarcula marismortui* et de *Thermus thermophilus*. La prédiction des motifs fonctionnels dans les structures secondaires des riborégulateurs FMN m'a permis de définir des contraintes structurales aidant à la modélisation structurale des riborégulateurs et donc à la compréhension sur les fonctions de ces molécules. Quoique *MC-Motifs* présente des défauts, j'ai suggéré des solutions pour les remédier.

L'habileté de *MC-Motifs* dépend des informations disponibles dans *bdMotifs*, je chercherai et ajouterai d'autres motifs fonctionnels provenant de la littérature dans cette base de données. De plus, j'effectuerai l'annotation des structures 3D d'ARN afin de rechercher des séquences pour chaque motif cyclique. Je désire connaître la distribution des séquences dans chaque motif cyclique.

Certes, les motifs cycliques peuvent faire partie de l'ontologie d'ARN car ils sont basés sur les séquences d'ARN, les structures secondaires et tertiaires. Ils permettent des description précises des motifs d'ARN.

Bibliographie

Aho AV, Hopcroft JE, Ullman JD. 1974. *The Design and Analysis of Computer Algorithms*: Reading, MA, USA Addison-Wesley

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology 215*:403-410.

Auffinger P, Westhof É. 2000. *RNA tertiary structure, "Encyclopedia of Analytical Chemistry"*: John Wiley & Sons Ltd, Chichester.

Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. 2000. The complete atomic structure of the large ribosomal subunit at 2.4 A resolution. *Science 289*:905-920.

Batey RT, Rambo RP, Doudna JA. 1999. Tertiary motifs in RNA structure and folding. *Angewandte Chemie-International Edition 38*:2327-2343.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res 28*:235-242.

Bermúdeza CI, Dazab EE, Andrade E. 1999. Characterization and comparison of Escherichia coli transfer RNAs by graph theory based on secondary structure. *Journal of Theoretical Biology 197*:193-205.

Carter AP, Celmons WM, Brodersen DE, Morgan-Warren RJ, Wimberly BT, Ramakrishnam V. 2000. Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature 407*:340-348.

Cate JH, Gooding AR, Podell E, Zhou KH, Golden BL, Kundrot CE, Cech TR, Doudna JA. 1996. Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science 273*:1678-1685.

Cech TR. 2000. Structural biology - The ribosome is a ribozyme. *Science 289*:878-879.

Correll CC, Swinger K. 2003. Common and distinctive features of GNRA tetraloops based on a GUAA tetraloop structure at 1.4 Å resolution *RNA 9*:355-363.

Crick F. 1970. Central Dogma of Molecular Biology. *Nature 227*:561-563.

Edwards TE, Klein DJ, Ferré-D'Amaré AR. 2007. Riboswitches: small-molecule recognition by gene regulatory RNAs. *Current Opinion in Structural Biology 17*:273-279.

Epshtein V, Mironov AS, Nudler E. 2003. The riboswitch-mediated control of sulfur metabolism in bacteria. *Proceedings of the National Academy of Sciences of the United States of America 100*:5052-5056.

Gan HH, Pasquali S, Schlick T. 2003. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Res 31*:2926–2943.

Gautheret D, Gutell R. 1997. Inferring the conformation of RNA base pairs and triples from patterns of sequence variation *Nucleic Acids Research 25*:1559-1564.

Gautheret D, Konings D, Gutell RR. 1995. G.U base pairings motifs in ribosomal RNA. *RNA 1*:807-814.

Gautheret D, Lambert A. 2001. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *Journal of Molecular Biology 313*:1003-1011.

Gautheret D, Major F, Cedergren R. 1990. Pattern searching alignment with RNA primary and secondary structures: an effective descriptor for tRNA *Computer applications in the biosciences 6*:325-331.

Geary C, Baudrey S, Jaeger L. 2008. Comprehensive features of natural and in vitro selected GNRA tetraloop-binding receptors. *Nucleic Acids Res 36*:1138-1152.

Gendron P, Lemieux S, Major F. 2001. Quantitative analysis of nucleic acid three-dimensional structures. *J Mol Biol 308*:919-936

Gotoh O. 1987. Pattern matching of biological sequences with limited storage. *Computer Applications in the Biosciences 3*:17-20.

Grillo G, Licciulli F, Liuni S, Sbisa E, Pesole G. 2003. PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Res 31*:3608-3612.

Gutell RR, Larsen N, Woese CR. 1994. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiological Reviews 58*:10-26.

Gutell RR, Lee JC, Cannone JJ. 2002. The accuracy of ribosomal RNA comparative structural models. *Curr Opin Struct Biol 12*:301 – 310

Hermann T, Westhof E. 1999. Non-Watson-Crick base pairs in RNA-protein recognition. *Chemistry & Biology 6*:R335-R343.

Heus HA, Pardi A. 1991. STRUCTURAL FEATURES THAT GIVE RISE TO THE UNUSUAL STABILITY OF RNA HAIRPINS CONTAINING GNRA LOOPS. *Science 253*:191-194.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte Fur Chemie 125*:167-188.

Hoffmann B, Mitchell GT, Gendron P, Major F, Andersen AA, Collins RA, Legault P. 2001. NMR structure of the active conformation of the Varkud satellite ribozyme cleavage site. *Proc Natl Acad Sci U S A 100*:7003-7008

Hogeweg P, Hesper B. 1984. Energy directed folding of RNA sequences. *Nucleic Acids Research 12*:67-74.

Horton J. 1987. A polynomial-time algorithm to find the shortest cycle basis of a graph. *Siam J Comput 16*:358-366.

Jaeger L, Verzemnieks EJ, Geary C. 2008. The UA_handle: a versatile submotif in stable RNA architectures. *Nucleic Acids Res 1*:1-16.

Kim SH, Sussman JL, Suddath FL, Quigely GJ, McPherson A, Wang AHJ, Seeman NC, Rich A. 1974. The General Structure of Transfer RNA Molecules. *Proc Natl Acad Sci USA 71*:4970-4974.

Klein DJ, Schmeing TM, Moore PB, Steitz TA. 2001. The kink-turn: a new RNA secondary structure motif. *Embo Journal 20*:4214-4221.

Klosterman PS, Hendrix DK, Tamura M, Holbrook SR, Brenner SE. 2004. Three-dimensional motifs from the SCOR, structural classification of RNA database: extruded strands, base triples, tetraloops and U-turns. *Nucl Acids Res 32*:2342-2352.

Krasilnikov AS, Mondragon A. 2003. On the occurrence of the T-loop RNA folding motif in large RNA molecules. *RNA-Publ RNA Soc 9*:640-643.

Krasilnikov AS, Yang X, Pan T, Mondragon A. 2003. Crystal structure of the specificity domain of ribonuclease P. *Nature 421*:760-764.

Laferriere A, Gautheret D, Cedergren R. 1994. An RNA pattern-matching program with enhanced performance and portability. *Computer applications in the biosciences 10*:211-212.

Lambert A, Fontaine J-F, Legendre M, Leclerc F, Permal E, Major F, Putzer H, Delfour O, Michot B, Gautheret D. 2004. The ERPIN server: an interface to profile-based RNA motif identification. *Nucl Acids Res 32*:W160-165.

Lee JC, Cannone JJ, Gutell RR. 2003. The lonepair triloop: A new motif in RNA structure. *J Mol Biol 325*:65-83.

Lemieux S, Chartrand P, Cedergren R, Major F. 1998. Modeling active RNA structures using the intersection of conformational space: Application to the lead-activated ribozyme. *RNA 4*:739-749.

Lemieux S, Major F. 2006. Automated extraction and classification of RNA tertiary structure cyclic motifs. *Nucleic Acids Res 34*:2340-2346.

Lengauer T. 2007. *Bioinformatics - from Genomes to Therapies*: John Wiley & Sons.

Leontis NB, Stombaugh J, Westhof E. 2002. Motif prediction in ribosomal RNAs: lessons and prospects for automated motif prediction in homologous RNA molecules. *Biochimie 84*:961-973.

Leontis NB, Westhof E. 1998. A Common Motif Organizes the Structure of Multi-helix Loops in 16 S and 23 S Ribosomal RNAs. *Journal of Molecular Biology 283*:571-583.

Leontis NB, Westhof E. 2001. Geometric nomenclature and classification of RNA base pairs. *Rna-a Publication of the Rna Society 7*:499-512.

Lescoute A, Leontis NB, Massire C, Westhof E. 2005. Recurrent structural RNA motifs, isostericity matrices and sequence alignments. *Nucleic Acids Research 33*:2395-2409.

Lewin B. 1997. *Genes V*. New York: Oxford : Oxford University Press.

Lisi V, Major F. 2007. A comparative analysis of the triloops in all high-resolution RNA structures reveals sequence-structure relationships. *RNA-Publ RNA Soc 13*:1537-1545.

Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. 2001. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res 29*:4724-4735.

Major F, Thibault P. 2007. RNA Tertiary Structure Prediction. In Bioinformatics: From Genomes to Therapies, T. Lengauer. In: Weinheim G, Wiley-VCH, ed. pp 491-539.

Major F, Turcotte M, Gautheret D, Lapalme G, Fillion E, Cedergren R. 1991. The Combination of Symbolic and Numerical Computation for Three-Dimensional Modeling of RNA. *Science 253*:1255-1260.

Mandal M, Boese B, Barrick JE, Winkler WC, Breaker RR. 2003. Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. *Cell 113*:577-586.

Marintchev A, Wagner G. 2004. Translation initiation: structures, mechanisms and evolution. *Quarterly Reviews of Biophysics 37*:197-284.

Michel F, Westhof E. 1990. Modelling of the three-dimensional architecture of group-I catalytic introns based on comparative sequences analysis. *Journal of Molecular Biology 216*:585-610.

Nagaswamy U, Fox GE. 2002. Frequent occurrence of the T-loop RNA folding motif in ribosomal RNAs. *Rna-a Publication of the Rna Society 8*:1112-1119.

Nahvi A, Sudarsan N, Ebert MS, Zou X, Brown KL, Breaker RR. 2002. Genetic control by a metabolite binding mRNA. *Chemistry & Biology 9*:1043-1049.

Nissen P, Ippolito JA, Ban N, Moore PB, Steitz TA. 2001. RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. *Proceedings of the National Academy of Sciences of the United States of America 98*:4899-4903.

Noller HF. 2005b. RNA Structure: Reading the Ribosome *Science 309*:1508-1514.

Noller HF, Hoang L, Fredrick K. 2005. The 30S ribosomal P site: a function of 16S rRNA. *FEBS Letters 579*:855-858.

Noller HF, Hoffarth V, Zimniak L. 1992. Unusual Resistance of Peptidyl Transferase to Protein Extraction Procedures. *Science 256*:1416-1419.

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

Pesole G, Liuni S, D'Souza M. 2000. PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics 16*:439-450.

Petersheim M, Turner D. 1983. Base-stacking and base-pairing contributions to helix stability: thermodynamics of double-helix formation with CCGG, CCGGp, CCGGAp, ACCGGp, CCGGUp, and ACCGGUp. *Biochemistry 22*:256-263.

Pley H, Flaherty K, McKay D. 1994. Three-dimensional structure of a hammerhead ribozyme. *Nature 372*:68-74.

Puglisi JD, Blanchard SC, Green R. 2000. Approaching translation at atomic resolution. *Nature Structural Biology 7*:855-861.

Quigley G, Rich A. 1976. Structural domains of transfer RNA molecules. *Science 194*:796-806.

Saenger W. 1984. *Principles of Nucleic Acid Structure*. New York: Springer-Verlag.

Spackova N, Sponer J. 2006. Molecular dynamics simulations of sarcin-ricin rRNA motif. *Nucleic Acids Research 34*:697-708

Szewczak AA, Moore PB, Chan YL, Wool IG. 1993. The conformation of the sarcin ricin loop from 28S ribosomal-RNA. *Proceedings of the National Academy of Sciences of the United States of America 90*:9581-9585.

Torres-Larios A, Dock-Bregeon AC, Romby P, Rees B, Sankaranarayanan R, Caillet J, Springer M, Ehresmann C, Ehresmann B, Moras D. 2002. Structural basis of

translational control by Escherichia coli threonyl tRNA synthetase. *Nature Structural Biology 9*:343–347

Turner B, Lilley DMJ. 2008. The Importance of G•A Hydrogen Bonding in the Metal Ion- and Protein-induced Folding of a Kink Turn RNA. *Journal of Molecular Biology 381*:431-442

Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS. 2004. Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends in Genetics 20*:44-50.

Westhof É, Fritsch V. 2000. RNA folding: beyond Watson-Crick pairs. *Structure 8*:55-65.

Wickiser JK, Winkler WC, Breaker RR, Crothers DM. 2005. The Speed of RNA Transcription and Metabolite Binding Kinetics Operate an FMN Riboswitch. *Molecular Cell 18*:49-60.

Wimberly BT, Brodersen DE, Clemons WM, Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V. 2000. Structure of the 30S ribosomal subunit. *Nature 407*:327-339.

Winkler W, Nahvi A, Breaker RR. 2002a. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature 419*:952-956.

Winkler WC, Breaker RR. 2003. Genetic control by metabolite-binding riboswitches. *Chembiochem 4*:1024-1032.

Winkler WC, Cohen-Chalamish S, Breaker RR. 2002b. An mRNA structure that controls gene expression by binding FMN. *Proceedings of the National Academy of Sciences of the United States of America 99*:15908-15913.

Zhuang ZY, Jaeger L, Shea JE. 2007. Probing the structural hierarchy and energy landscape of an RNA T-loop hairpin. *Nucleic Acids Research 35*:6995-7002.

Annexe A : Algorithme de *pdb2NCM.pl*

L'algorithme de *pdb2NucleotideCyclicMotif.pl* ou *pdb2NCM.pl*, écrit en langage Perl, permet à l'utilisateur de représenter un motif structural d'ARN généré par *MC-Search* en termes de motifs cycliques. L'algorithme comprend deux étapes. La première étape consiste à représenter la structure tertiaire du motif en chaîne de points et parenthèses. Le point représente une base non appariée et la parenthèse, qu'elle soit gauche ou droite, représente une base appariée. La deuxième étape consiste à traduire l'annotation en termes de motifs cycliques.

Première étape : Représentation du motif en chaîne de points et parenthèses

(A) *pdb2NCM.pl* prend en entrée la structure tertiaire du motif

Dans la *figure A1*, *pdb2NCM.pl* prend entrée un motif structural (ex : sarcin-ricin) généré par *MC-Search* (*figure A1-a*). *pdb2NCM.pl* fait appel à *MC-Annotate* pour obtenir le graphe d'interactions du motif (*figure A1-b*). Il garde les arêtes qui représentent les interactions entre les bases appartenant aux brins complémentaires (*figure A1-c*), afin de décrire facilement la structure du motif en divers motifs cycliques.

(B) Représentation du graphe d'interactions en chaîne de points et parenthèses

pdb2NCM.pl assigne les symboles « (» et «) » aux bases reliées par les arêtes et des « . » aux bases non appariées. La structure du sarcin-ricin en chaîne de points et parenthèses est donc « (. (( ))) » (*figure A2*).

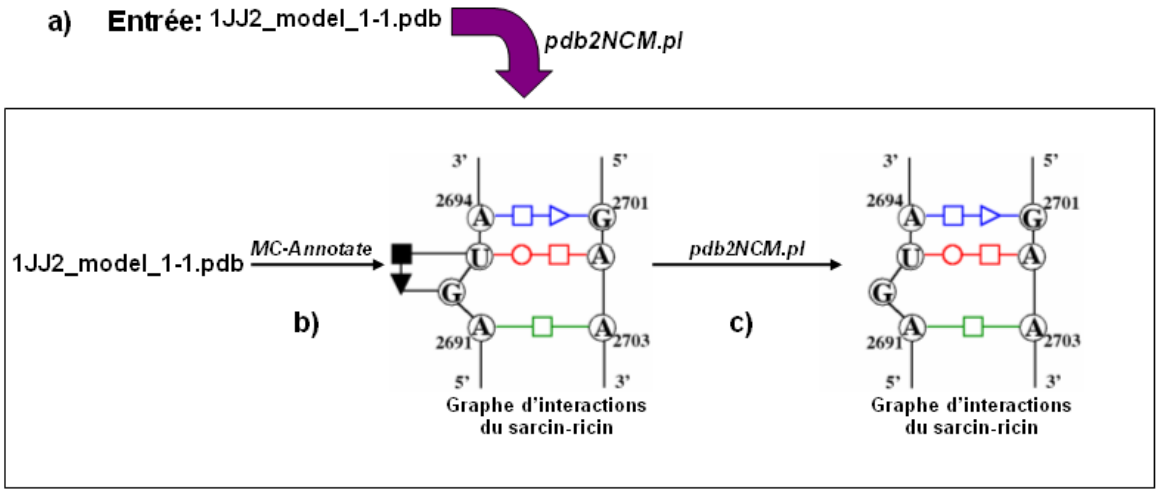


Figure A1 : Prise de données par *pdb2NCM.pl*. **(a)** *pdb2NCM.pl* prend entrée le motif structural sarcin-ricin ressorti par *MC-Search*. **(b)** Le programme retient seulement les arêtes reliant les bases appartenant aux brins complémentaires. **(c)** Dans le sarcin-ricin, l'appariement S/W *cis* entre G2692 et U2693 est retiré.

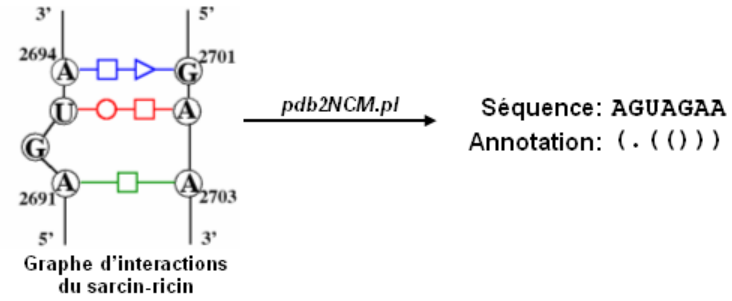


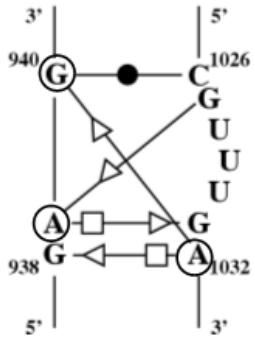
Figure A2 : Représentation de la structure de sarcin-ricin en chaîne de points et parenthèses par *pdb2dotsBrackets.pl*.

Cas exceptionnel pour certains motifs

Les étapes **(A)** et **(B)** fonctionnent lorsque chaque base d'un motif est appariée à une base du brin complémentaire. Mais, qu'arrive-t-il s'il existe des bases qui s'apparient à plus d'une base du brin complémentaire? Il faut s'assurer que la structure du motif puisse être représenté en divers motifs cycliques. Dans ce cas, *pdb2NCM.pl* choisit, dans le

graphe d'interactions du motif, un ou des ensembles d'arêtes qui ne s'entrecroisent pas avec d'autres arêtes; et, des arêtes n'appartenant pas à une même base.

Prenons l'exemple d'un kink-turn de l'ARNr 23S d'*Haloarcula marismortui*. Ce motif contient trois nucléotides qui sont appariés à plus d'une base dans le brin complémentaire allant de 5' → 3' : A939, G940 et A1032 :



structure secondaire
du kink-turn

Dans le graphe d'interactions du motif (*figure A3-b)*, *pdb2NCM.pl* choisit des ensembles d'arêtes qui ne s'entrecroisent pas avec d'autres arêtes, et, des arêtes n'appartenant pas à un même sommet (*figures A3-c* et *A3-d*). Dans le graphe d'interactions du kink-turn, les arêtes du graphe sont numérotées afin de mieux comprendre la figure.

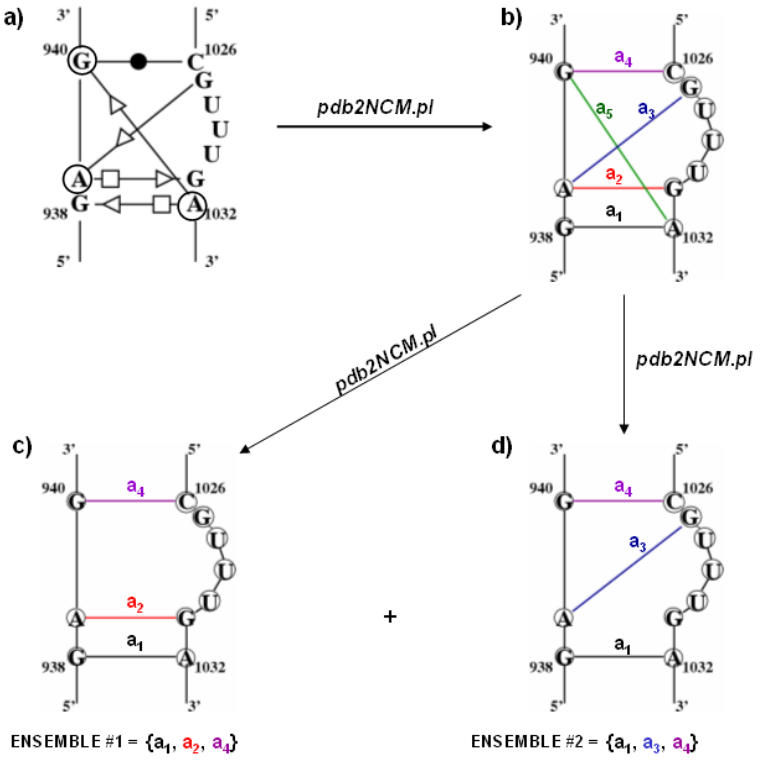


Figure A3 : Ensembles d'arêtes du graphe d'interactions du kink-turn. **(a)** Structure secondaires du kink-turn de l'ARNr 23S d'*Haloarcula marismortui*. **(b)** Graphe d'interactions du kink-turn. **(c et d)** Deux ensembles d'arêtes du graphe d'interactions où chaque ensemble contient des arêtes qui ne s'entrecroisent pas.

Pour le graphe d'interactions du kink-turn, il y a deux ensembles d'arêtes (*figures A3-c* et *A3-d*). L'ensemble d'arêtes #1 contenant a_1 (noir), a_2 (rouge) et a_4 (mauve) et l'ensemble d'arêtes #2 contenant a_1 (noir), a_3 (bleu) et a_4 (mauve). Remarquez que, dans chacune de ces ensembles, chaque base est rattachée qu'à une seule arête et que les arêtes ne s'entrecroisent pas. Ainsi, certaines occurrences d'un motif, comme celui du motif kink-turn de la *figure A3-a*, possèdent plus qu'une représentation de leur structure en chaîne de points et parenthèses (*figure A4*).

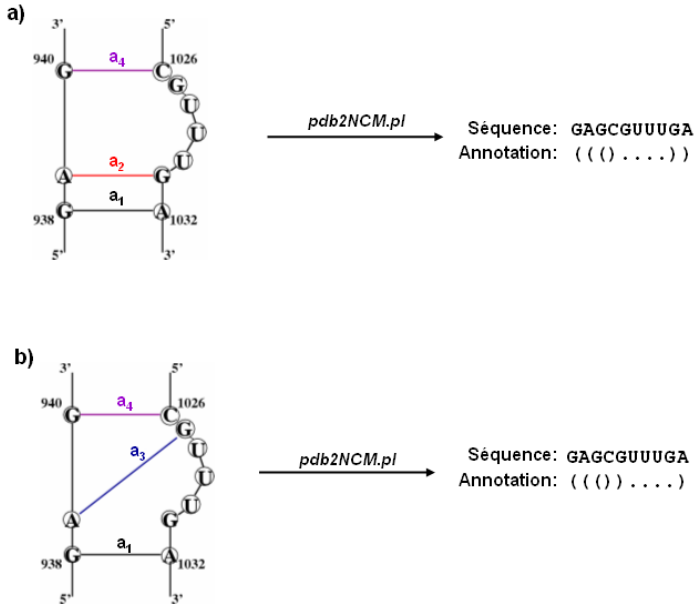


Figure A4 : Représentations du motif kink-turn en chaîne de points et parenthèses. Le motif kink-turn de l'ARNr 23S d'*H. marismortui* a deux représentations de sa structures en chaîne de points et parenthèses : **(a)** « ((( ) . . . .)) » et **(b)** « ((( )) . . . .) ».

Voici d'autres exemples de *pdb2NCM.pl* pour la représentation structurale du motif C de type 6x4 (*figure A5*) et de la tétraboucle GNRA (*figure A6*) en chaînes de points et parenthèses :

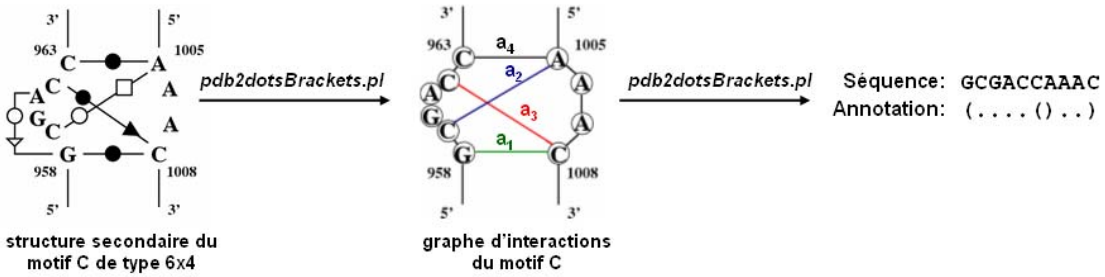


Figure A5 : Représentation du motif C en chaîne de points et parenthèses. La chaîne de points et parenthèses du motif C de type 6x4, est « (. . . . ( ) . .) ». *pdb2NCM.pl* ne tient pas compte des arêtes a_2 (bleu) et a_3 (rouge), car elles s'entrecroisent.

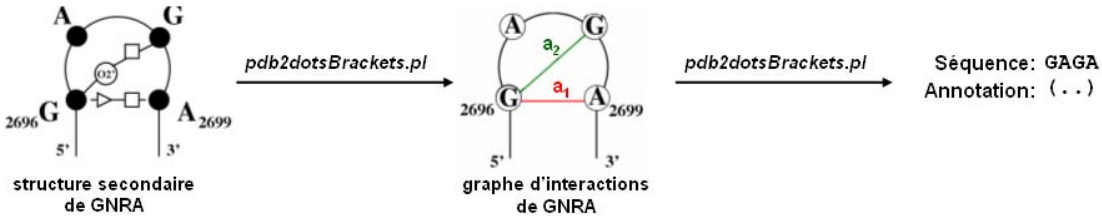


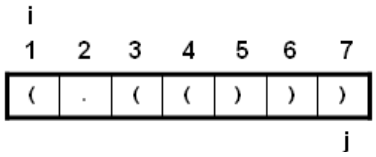
Figure A6 : Représentation de la tétraboucle GNRA en chaîne de points et parenthèses. La chaîne de points et parenthèses du motif GNRA est « (. .) ». *pdb2NCM.pl* ne tient pas compte de l'arête a_2 (vert), car elle appartient à un même sommet, G2696, que l'arête a_1 (rouge).

Deuxième étape : Transformation de la chaîne de points et parenthèses d'un motif en chaîne de motifs cycliques

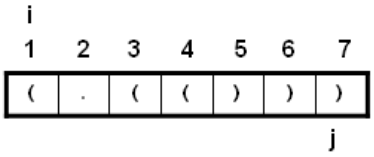
pdb2NCM.pl fait appel à *NCMsInSequence*, un module écrit en Java, qui transforme la chaîne de points et parenthèses en une chaîne de motifs cycliques. *NCMsInSequence* fait partie du moteur de recherche de motifs *MC-Motifs*. Les figures suivantes illustrent l'algorithme de *NCMsInSequence*. Pour décrire l'algorithme, la chaîne de points et parenthèses du sarcin-ricin est utilisée comme exemple :

```
5'-AGUAGAA-3'
 (.(()))
```

Dans l'algorithme de *NCMsInSequence*, l'entrée est un tableau T[1…n] contenant la chaîne de points et parenthèses, i et j sont les indices du tableau. Ces indices désignent les éléments du tableau. L'indice i est initialisé à la position du premier caractère de la chaîne alors que l'indice j à la position du dernier caractère :



(Étape 1) Tant que l'élément de T[i] est différent de «) » et que celui de T[j] est différent de « (», *NCMsInSequence* cherche deux paires de « (» et «) » afin d'identifier le type de motif cyclique à double brin. Dans la figure ci-dessous, *NCMsInSequence* trouve la première paire de « (» et «) » aux indices 1 et 7 respectivement. Les éléments de T[1] et T[7] sont sauvegardés respectivement dans les variables nommés $symbole_i$ et $symbole_j$:

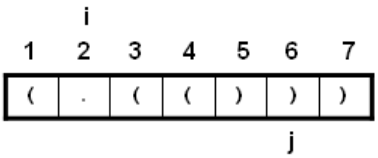


$symbole_i$ = « (»

$symbole_j$ = «) »

Paires de « (» et «) » trouvés = 1

(Étape 2) L'indice i est augmenté de 1, alors que l'indice j est diminué de 1 :

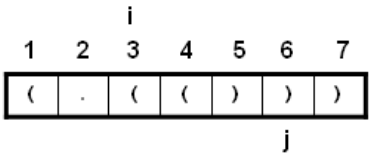


$symbole_i$ = « (»

$symbole_j$ = «) »

Paires de « (» et «) » trouvés = 1

Si l'élément de T[i] équivaut à « . », alors l'élément est ajouté dans la variable $symbole_i$ et l'indice i est augmenté de 1. Sinon, si c'est l'élément de T[j] qui équivaut à « . », alors l'élément est ajouté dans la variable $symbole_j$ et l'indice j est diminué de 1. Dans la figure ci-dessous, l'élément « . » se trouve à l'indice i = 2. L'élément est alors ajouté dans le $symbole_i$ et i est augmenté de 1 :

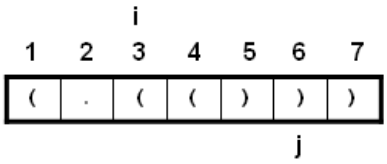


$symbole_i$ = « (. »

$symbole_j$ = «) »

Paires de « (» et «) » trouvés = 1

NCMsInSequence reprend l'étape 1 et il trouve une secondaire paire de « (» et «) » à i = 3 et j = 6 respectivement. L'éléments de T[3] est ajouté dans le $symbole_i$ et celui de T[6] dans le $symbole_j$:

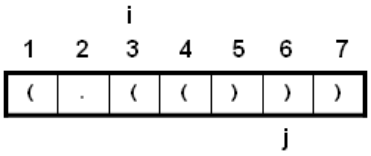


$symbole_i$ = « (. (»

$symbole_j$ = «)) »

Paires de « (» et «) » trouvés = 2

(Étape 3) Comme deux paires de « (» et «) » a été détecté, *NCMsInSequence* assigne un code d'identification (ID), m_n, pour identifier le type de motif cyclique à double brin. Ce code est formé par l'expression « #$symbole_i$ _#$symbole_j$ » où #$symbole_x$ indique le nombre de caractères dans la variable. Donc, le $symbole_i$ contient 3 caractères et le $symbole_j$ contient 2 caractères. Le premier motif cyclique identifié est alors 3_2 :



$symbole_i$ = « (. (»

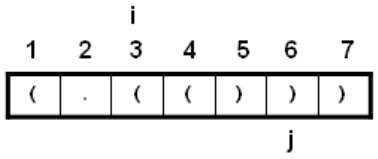
$symbole_j$ = «)) »

ID = #$symbole_i$_ #$symbole_i$

ID = 3_2

Chaîne de motifs cycliques = 3_2

(Étape 4) *NCMsInSequence* réinitialise les variables $symbole_i$ et $symbole_j$ à « », et le nombre de paires de « (» et «) » trouvé à 0 afin de chercher les prochains groupes de deux « (» et «) » :

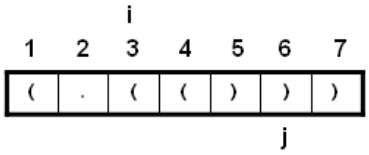


$symbole_i$ = « »

$symbole_j$ = « »

Paires de « (» et «) » trouvés = 0

NCMsInSequence reprend l'étape 1 et il trouve une paire de « (» et «) » à i = 3 et j = 6 respectivement. Les éléments de T[3] et T[6] sont sauvegardés respectivement dans les $symbole_i$ et $symbole_j$:

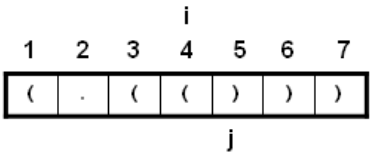


$symbole_i$ = « (»

$symbole_j$ = «) »

Paires de « (» et «) » trouvés = 1

NCMsInSequence reprend l'étape 2 : l'indice i est augmenté de 1, alors que l'indice j est diminué de 1 :

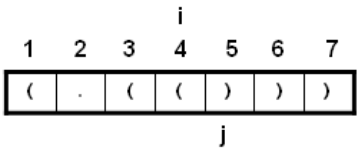


$symbole_i$ = « (»

$symbole_j$ = «) »

Paires de « (» et «) » trouvés = 1

NCMsInSequence trouve une secondaire paire de « (» et «) » à i = 4 et j = 5 respectivement. L'éléments de T[4] est ajouté dans le $symbole_i$ et celui de T[5] dans le $symbole_j$:

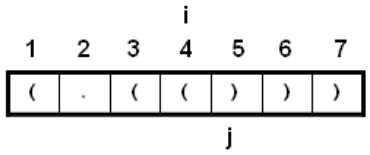


$symbole_i$ = « ((»

$symbole_i$ = «)) »

Paires de « (» et «) » trouvés = 2

Comme deux paires de « (» et «) » a été détecté, *NCMsInSequence* assigne un code ID, m_n, au second type de motif cyclique à double brin. Le second motif cyclique identifié est 2_2 :



$symbole_i$ = « ((»

$symbole_j$ = «)) »

ID = #$symbole_i$ _ #$symbole_j$

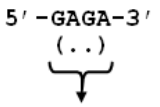
ID = 2_2

Chaîne de motifs cycliques = 3_2 + 2_2

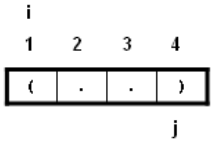
Le sarcin-ricin est alors composé de motifs cycliques 3_2 et 2_2 :

Représentation du sarcin-ricin en motifs cycliques

Si, dans une chaîne de points et parenthèses d'un motif donné, *NCMsInSequence* ne trouve qu'un seul un groupe de « (» et «) », cela signifie que le motif cyclique représente une boucle. *NCMsInSequence* assigne un code d'identification (ID), m, pour identifier le type de motif cyclique à simple brin. Ce code est formé par la somme des nombres de caractères dans $symbole_i$ et dans $symbole_j$ Le schéma ci-dessous affiche cet exemple. Celui du motif GNRA (5'-GAGA-3') dans l'ARNr 23S de *Haloarcula marismortui* :



Initialisation des variables



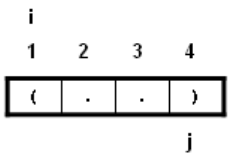
$symbole_i$ = «»

$symbole_j$ = « »

Paires de « (» et «) » trouvés = 0

Tant que l'élément de T[i] est différent de «) » et que l'élément de T[j] est différent de « (», *NCMsInSequence* cherche deux paires de « (» et «) ».

Première paire de « (» et «) » détectée

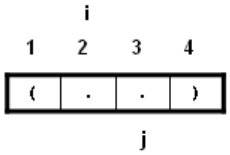


$symbole_i$ = « (»

$symbole_j$ = « (»

Paires de « (» et «) » trouvés = 1

L'indice i est augmenté de 1, l'indice j est diminué de 1



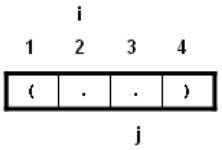
$symbole_i$ = « (»

$symbole_j$ = «) »

Paires de « (» et «) » trouvés = 1

L'élément de T[2] équivaut à « . », l'élément est ajouté dans $symbole_i$

L'indice i est augmenté de 1



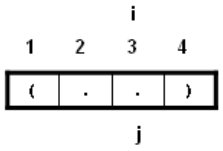
$symbole_i$ = « (. »

$symbole_j$ = «) »

Paires de « (» et «) » trouvés = 1

L'élément de T[3] équivaut à « . », l'élément est ajouté dans $symbole_i$

L'indice i est augmenté de 1

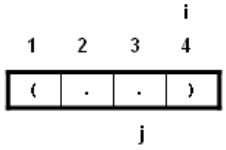


$symbole_i$ = « (. . »

$symbole_j$ = «) »

Paires de « (» et «) » trouvés = 1

NCMsInSequence a terminé de parcourir le tableau, car T[4] n'est pas différent de «) ».



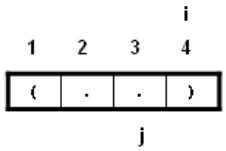
$symbole_i$ = « (. . »

$symbole_j$ = «) »

Paires de « (» et «) » trouvés = 1

NCMsInSequence n'a trouvé qu'une seule paire de « (» et «) ».

NCMsInSequence assigne un code ID, m, pour identifier le type de motif cyclique à simple brin.



$symbole_i$ = « (. . »

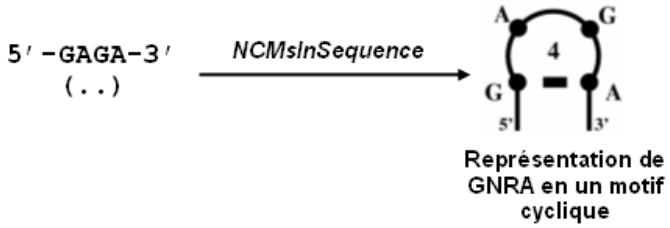
$symbole_j$ = «) »

ID = #$symbole_i$ + #$symbole_j$

ID = 4

Chaîne de motifs cycliques = 4

Le motif GNRA (5'-GAGA-3') est composé d'un motifs cyclique 4:



Représentation de GNRA en un motif cyclique

Le temps d'exécution de *NCMsInSequence* se fait en O(n)[7] où n correspond au nombre de caractères dans la chaîne de points et parenthèses. En d'autres mots, le temps d'exécution dépend de la longueur de la chaîne.

La *table A7* affiche quelques résultats sur les représentations des motifs structuraux en terme de motifs cycliques par *NCMsInSequence* via *pdb2NCM.pl*.

MOTIFS	SÉQUENCES (5' → 3')	MOTIF(S) CYCLIQUE(S)
GNRA	GAGA	4
GNRA	GCGA	4
GNRA	GCAA	4
GNRA	GUGA	4
GNRA	GGAA	4
GNRA	GAAA	4
kink-turn (Kt-7)	GAG-CGAAGA	2_5 + 2_2
kink-turn (Kt-7)	GAG-CGAAGA	2_2 + 2_5
kink-turn (Kt-15)	GAAG-CAAUGU	2_5 + 3_2
kink-turn (Kt-15)	GAAG-CAAUGU	2_2 + 3_5
kink-turn (Kt-23)	UAG-CGCAGA	2_2 + 2_5
kink-turn (Kt-23)	UAG-CGCAGA	2_5 + 2_2
kink-turn (Kt-38)	GAG-CGUUUGA	2_6 + 2_2
kink-turn (Kt-38)	GAG-CGUUUGA	2_2 + 2_6
kink-turn (Kt-42)	CCUAGA-GAG	2_2 + 5_2
kink-turn (Kt-42)	CCUAGA-GAG	5_2 + 2_2

[7] Notation utilisée en informatique pour exprimer l'ordre de grandeur des temps d'exécution quand *n* devient très grand.

kink-turn (Kt-46)	GAUGGA-GAC	2_2 + 5_2
kink-turn (Kt-46)	GAUGGA-GAC	5_2 + 2_2
kink-turn (Kt-58)	GAAGC-GCAGGA	2_5 + 4_2
kink-turn (Kt-58)	GAAGC-GCAGGA	4_5 + 2_2
motif C type 5x2	GCAAU-AC	5_2
motif C type 5x3	GCACU-AGC	5_3
motif C type 5x3	CCACU-ACG	5_3
motif C type 5x4	GCACU-AAAC	5_4
motif C type 6x4	GCGACC-AAAC	6_4
sarcin-ricin	AGUA-GAA	3_2 + 2_2
sarcin-ricin	AGUA-AAA	3_2 + 2_2
boucle T	UGAGA	5
boucle T	UGCAA	5
boucle T	AGAGA	5
boucle T	CGAAA	5
UA_handle (type I)	CU-AUG	2_3
UA_handle (type I)	UU-AAA	2_3
UA_handle (type II)	GU-AUAC	2_4
UA_handle (type II)	AAAC-GU	4_2
UA_handle (type III)	UU-AAAAA	2_5
UA_handle (type III)	AGAAC-GU	5_2

Table A7: Représentation des structures des motifs en termes de motifs cycliques.

Annexe B : Les modules de *MC-Motifs*

MC-Motifs a été développé en langage Java. Il est composé de sept fichiers considérés comme des modules.

NOMS DE FICHIERS	DESCRIPTION
motifsInSequence.java	Module principal recevant en entrée le nom de la molécule d'ARN, sa structure secondaire de l'ARN en chaîne de points et parenthèses, sa séquence de nucléotides. Il active le module *mcBloc* et il fait appel à *NCMsInSequence* pour la recherche de motifs. Il affiche à l'utilisateur les motifs détectés dans la chaîne de points et parenthèses.
mcBloc.java	Module séparant la structure secondaire de l'ARN en hélices et tiges-boucles. Chaque élément structural est représenté par leur chaîne de points et parenthèses.
Hairpin.java	Module contenant la séquence d'une tige-boucle, sa structure secondaire en chaîne de points et parenthèses, et sa position au sein de la structure secondaire de l'ARN.
Stem.java	Module contenant la séquence d'une hélice, sa structure secondaire en chaîne de points et parenthèses, et sa position au sein de la structure secondaire de l'ARN.

NCMsInSequence.java	Ce module représente la chaîne de points et parenthèses en chaîne de motifs cycliques. Par la suite, il recherche la présence possible des motifs fonctionnels à l'aide de *dbIUPAC*.
baseDeDonnees.java	Module ayant accès aux bases de données *bdMotifs* et *dbIUPAC* .
dbIUPAC.java	Ce module renvoie à *NCMsInSequence* la séquence de chaque motif selon la description structurale de celui-ci dans *bdMotifs*.

Annexe C : Pseudocode pour convertir une chaîne de points et parenthèses en chaîne de motifs cycliques

Dans ce pseudocode, provenant du module *NCMsInSequence*, l'entrée est un tableau T[1…n] contenant la chaîne de points et parenthèses, i et j sont les indices du tableau:

NCMsInSequence(Entrée : T[1…n]; sortie : chaîneCycles)

1. **Début**
2. i ← 1;
3. j ← longueur du tableau T[];
4. $quantité_i$ ← 0; variable contenant la quantité de caractères « (» et « . »
5. $quantité_j$ ← 0; variable contenant la quantité de caractères «) » et « . »
6. quantitéPB ← 0; nombre de paires de bases identifiées
7. PBMAX ← 2; nombre de paires de bases maximales pour identifier un motif pour identifier un motif cyclique double brin
8. ncmID ← «»; code d'identification du motif cyclique
9. chaîneCycles ← «»; chaîne de motifs cycliques (ncmID) représentant la chaîne de points et parenthèses

Identification des motifs cycliques à double brin

10. Tant que T[i] est différent de «) » et T[j] est différent de « (»
11. Si quantitéPB est inférieur ou égal à PBMAX alors
12. Si T[i] équivaut à « (» et T[j] équivaut à «) » alors
13. quantitéPB ← quantitéPB + 1;
14. $quantité_i$ ← $quantité_i$ + 1;
15. $quantité_j$ ← $quantité_j$ + 1;
16. i ← i +1, j ← j -1;
17. Sinon
18. Si T[i] équivaut à « . » alors
19. $quantité_i$ ← $quantité_i$ + 1;
20. i ← i +1;
21. Sinon si T[j] équivaut à « . » alors

22. quantité$_j$ ← quantité$_j$ + 1;
23. j ← j -1;
24. Fin de Si
25. Fin de Si
26. Fin de Si

Ajout des motifs cycliques à double brin dans la chaîneCycles

27. Si quantitéPB équivaut à PBMAX alors
28. ncmID ← quantité$_i$ + « _ » + quantité$_j$;
29. chaîneCycles ← chaîneCycles + ncmID;
30. Réinitialiser les variables quantitéPB, quantité$_i$ et quantité$_j$ à 0;
31. i ← i -1, j ← j +1;
32. Fin de Si
33. Fin de Tant que

Ajout d'un motif cyclique à simple brin dans la chaîneCycles, s'il y a lieu

34. ncmID ← quantité$_i$ + quantité$_j$;
35. Si ncmID est différent de 2 alors
36. chaîneCycles ← chaîneCycles + ncmID;
37. Fin de Si
38. Afficher chaîneCycles;
39. **Fin**

Annexe D : La base de données de *bdMotifs*

Les lignes de codes ci-dessous permettent de créer les bases de données pour *MC-Motifs* dans *MySQL* : *bdMotifs* pour la base de données des motifs fonctionnels d'ARN (*bdMotifs* utilise une seconde base de données nommée *TYPEMOTIF* pour identifier le type d'élément structural du motif).

```
--
-- Host: mysql        Database: bdMotifs
----------------------------------------------------------------------
--
DROP DATABASE IF EXISTS bdMotifs;
CREATE DATABASE bdMotifs;
USE bdMotifs;


CREATE TABLE `bdMotifs` (
      `motif_id` int(11) NOT NULL auto_increment,
      `motif_name` varchar(100) NOT NULL default '',
      `motif_type` varchar(10) default NULL,
      `sequence` varchar(100) NOT NULL default '',
      `dots_brackets` varchar(100) NOT NULL default '',
      `ncm_blocs` varchar(100) NOT NULL default '',
      `description` varchar(100) default NULL,
      PRIMARY KEY  (`motif_id`)
)


CREATE TABLE `TYPEMOTIF` (
      `type_id` varchar(10) NOT NULL default '',
      `description` varchar(30) default NULL
)
INSERT INTO table_name (type_id, description)
VALUES ("A", "Helice");
INSERT INTO table_name (type_id, description)
VALUES ("B", "Hairpin loop");
INSERT INTO table_name (type_id, description)
VALUES ("C", "Internal loop");
```