Université de Montréal

Direction des bibliothèques

AVIS

Ce document a été numérisé par la Division de la gestion des documents et des archives de l'Université de Montréal.

L'auteur a autorisé l'Université de Montréal à reproduire et diffuser, en totalité ou en partie, par quelque moyen que ce soit et sur quelque support que ce soit, et exclusivement à des fins non lucratives d'enseignement et de recherche, des copies de ce mémoire ou de cette thèse.

L'auteur et les coauteurs le cas échéant conservent la propriété du droit d'auteur et des droits moraux qui protègent ce document. Ni la thèse ou le mémoire, ni des extraits substantiels de ce document, ne doivent être imprimés ou autrement reproduits sans l'autorisation de l'auteur.

Afin de se conformer à la Loi canadienne sur la protection des renseignements personnels, quelques formulaires secondaires, coordonnées ou signatures intégrées au texte ont pu être enlevés de ce document. Bien que cela ait pu affecter la pagination, il n'y a aucun contenu manquant.

NOTICE

This document was digitized by the Records Management & Archives Division of Université de Montréal.

The author of this thesis or dissertation has granted a nonexclusive license allowing Université de Montréal to reproduce and publish the document, in part or in whole, and in any format, solely for noncommercial educational and research purposes.

The author and co-authors if applicable retain copyright ownership and moral rights in this document. Neither the whole thesis or dissertation, nor substantial extracts from it, may be printed or otherwise reproduced without the author's permission.

In compliance with the Canadian Privacy Act some supporting forms, contact information or signatures may have been removed from the document. While this may affect the document page count, it does not represent any loss of content from the document.

Université de Montréal

Analyse de motifs d'ARN

Par

Louis-Philippe Lavoie

Département d'Informatique et de Recherche Opérationnelle

Institut de Recherche en Immunologie et Cancer

Faculté de Médecine

Mémoire présenté à la Faculté de Médecine

en vue de l'obtention du grade de *Magister Scientiæ* (M. Sc.)

en Bioinformatique

Août 2007

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES
10 JAN. 2008
Université de Montréal

© Louis-Philippe Lavoie, 2007

Université de Montréal
Faculté de Médecine

Ce mémoire intitulé :

Analyse de motifs d'ARN

Présenté par :

Louis-Philippe Lavoie

a été évalué par un jury composé des personnes suivantes :

Pascal Chartrand
Président-rapporteur

François Major
Directeur de recherche

Serguei Chteinberg
Membre du jury

Résumé

L'acide ribonucléique est une chaîne composée d'unités toute simples formant des ensembles d'une grande complexité tant structurelle que fonctionnelle, deux aspects qui sont intimement liés. Des nucléotides à la molécule, il est pertinent de se demander si un niveau médian de complexité ne pourrait être atteint par la formulation d'un *méta-élément* qui représenterait une étape intermédiaire vers la compréhension de la formation des structures. Avec la publication de structures d'ARN augmentant de façon quasi-exponentielle, la réalisation que ces structures sont justement composées d'une multitude de ces éléments, appelés *motifs*, provoqua une avancée soudaine dans le domaine.

Ce mémoire porte essentiellement sur des travaux reliés à l'étude de ces motifs. Une présentation de l'acide ribonucléique, la méthodologie de son analyse et du concept de motifs d'ARN tiennent lieu d'introduction. La partie principale fait ensuite état, sous forme d'article, des résultats de la recherche et la caractérisation d'un de ces motifs, la tétraboucle dite « YNMG » qui est un élément présent dans plusieurs structures d'ARN. Cette analyse démontre que ce motif, mais aussi l'ARN en général, peut facilement ajuster sa séquence et même sa structure pour répondre aux contraintes locales, sans impact significatif sur l'environnement global. Une application développée expressément pour l'étude de motifs d'ARN, et utilisée dans les travaux présents, est ensuite décrite. Finalement, un article relatant une analyse de motifs d'ARN dans les virus à laquelle l'auteur à contribué est joint en annexe.

Mots-clés : tétraboucle, UNCG, YNMG, cycles, MC-View, visualisation de motifs.

Abstract

Ribonucleic acid is a chain composed of deceptively simple building blocks, folding into assemblages of great complexity as much structurally as functionally, two aspects intricately connected. Between the nucleotides and the molecule, it is worthwhile to speculate on the existence of an intermediate level of complexity embodied by a *meta-element*, a stepping stone toward full comprehension of structural folding. With the publication of high-quality RNA structures increasing at a seemingly exponential rate, the discovery that those structures are indeed made up of a myriad of these elements, labeled *motifs*, has contributed new impetus to the field.

This document is essentially about studies of these motifs. A presentation of ribonucleic acid, the associated research methodology, and the motif concept are offered as introduction. The body of the document then relates, in article form, the results of the study and characterization of one type of motif, the « YNMG » tetraloop, found in a large number of RNA structures. This analysis demonstrates that the motif, and RNA in general, can easily alter its sequence or even structure to accommodate local constraints, with no significant effect on its global environment. Software created expressively for RNA motif analysis, and used in the present work, is described next. Last, an article reporting on viral RNA motif analysis, to which the author has contributed, is submitted in an appendix.

Keywords: tetraloops, UNCG, YNMG, cycles, MC-View, motif visualisation

Table des matières

Liste des tableaux

Liste des figures

Liste des principaux sigles et abréviations

A : Adénosine

ADN : Acide désoxyriboNucléique

ARN : Acide riboNucléique

ARNm : ARN messager

ARNr : ARN ribosomique

ARNt : ARN de transfert

C : Cytidine

G : Guanosine

H : Face Hoogstein d'une base

IUPAC : International Union of Pure and Applied Chemistry

MC- : Macromolecular Conformation (préfixe de la suite d'outils du laboratoire)

O2' : Atome d'oxygène sur le sucre ribose d'un nucléotide

O2P : Atome d'oxygène sur le groupement phosphate d'un nucléotide

P : Purine

PDB : Protein Data Bank

RMN : Résonance magnétique nucléaire

RMSD : *Root mean square deviation*

S : Face Sucre d'un nucléotide

U : Uridine

W : Face Watson-Crick d'un nucléotide

WC : Appariement de type Watson-Crick canonique

W/W : Appariement générique entre faces Watson

Y : Pyrimidine

À Isabelle.

Remerciements

Je tiens à remercier mon directeur de recherche, François Major, pour le support, les discussions et les barbecues. Merci à tous les membres du labo pour la bonne entente. Je dois aussi remercier l'Université et les IRSC pour la bonne idée qu'ils ont eue en créant le programme de bourses biT, sans quoi je n'aurai pu compléter mon diplôme dans les deux ans prescrits. Il faut aussi souligner le travail extraordinaire, le dévouement et l'initiative d'Élaine Meunier, qui est l'archétype même du phare dans la tempête administrative.

Merci tout spécial à Emmanuelle Permal pour les encouragements, les rires, et les gâteaux.

Merci à Pascal de me montrer le chemin.

Merci à Isabelle. Pour trop de choses qu'il serait impossible d'énumérer.

Introduction

Les travaux de maîtrise qui seront présentés dans ce mémoire portent sur l'analyse structurale de motifs présents dans l'acide ribonucléique (ARN). Pour bien comprendre la nature et la portée des résultats, il est essentiel d'introduire certains concepts tel le rôle de l'ARN et les principaux éléments avec lequel il interagit dans la cellule, ainsi que des notions d'analyse structurale. Suivront ensuite une introduction des motifs d'ARN et de celui qui m'a tout particulièrement intéressé, le motif YNMG.

Les deux chapitres suivants présentent des articles qui font état des résultats de l'analyse du dit motif, et d'un logiciel crée expressément pour l'analyse de motifs d'ARN. En annexe après la conclusion est placé un troisième article rapportant les travaux d'une collègue sur les motifs d'ARN dans les virus, travaux auxquels l'auteur a participé.

Les mots soulignés dans le texte sont définis dans le lexique en fin de document.

A moins d'indication contraire, les images de structures ont étés réalisées avec PyMOL (DeLano 2002) et GIMP (*The GNU Image Manipulation Program*).

"The most beautiful thing we can experience is the mysterious; it is the source of all true art and science"

- Albert Einstein

1. L'acide ribonucléique

La chimie organique se distingue de la chimie générale par la présence des atomes de carbone, qui en plus d'être assez abondant (4e élément en abondance dans l'univers) possède des propriétés d'interactions atomiques particulières propice au développement d'une variété impressionnante de molécules (plus de 10 millions de composés du carbone sont connus). Le carbone peut former des liaisons covalentes extrêmement stables, simples, doubles ou triples avec une grande variété d'autres éléments ou avec lui-même, et sa réactivité aux températures moyennes sur Terre est telle que les composés sont stables, mais relativement faciles à modifier par plusieurs mécanismes de réaction. Par force de temps et d'essais, la vie s'est développée autour du carbone et des autres éléments présents sur Terre pour mener à la biodiversité que l'on connait aujourd'hui. La plupart des plastiques commerciaux sont des polymères d'unités de carbone, d'oxygène et d'azote. Les hydrocarbures (carburants, pétrole, ...) sont par définition des chaînes composées de carbone et d'hydrogène. Les polymères de carbone sont non seulement à la base de la vie mais aussi une partie importante de notre industrie. Ainsi, la phénoménale complexité de la chimie du carbone à donné naissance à l'Homme, et lui à fourni de quoi exercer sa curiosité pour encore plusieurs siècles.

Deux groupes de polymères du carbone sont particulièrement présents en biochimie : Les protéines et les acides nucléiques. De ces derniers on compte deux types : l'acide désoxyribonucléique (ADN) et l'acide ribonucléique (ARN). Alors que l'ADN est habituellement vu comme le dépositaire de l'information génétique et les protéines sont les principaux effecteurs de l'activité biochimique, l'ARN était jusqu'à tout récemment considéré simplement comme un intermédiaire éphémère entre les deux autres. Ces trois bio-polymères sont en effet reliés par ce qu'on appelle le *dogme central de la biologie moléculaire* (Crick 1958 & 1970), qui énonce en un concept simple l'ordre fondamental de la vie :

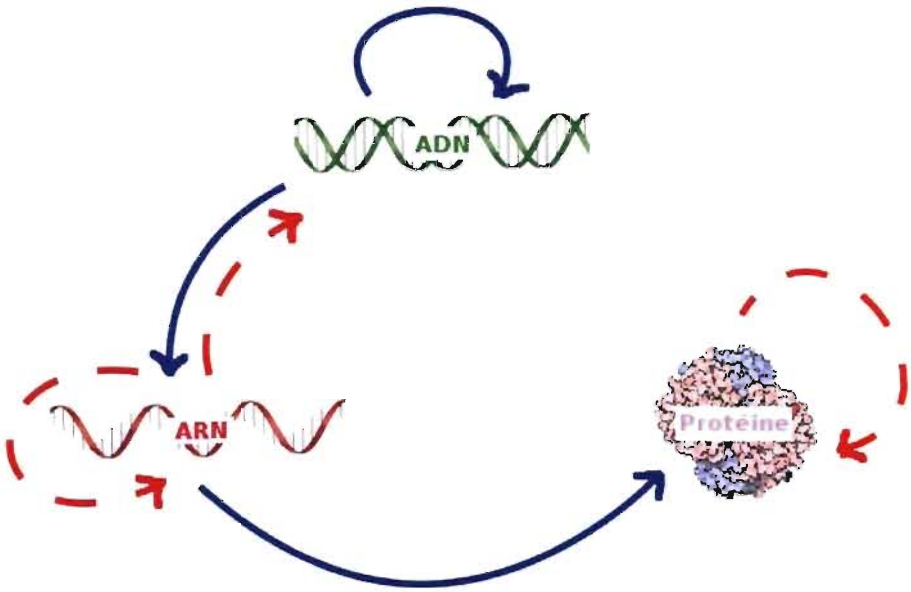


Figure 1 : Le *dogme central de la biologie moléculaire* montre la relation entre ADN, ARN et Protéines (ici, l'hémoglobine, image de David S. Goodsell tirée de « Molecule of the Month » du RCSB Protein Data Bank). Les flèches bleues indiquent le dogme tel qu'il à longtemps été énoncé. Les dernières décennies ont démontrés que d'autres voies étaient possibles (flèches rouges). Les voies sont expliquées dans le texte de cette page.

À l'origine, cette relation était considérée comme unidirectionnelle (voir **figure 1**, flèches bleues) : L'ADN se réplique et est transcrit en ARN, qui est traduit en protéines, ces dernières étant alors les pierres d'angle de la vie. Mais les dernières décennies ont mis en lumière de nouveaux rôles pour l'ARN (flèches rouges). L'ARN peut ainsi être réintégré en ADN comme dans le cas de la *rétro-transcription*, utilisé par exemple par le virus HIV, ou le cas des *rétrotransposons*. Certains virus peuvent aussi répliquer leur génome d'ARN sans passer par l'étape d'ADN. La découverte des prions (Prusiner 1982), avec leur pseudo-réplication, pourrait être considéré comme une autre transgression au dogme original car les protéines sont ainsi porteuses d'information (mais *structurelle*) et peuvent propager cette information. Nous verrons plus loin la structure de l'ARN plus en détail.

1.1 Rôles et importance de l'ARN

Ainsi, le rôle de l'ARN dans la cellule est maintenant reconnu comme étant beaucoup plus important que ne le laisse croire le *dogme*. L'ARN offre en effet les caractéristiques à la fois de l'ADN (support d'information génétique) et des protéines (activité catalytique, régulation d'expression), comme il sera vu plus en détail au cours de ce chapitre.

Dans les années 1980, les équipes de Thomas R. Cech et Sydney Altman co-découvrent les premiers ribozymes (Kruger et al. 1982 ; Guerrier-Takada et al. 1983), ce qui leur vaudra conjointement le prix Nobel de chimie en 1989. Ce type d'ARN démontre des capacités catalytiques (parfois même auto-catalytiques), jusqu'ici l'apanage exclusif des protéines. Il apparait aujourd'hui que bon nombre de mécanismes cellulaires importants, y compris une grande partie de la synthèse de protéines, sont catalysés ou orchestrés par des brins d'ARN. Voici des exemples parmi les plus connus :

- Le *ribosome*, un complexe de protéines et d'ARN à la base de synthèse des protéines, possède un cœur catalytique formé d'ARN.

- Lors de la création de protéines, les acides aminés sont apportés aux ribosomes par des molécules d'ARN (les ARN de transfert). Les ARN messagers dictent la séquence de la nouvelle protéine.

- L'épissage des introns dans les ARN messagers est effectué par le *spliceosome*, un autre complexe protéine-ARN.

- La machinerie d'expression des gènes pourrait être contrôlée par des petits brins d'ARN, le phénomène d'ARN d'interférence (décrit plus bas).

- *Xist*, un transcrit d'ARN de 18 kb est nécessaire pour l'inactivation du chromosome X chez les mammifères (Ng et al. 2007).

1.1.1. ARN non-codant

Le nombre de gènes codant pour des protéines chez l'humain (H. sapiens), estimé à environ 20 000 (IHGSC 2004 ; Pennisi 2007), est largement inférieur à ce qu'on pourrait s'attendre par la complexité et la taille du génome, qui même à 3.2 milliards de paires de bases est loin derrière le poids lourd *Amoeba Dubia*, un prokaryote de 670 milliards de paires de bases (Ussery & Hallin 2004). Alors qu'on a longtemps considéré que la majorité du génome d'ADN était composé de rebuts généré par l'évolution ou de régions silencieuses (« *junk DNA* »), on s'aperçoit aujourd'hui que de nombreux segments remplissent en fait des fonctions de régulation de l'expression génétique et du métabolisme. Lorsque le transcrit d'ARN ne mène pas à la synthèse d'une protéine, on parle d'ARN non-codant et la séquence d'ADN dont il est issu est parfois appelée gène d'ARN. Des études suggèrent que la quasi-totalité des transcrits d'ARN produits, soit 98%, sont non-codants (Mattick 2001) et on estime qu'il existe, chez la souris, environ 28 000 ARN non-codants (Liu et al. 2006). C'est donc la quasi-totalité de l'ARN excepté l'ARN messager. Mais non-codant n'équivaut pas à non-fonctionnel ! Bien que la nature et le mécanisme d'action de la plupart des ARN non-codants ne soient pas encore élucidés, leur importance ne fait plus aucun doute et nous verrons quelques exemples.

1.1.2. Types de molécules d'ARN

Il existe plusieurs types de molécules d'ARN. Le plus connu, celui auquel le *dogme central* fait référence, est l'ARN messager (ARNm) (voir **figure 1**). Il est une réplique fidèle mais complémentaire de l'ADN, c'est-à-dire que les nucléotides qui composent sa séquence sont celles qui s'apparient à l'ADN duquel il est issu. C'est le seul type qui est considéré comme étant « codant ». L'ARNm contient des régions non-traduites en 5' et 3' et des introns, qui peuvent contenir des régions de régulation d'expression, de stabilité, de localisation et autres. Déjà on note donc la présence d'ARN non-codant, et ces régions peuvent présenter une certaine structure (dite *secondaire*) ainsi que des régions biologiquement actives.

Tous les autres types forment le groupe des ARN non-codants. Un membre important est celui des ARN de transfert (ARNt). Ils sont impliqués dans la formation des protéines en transportant les acides aminés appropriés vers le cœur du ribosome pour qu'ils soient transférés sur la chaîne. Les ARNt présentent une structure déjà plus complexe.

Un troisième type de molécules d'ARN important à définir est l'ARN ribosomal. C'est celui qui forme les ribosomes, la principale machine de production des protéines, et comme ce sont parmi les plus grosses structures connues impliquant un acide ribonucléique (environ 3000 unités), ils sont particulièrement indiqués pour étudier la structure de l'ARN. De par sa fonction hautement critique, l'ARN ribosomal est parmi les transcrits d'ARN les plus conservés à travers de toutes les espèces vivantes, et est donc aussi très utilisé en phylogénie.

Les premières structures tridimensionnelles de ribosomes n'ont été disponibles qu'en 2000 (Ban et al. 2000), et encore aujourd'hui l'éventail est limité par rapport aux protéines (Berman et al. 2000). Bien qu'il existe plusieurs centaines de séquences disponibles pour l'étude, on ne compte que quatre espèces dont les ribosomes ont été cristallisés : *Haloarcula Marismortui* (archaebactérie) (Ban et al. 2000), *Thermus Thermophilus* (bactérie) (Wimberly et al. 2000), *Escherichia Coli* (bactérie) (Vila-Sanjurjo et al. 2003) et *Deinococcus Radiodurans* (bactérie) (Harms et al. 2001). On compte plusieurs copies avec différents antibiotiques, ARNt et/ou ARNm et il est donc possible de voir certains changements structuraux selon l'état du ribosome ou le ligand.

1.2 L'ARN comme sujet d'étude ; impact & applications

L'étude de l'ARN suscite de plus en plus d'intérêt en recherche, pour en élucider les mécanismes auxquels il est relié mais aussi d'un point de vue thérapeutique. Par les exemples suivant et plusieurs autres, il est facile d'illustrer l'importance d'une recherche soutenue de l'ARN. Du fait de l'éveil plutôt récent de la science à la complexité de cette molécule, c'est encore un domaine en pleine expansion et il est excitant d'y prendre part.

1.2.1. Cible thérapeutique

Déjà depuis quelques temps, des antibiotiques existent qui ciblent spécifiquement les ribosomes des procaryotes, structurellement assez différents du ribosome eukaryote pour offrir un vecteur de traitement fiable et sans trop d'effets secondaires (Recht et al. 1999). Cependant, ce type de traitement n'est pas encore très spécifique et un accroissement des connaissances les structurales de l'ARN promet de raffiner les sites liaisons, ouvrant la porte à une palette plus variée d'antibiotiques spécifiques à un seul type de bactérie ou même d'eucaryotes pathogènes.

1.2.2. Virus à génome d'ARN

Ce serait une faute que de passer sous silence une autre classe importante de structures d'ARN, celle composée de génomes de virus. Plusieurs types ont évolué pour exploiter des niches particulières de la machinerie de transcription, et nous retrouvons aujourd'hui des génomes double-brins, simple-brin positif et simple-brin négatif. Ces deux dernières catégories se distinguent par l'existence d'un génome directement fonctionnel (donc similaire à un ARN messager) ou bien complémentaire, qui doit être retranscrit par une polymérase.

Les *picornavirus* sont une famille de virus à génome d'ARN simple-brin positif qui inclut (entre autres) les *poliovirus* et autres *enterovirus*, le virus de l'hépatite A ainsi que les *rhinovirus*, eux-mêmes impliqués dans le rhume commun et plus de la moitié des infections

des voies respiratoires (Greenberg & Obin 2003). Les *coronavirus* ont aussi un génome d'ARN simple-brin positif. Le membre le plus connu est sans doute celui qui cause le SRAS qui créa une épidémie en 2003. D'autres virus à génome d'ARN notables causent la fièvre dengue, la fièvre du Nil, l'hépatite C, la fièvre jaune et l'influenza, pour ne nommer que les maux les plus sérieux. Il est donc clair que ce sont des cibles de recherches extrêmement intéressantes, tant au niveau de leur structure que de leurs interactions avec des protéines endogènes ou virales. Le motif étudié au chapitre 2 est un élément vital du génome des *picornavirus*.

1.2.3. Transcriptome et épigénome

Comme vu précédemment, la grande majorité des transcrits d'ARN sont non-codants et servent potentiellement à la régulation de l'expression et du métabolisme. L'ensemble des transcrits d'ARN d'une cellule est appelé transcriptome, par similitude au mot *génome*. Un concept relié est l'épigénome, qui est l'état d'expression des gènes tel que caractérisé par des phénomènes propres à l'expression génique et non des changements de la séquence du génome elle-même : méthylation de l'ADN, structure de la chromatine, acétylation des histones (Reik & Walter 2001 ; Bird 2002 ; Li 2002). Ces états sont plus facilement modifiables par l'environnement (stress, diète, etc.) et peuvent modifier le phénotype sur plusieurs générations (Bernstein et al. 2007) ou causer des maladies y compris le cancer (Jones & Baylin 2007). Il s'agit donc d'une forme d'hérédité variable sans modification de l'ADN, ce qui n'était pas pris en compte par le modèle de l'hérédité de Mendel. Ces deux domaines représentent donc des champs d'études à très fort impact et intimement liés à l'étude de l'ARN (Bernstein & Allis 2005).

1.2.4. ARN d'interférence

L'ARN d'interférence (ARNi), très probablement parmi les avenues les plus prometteuses de la recherche biochimique des récentes années, en est à ses balbutiements et pour le moment la technique n'est pas assez précise pour être utilisée à des fins thérapeutiques à grande échelle (Qiu et al. 2005).

Dans les années 90, des chercheurs ont remarqué que des brins d'ARN pouvaient inhiber l'expression d'un gène chez les plantes (Napoli et al. 1990 ; Romano & Macino 1992). Le mécanisme suscita beaucoup d'intérêt mais demeura inconnu jusqu'à la publication des travaux de l'équipe de Fire et Mello (Fire et al. 1998), qui identifiait des molécules d'ARN double-brin comme l'élément inhibiteur. L'ARN d'interférence est un mécanisme cellulaire conservé par lequel des tige-boucles d'ARN viennent réguler l'expression de gènes, et les applications potentielles en médecine et en biotechnologie sont énormes. Cette découverte a valu à Andrew Fire et Craig C. Mello le prix Nobel de médecine en 2006.

Dans un surprenant rebondissement, des études récentes (Li et al. 2006 ; Janowski et al. 2007) ont aussi fait état d'*activation* de gène par de petits ARN (*ARNa*, par analogie à ARNi). Bien que le mécanisme soit encore disputé, si elle s'avère exacte la nouvelle pourrait bien multiplier encore plus l'enthousiasme pour les molécules d'ARN.

1.2.5. Théorie « RNA world »

De transitoire, certains avancent maintenant que l'ARN fut possiblement le précurseur de la vie, la première molécule capable de se répliquer et de se propager avant l'apparition de l'ARN et des protéines, selon l'hypothèse du « RNA World » (Gilbert 1986). Que cette théorie soit fondée ou non, la communauté scientifique réalise aujourd'hui que l'acide ribonucléique est un acteur beaucoup plus important qu'il était imaginé au moment de l'énoncé original du *dogme central*.

1.3 Principes de structure d'ARN

1.3.1. Code IUPAC

Pour désigner les bases d'ARN, un alphabet standard a été décidé par le regroupement *IUPAC (International Union of Pure and Applied Chemistry)*. Cette convention sera utilisée dans ce document.

Tableau I : Liste des symboles IUPAC pour l'ARN.

Code	Signification	Code	Signification
A	Adénine	G	Guanine
C	Cytosine	U	Uracile
M	A ou C	S	C ou G
R	A ou G	Y	C ou G
W	A ou U	K	G ou U
V	A, C ou G	H	A, C ou U
D	A, G ou U	B	C, G ou U
N	A, C, G ou U		

1.3.2. Nomenclature

Les acides nucléiques sont des polymères d'unités appelées *nucléotides*, elles-mêmes formées d'un *ribose* – un anneau de sucre à quatre carbones et un oxygène, d'une *base azotée* (ou *nucléobase*) variable, et d'un groupement phosphate. La combinaison d'une base azotée et d'un ribose sans le groupe phosphate est appelée *nucléoside* (voir **figure 2a**). Alors que le phosphate se retrouve dans toutes les unités et la plupart des bases sont communes, l'ADN et l'ARN se différencient par leur ribose : L'acide ribonucléique possède un groupement hydroxyle supplémentaire sur le carbone en position 2' du ribose, ce qui le rends plus réactif, et par le fait même plus fragile à l'action de certains solvants ou

molécules organiques. En tant que support d'information génétique, l'ARN est donc moins stable que l'ADN. De plus, en partie grâce à son groupement supplémentaire, l'ARN est capable d'activité (auto-) enzymatique (Krasilnikov et al. 2004), au contraire de l'ADN.

Les quatre bases de l'ARN sont : adénine, cytosine, guanine et uracile (voir **figure 2b**). Les trois premières se retrouvent aussi dans l'ADN et l'uracile y est remplacé par la thymine. La cytosine, l'uracile et la thymine sont des dérivés de pyrimidines, soit des cycles de six atomes avec un azote en position 1 et 3 ($C_4H_4N_2$). L'adénine et la guanine sont dérivés des purines, soit des composés formés d'un anneau de pyrimidine et un d'imidazole, un cycle de cinq atomes aussi avec deux d'azote ($C_3H_4N_2$). Les bases sont reliées à leur ribose par le lien glycosidique, entre l'atome C1' du ribose et l'atome N9 (chez les purines) ou N1 (pyrimidines) (aussi en orange dans la **figure 3**).

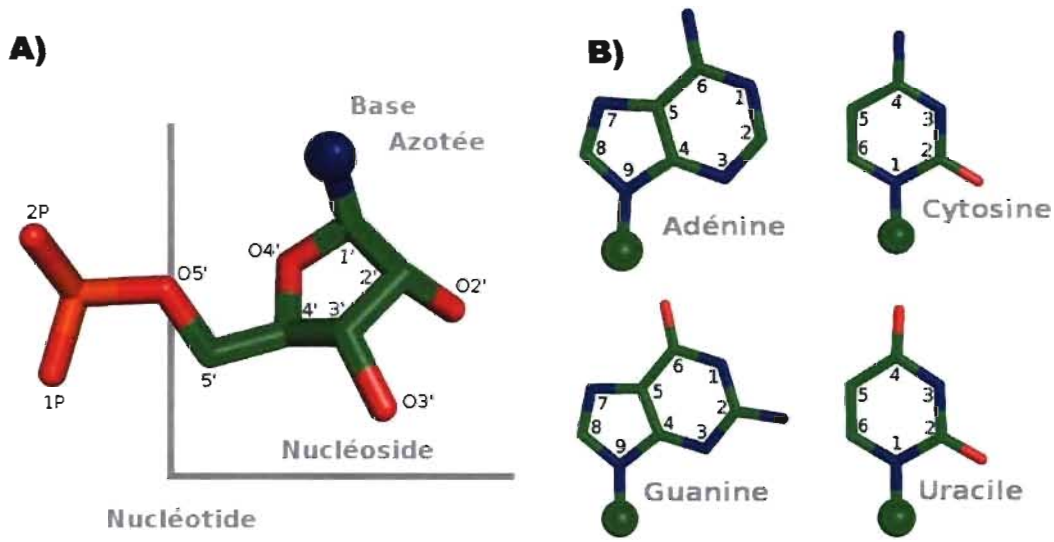


Figure 2 : A) Un nucléotide, formé du groupe phosphate, du ribose et d'une baze azotée (représentée par une sphère bleue). B) Les quatre bases de l'ARN, affichant la numérotation standard des atomes. Le carbone C1', point d'attache sur le ribose, est représenté comme une sphère verte. Dans les deux cas les atomes sont colorés pour identification (carbone : vert ; azote : bleu ; oxygène : rouge ; phosphate : orange).

Les nucléotides associés aux nucléobases sont adénosine, cytidine, guanosine, uridine et thymidine. Les monomères formant la chaîne d'acides nucléiques sont reliés entre eux par un lien phosphodiester entre les atomes O3' et un phosphate (1P et 2P sont équivalents). Il est donc possible de parcourir le polymère en sautant sur les atomes 5'-O5'-P-O3'-3'.

À cause de contraintes stériques dues aux angles des liens covalents présents dans le cycle, la géométrie du ribose n'est pas planaire. Ils possèdent cinq angles de torsion (en cyan dans la **figure 3**) mais ces angles sont interdépendants et peuvent être définis de façon non-ambigue par un angle de *pseudorotation*, ρ (Altona & Sundaralingam 1972).

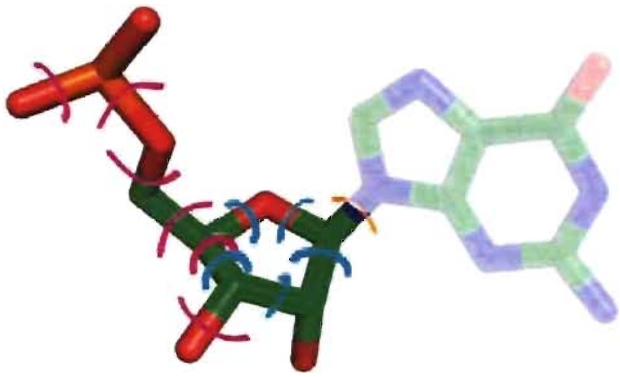


Figure 3 : Les angles de torsion (en magenta) fixant la position d'un nucléotide (ici une guanine). L'un d'entre eux est redondant avec un des angles du ribose (en cyan). Lorsque le squelette est fixé, il ne reste que l'angle sur le lien glycosidique (orange) pour déterminer la position de la base.

La notation recommandée par le regroupement *IUPAC* (*International Union of Pure and Applied Chemistry*) pour annoter la conformation du ribose est la notation E/T : Si quatre atomes du cycle se trouvent dans un même plan, ce plan devient la référence et le ribose est dit en forme d'enveloppe (E). Sinon, le plan de référence est celui des trois atomes le plus près du plan moyen et la conformation est dite « twist » (T). Les atomes qui sont au-dessus du plan (du côté où la numérotation est dans le sens des aiguilles d'une montre) sont en exposé avant la lettre, et les autres atomes en indice après la lettre. Par exemple : $^{3}T_{2}$ ou ^{3}E. Une autre nomenclature aussi couramment utilisée, y compris dans ce document, est la notation *endo/exo*, où le premier terme désigne les atomes au dessus du

plan et le second ceux en dessous. L'exemple devient donc : C3'-endo/C2'-exo ou C3'-endo. La configuration C3'-endo est celle préférée dans la formation de doubles hélices d'ARN de type A, et est donc retrouvée la plus fréquemment dans les structures (Dickerson & Ng, 2001). La plupart des conformations sont toutefois possibles.

Contrairement aux riboses, les cycles des nucléobases sont planaires et les bases sont considérées comme des corps rigides. Les seuls points variables sont alors les angles de torsion sur les liens non membres d'un cycle, en plus de l'angle unique de pseudorotation du ribose. Le squelette phosphodiester présente donc six angles de torsion pour être complètement déterminé dans l'espace, plus un angle pour placer la base sur son ribose (en magenta et orange dans la **figure 3**). Cependant, des travaux semblent indiquer que certains des angles restants sont aussi interdépendants, varient très peu ou dans un intervalle précis (Schneider et al. 2004 ; Hershkovitz et al. 2006 ; Murray et al. 2003 & 2005).

La rotation de la base autour du lien glycosidique peut être qualifiée de manière générale par les termes *syn*, si la base est en direction du ribose et de son phosphate, et *anti* dans le case inverse.

Finalement, dans le but de bien spécifier les relations entre bases la nomenclature définit trois faces d'interaction (voir **figure 4**). Des symboles sont associés à chacune pour permettre une notation abrégée. La face Sugar (▲), du côté du sucre, la face Watson (●), qui forme les paire canoniques Watson-Crick, et la face Hoogstein (■) (Leontis & Westhof, 2001).

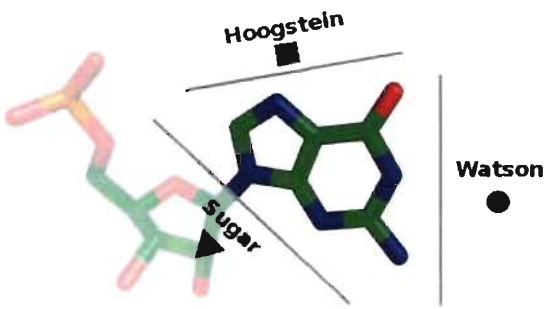


Figure 4 : Les trois faces d'interaction d'un nucléotide (ici une guanine), avec leur symbole correspondant.

1.3.3. Interactions entre nucléotides

Maintenant que les nucléotides, les unités composant les chaînes d'acides nucléiques, sont bien définies, il faut définir les types d'interactions possibles entres elles.

L'interaction la plus simple, qu'on oublie souvent de considérer tant elle est évidente, est l'adjacence, formée le long du squelette par le lien phosphodiester. C'est la plus facile à caractériser car elle ne comporte aucune ambiguïté (il y a présence ou non, sans valeurs intermédiaires)

Nous avons ensuite les empilements de bases. Les bases sont considérées comme planaires et rigides et nous pouvons donc définir un vecteur normal au plan de la base, selon le sens de la numérotation des atomes du cycle de pyrimidine. Avec ce vecteur qui fixe l'orientation de chaque base, il existe quatre façons de les superposer (voir **figure 5**) (Major & Thibault 2007). Si nous nommons deux bases A et B avec chacune leur vecteur normal, on peut alors empiler les bases de façon à ce que les normales soient : vers l'intérieur (*inward*), vers l'extérieur (*outward*) ou dans le même sens. Cette dernière façon présente une distinction des cas « B sur A » (*upward*) et « A sur B » (*downward*), qui dénotent l'orientation relative de l'empilement (5' ou 3'). De façon textuelle, on note aussi l'empilement avec les signes > et <.

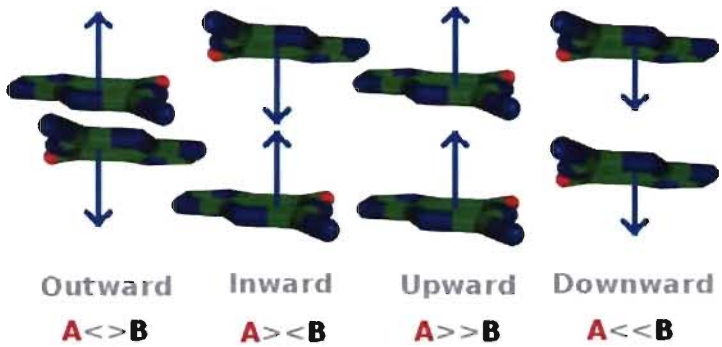


Figure 5 : Les quatre types d'empilements de bases. Le vecteur normal (en bleu) de chaque base détermine son orientation, et l'annotation d'empilement détermine ensuite de façon non-ambigüe l'orientation relative des deux bases (rouge et noir).

Comme les bases peuvent tourner librement autour du lien glycosidique, l'orientation de l'empilement n'est pas corrélée avec l'orientation relative des brins d'ARN des bases impliquées.

Le troisième et dernier type d'interaction et l'appariement. Il existe trois types d'appariements canoniques, soit les paires G:C, les paires A:U et nous considérons aussi les paires G:U (*wobble*) comme canoniques. Pour ces trois types, les faces en interactions sont les faces Watson. Tous les autres types, non-canoniques, peuvent aussi être décrits par leurs faces d'interaction (voir **figure 6**).

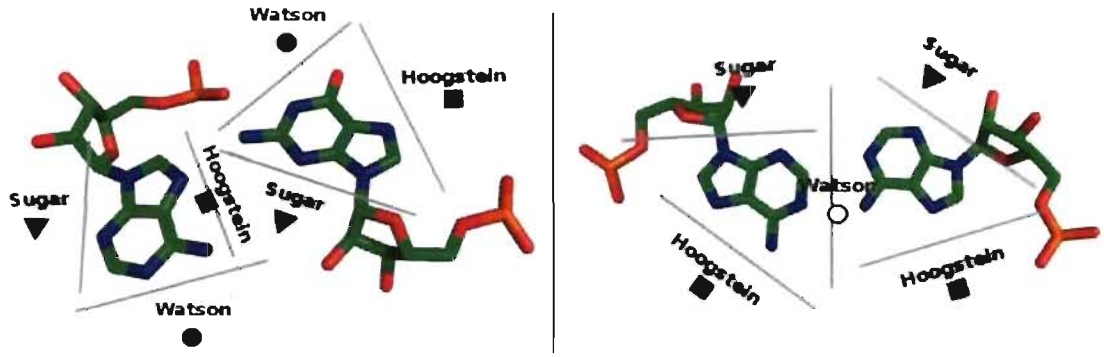


Figure 6 : Deux exemples d'appariements. A gauche, une paire A-G antiparallèle *trans* à travers les faces d'interactions H et S. À droite une paire A-A antiparallèle *cis* avec faces W/W. Les deux sont non-canoniques.

Pour qualifier d'avantage la relation d'appariement, deux autres concepts ont étés formulés. D'abord, comme ce type de relation place généralement les bases sur un même plan approximatif, nous pouvons préciser l'orientation des bases entre elles à l'aide de leur vecteur normal : une paire parallèle en est une ou les normales des deux bases pointent du même côté du plan d'appariement. Sinon, la paire est antiparallèle. Finalement, toujours en assumant que la paire forme un plan, il est possible d'annoter la position du ribose par rapport à un axe parallèle à ce plan (voir **figure 6**). Si les deux riboses sont du même côté, l'appariement sera désigné *cis*, et *trans* dans le cas contraire. Cette annotation peut être jointe aux symboles d'annotation des faces (▲,● et ■) qui seront noirs pour un appariement *cis* et blancs dans le cas *trans* (donc Δ, ○ et □).

Certaines paires, par exemple les canoniques, occupent un volume dans l'espace qui est équivalent d'une paire à l'autre. Dans ces cas, il est possible de les intervertir sans modifier la structure environnante et les deux appariements sont dits isostériques (Lescoute et al. 2005).

L'adjacence est une relation discrète, deux nucléotides sont ou bien liés de façon covalente ou bien ils ne le sont pas. L'empilement et l'appariement se présentent sur un intervalle de valeurs continues. Lorsqu'on annote des structures, il devient nécessaire d'utiliser une valeur limite pour décider si on considère une relation comme présente ou non. Gabb et al. (1996) ont publié les résultats d'une étude qui sert de référence pour les logiciels d'annotation de notre laboratoire, où la distance entre les cycles (imidazole et pyrimidine) ainsi que les angles entre les plans (angle dihédral) sont évalués pour prendre une décision sur l'annotation d'empilement.

De même pour les appariements, la décision d'annotation dépend d'études d'apprentissage machine réalisées antérieurement dans notre laboratoire (Gendron et al. 2001 ; Lemieux & Major 2002). Le biais qui découle de la prise de décision pour ces deux types d'annotation sera discuté plus en détail dans les prochaines sections de ce document.

1.3.4. Bases modifiées

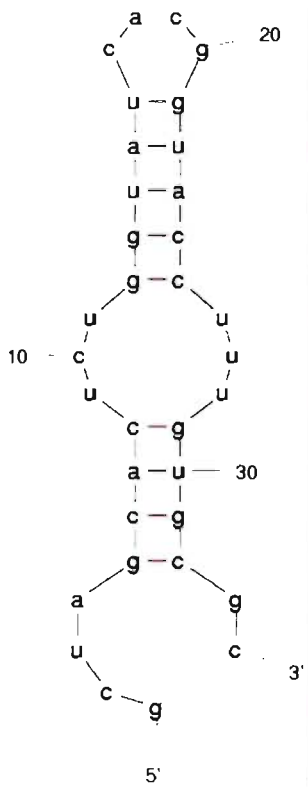
Divers voies chimiques, métaboliques ou autres (radiations, mutagènes) peuvent venir modifier les quatre bases typiques de l'ARN, et certaines de ces bases modifiées sont des éléments critiques de sites fonctionnels de molécules d'ARN, particulièrement dans les ARNt, comme la pseudouridine (symbole : ψ) (Limbach et al. 1994). D'autres comme la xanthine sont des produits de dégradation des bases principales. Il existe près d'une centaine de bases modifiées connues à l'heure actuelle, et lorsqu'elles sont présentes dans une structure elles en compliquent grandement l'analyse : Les modes d'appariement et même d'empilement précis de ces bases ne sont pas connus, et donc la plupart sont présentement ignorées par les méthodes d'annotation courantes.

2. Notions d'analyse structurale

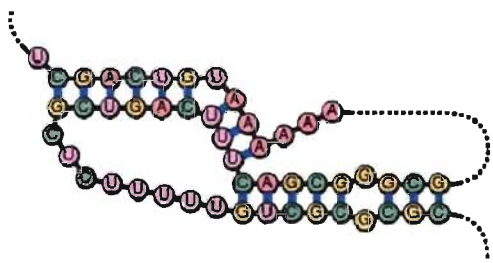
2.1 Structure primaire

Le terme structure primaire désigne la séquence d'un biopolymère. Pour les protéines particulièrement, il est possible d'inférer une certaine quantité d'information sur la forme finale à partir de la séquence, dont les régions adoptant des configurations en hélice-α ou en feuillet-β (Ginalski et al. 2005 ; Parisien & Major 2005). Pour l'ARN, on peut déduire avec un bon support d'informations (ex. : énergie minimale, alignements de multiples séquences, etc.) les paires susceptibles de former ensemble des appariements canoniques Watson-Crick, et donc déterminer les hélices – la structure secondaire (Xu et al. 2007). Les *pseudo-nœuds*, qui ne forment pas des tiges-boucles simples (voir **figure 7**) sont présentement les régions qui représentent les embûches les plus importantes pour la prédiction de structure secondaire et nécessitent des méthodes spécifiques (Huang & Ali 2007).

2.2 Structure secondaire



Au sens strict, un diagramme de structure secondaire est une représentation planaire de la séquence de façon à ce que les nucléotides formant des appariements canoniques entre eux soient à proximité (voir **figure 8**). Dans cette définition, on voit qu'on ne peut pas facilement représenter les pseudo-nœuds tout en gardant le diagramme planaire.



►**Figure 7 :** Un pseudo-nœud. Cet exemple ce trouve dans la télomérase humaine. (Image du domaine publique produite par Philip Ronan).

◄**Figure 8 :** Structure secondaire de la séquence consensus de la tige-boucle D inférée d'un alignement de 60 séquences de rhinovirus et enterovirus (Zell et al. 2002). Seuls les liens canoniques GC (rouge), AU (bleu) ou GU (vert) sont indiqués. Image générée avec MFold (Zuker 2003).

Parce qu'il est possible de la générer algorithmiquement avec une assez bonne fiabilité, la structure secondaire est la représentation la plus courante pour les études structurales. Cependant, certains résultats laissent croire que c'est plutôt la structure 3D qui est sujette à la pression évolutive, tel que démontré de façon remarquable par la publication des structures cristallographiques du domaine de spécificité de la RNase P de type A et B (Krasilnikov et al. 2004 ; Torres-Larios et al. 2005). Alors que les deux types sont structurellement et fonctionnellement similaires, les structures secondaires démontrent des différences significatives (Westhof & Massire 2004).

La structure secondaire stricte présente un intérêt accru si nous y ajoutons les interactions tertiaires (hors de la même hélice), ce qui est fait couramment lorsque nous nous intéressons à la structure tridimensionnelle (voir **figure 9**). Avec ou sans les relations tertiaires, la structure secondaire ne présente pas les empilements de bases.

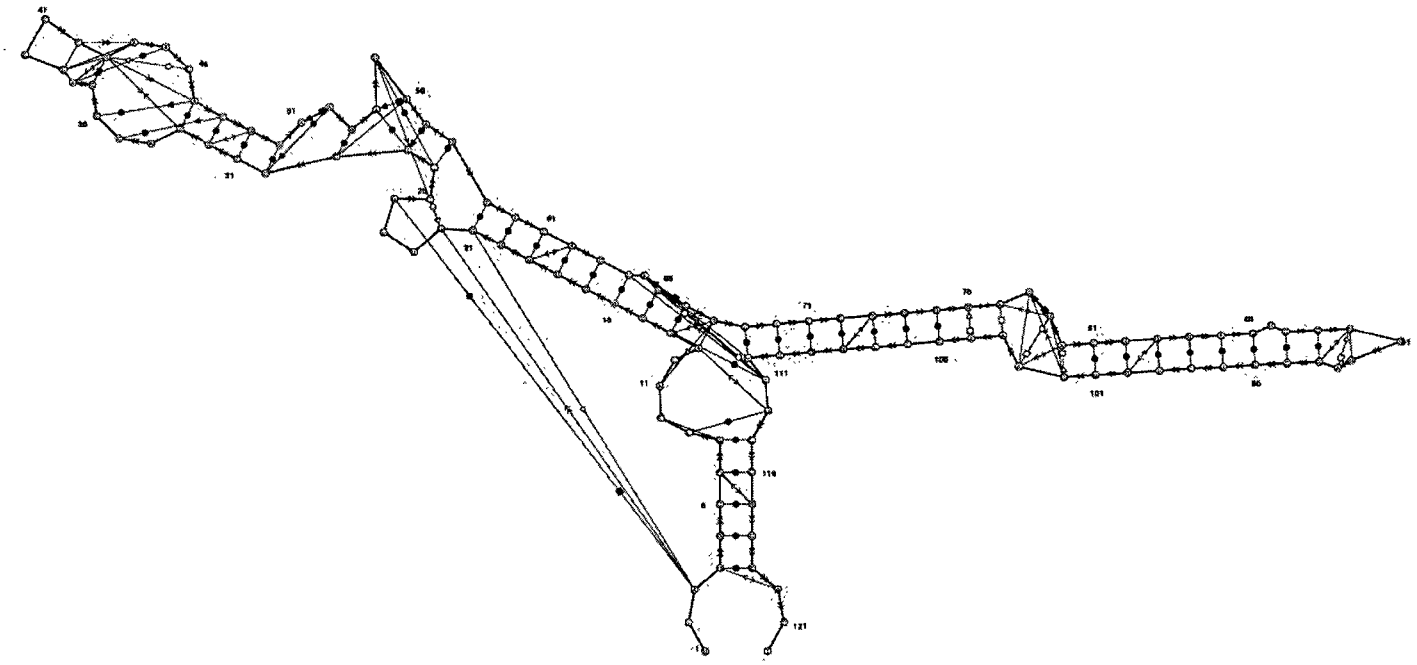
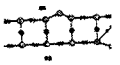
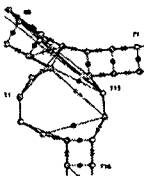


Figure 9 : Exemple de structure secondaire avec relations tertiaires (ici le brin 5S de *Haloarcula Marismortui*, tiré du PDB 1S72). Les symboles sont issus de la nomenclature décrite à la section précédente. Image générée avec *MC-View* (Lavoie et al., manuscrit disponible au chapitre 3).

2.3 Structure tertiaire

La structure tertiaire contient assez d'information pour représenter la molécule sous sa forme native tridimensionnelle. Certains séparent le terme *tertiaire* de *tridimensionnel* pour différencier un diagramme planaire avec interactions tertiaires, tel que vu en **figure 9**, d'une structure avec coordonnées 3D (X, Y, Z), respectivement. Encore ici, les annotations d'empilements ne sont possibles que si nous sommes en possession de la forme 3D. Nous distinguons habituellement quatre grandes classes d'éléments structurels, leur définition change peu de la structure secondaire à tertiaire.

Tableau II : Classes d'éléments de structure d'ARN

Type d'élément	Définition	Exemple
Tige-boucle	Brin d'ARN se repliant en deux pour formant une double hélice formée d'appariements canoniques, terminée par une boucle.	
Bourgeon	Un saut d'au moins un nucléotide dans la série d'appariements d'une même tige. Parfois aussi nommé « boucle interne » pour les sauts de plus de deux ou trois bases.	
Jonction	Les jonctions sont la rencontre de plus de deux hélices.	
Autre	Les régions sans structure distincte (habituellement les extrémités 5' et 3').	-

Un des buts de la biologie structurale est la prédiction de la structure tridimensionnelle directement à partir de la séquence, sans nécessairement passer par la structure secondaire ou bien en la calculant de façon implicite et automatique dans la méthode. Présentement, encore beaucoup d'ajustements manuels sont nécessaires pour obtenir des structures 3D (Shapiro et al. 2007).

2.4 L'ARN constitué de petits fragments récurrents

En plus d'étudier l'ARN au niveau des nucléotides (interactions, angles de torsion, alignements de séquences et de structures secondaires), la communauté scientifique s'est aussi efforcée de trouver un ordre à des niveaux plus élevés. En plus de l'avantage évident de diminuer la complexité du travail, la méthode avait jusqu'ici très bien fonctionné pour les protéines avec les descriptions des structures en termes de feuillets-β et d'hélices-α. Cet effort connait un certain succès pour l'ARN aussi.

Il est aujourd'hui largement accepté que l'ARN forme des molécules composées d'un amalgame de petits noyaux de nucléotides (Moore 1999 ; Hendrix et al. 2005 ; Leontis et al. 2006). Ces noyaux sont répétés et connectés ensemble pour former la structure globale de la molécule. Localement stables et présentant des interactions inter-nucléotides caractéristiques, ils accommodent tout de même une certaine flexibilité et un nombre relativement restreint de ces noyaux sont nécessaires pour reformer, en s'assemblant de façon combinatoire, la grande majorité des structures connues à ce jour.

Le terme normalement employé est « motif d'ARN». Il n'existe cependant pas qu'une seule définition précise car on peut caractériser divers aspects de la structure. Différents auteurs proposent différents motifs et réalisent leurs analyses sur des bases complètement distinctes, dans le but d'obtenir la théorie qui décrira avec succès l'ensemble des structures ou permettra la modélisation efficace. Nous verrons des approches alternatives à la section 2.9, après avoir introduit celle privilégiée au laboratoire. La présence de motifs répétés n'est plus disputée par personne, et de manière générale les motifs sont soit des fragments superposables entre eux soit des régions qui présentent le même type d'interactions (Moore 1999 ; Leontis & Westhof 2003). Aussi longtemps que les règles de repliement nous éluderons, toutes ces définitions sont à la fois valides car elles donnent des résultats intéressants.

2.5 Théorie des graphes ; application à l'ARN

Il faut maintenant introduire un concept qui est central à l'analyse de structure faite dans ce document, et qui nous vient de l'informatique et des mathématiques formelles. À partir de l'instant où nous sommes en possession d'une structure, qu'elle soit désignée primaire, secondaire, tertiaire ou tridimensionnelle, nous avons un ensemble de nucléotides avec des relations entre eux, ne serait-ce que la simple adjacence dans le cas d'une structure primaire. Il est alors possible de représenter cet ensemble de nucléotides sous forme de graphe de relations, composé de points (ou *vertex,* sommet, les nucléotides), et de relations entre ces points. On peut donc représenter toute structure d'ARN (aussi ADN, protéine) sous forme de graphe (voir **figure 10**). Chaque point et chaque relation peut avoir un nombre de propriétés illimitées, il est possible d'utiliser n'importe quelle propriété quantifiable. Dans le cas qui nous intéresse, les propriétés des points seront l'identité du nucléotide (A, C, G, U), sa conformation ribose, N-glycosidique, ses coordonnées 3D, etc. Les propriétés des relations indiqueront les faces impliquées et l'orientation relatives des deux partenaires, selon la nomenclature vue précédemment.

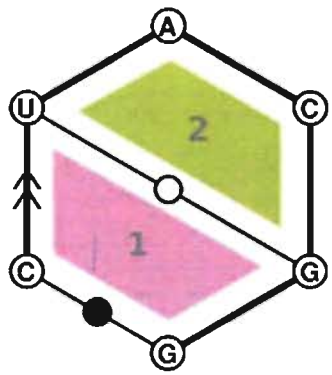


Figure 10 : Représentation simplifiée d'une tétraboucle YNMG, le motif qui nous à intéressé (voir chapitre 3). Les nucléotides et les relations sont indiqués selon la nomenclature énoncée dans le document. Les autres propriétés (orientation, conformation du sucre, etc.) ne sont pas dans ce genre de figure. Les liens d'adjacence sont représentés par des traits plus gras. L'appariement W/W entre les nucléotide U et G divise cette boucle en deux cycles de relations (rose et vert). Figure générée par *MC-View* (Lavoie et al., manuscrit disponible au chapitre 3).

2.6 Cycles

Dans la lancée de définition de motifs d'ARN, notre laboratoire n'est pas resté en marge. La théorie des graphes ouvre la porte à une définition neutre et mathématiquement élégante d'un fragment d'ARN : dans un graphe d'interaction, tout chemin dans le graphe circulaire et indivisible forme un cycle de nucléotides (voir **figure 10**). Les cycles deviennent donc des motifs, et des algorithmes existent pour diviser de façon complète un graphe en ses constituants de cycles (Horton 1987 ; Vismara 1997) qui peuvent ensuite être analysés globalement (Lemieux & Major 2006). Cette définition, comme toutes les autres, comporte des avantages et des inconvénients :

- (+) Très facile à manipuler algorithmiquement. Leur définition est simple et systématique, à l'inverse des motifs biologiques qui sont souvent définis de façon arbitraire et contextuelle. Les nouveaux cycles sont visibles immédiatement, au lieu d'avoir à chercher à la fois un motif et une définition de motif comme le font certains auteurs.

- (+) Traite tous les types d'interaction de façon égale, c'est-à-dire qu'il est possible de considérer par exemple seulement les empilements. Les définitions basées sur l'adjacence ne peuvent trouver les motifs montrant une insertion.

- (-) Détaché de la fonction biologique. Des efforts sont faits pour relier des cycles à leur fonction, mais la relation n'est pas claire.

2.7 Recherche de sous-graphes ; isomorphisme

A partir de là, il devient extrêmement aisé de rechercher des propriétés particulières dans le graphe par un algorithme de la théorie des graphes qui recherche les *isomorphismes*, c'est-à-dire les graphes équivalents (Ullmann 1976). Cette recherche est indépendante du type de propriété recherchée. Des exemples plausibles sont toutes les adénines reliées à des uraciles par un lien Watson-Crick, ou les guanines en conformation C3'-endo / *syn*. Il est possible de complexifier la recherche et déterminer tous les sous-ensembles (sous-graphes) d'une nature précise qui sont présent dans le graphe de départ (la molécule), et ainsi chercher rapidement des motifs particuliers dans n'importe quelle molécule, les comparer, les analyser.

2.8 Outils MC-*

En parallèle à l'élaboration d'une nomenclature précise, du concept d'annotation en graphe et de méthode de recherche par isomorphisme, plusieurs outils logiciels ont étés développés au laboratoire. Bon nombre de ces outils ont étés utilisés dans les travaux présentés dans ce mémoire et une brève description de chacun est donnée en page suivante.

Tableau III : Les logiciels de la suite d'outils du laboratoire

Outil	Fonction
MC-Core	Librairie de base, formant le cœur de tous les autres outils. La traduction en graphe est réalisée par ce module. Code source complet disponible gratuitement.
MC-Annotate	Affichage complet du graphe de relations et des propriétés d'une molécule. (Gendron et al. 2001)
MC-Search	Recherche de sous-graphes par isomorphisme (Olivier et al. 2005 ; Hoffmann et al. 2003)
MC-RMSD	Calcul de la déviation RMS selon une superposition optimale. Lorsque l'identité des nucléotides diffère, la RMSD se fait sur les atomes communs, selon l'algorithme de Kabsch (1976).
MC-Cycle	Énumération de l'ensemble des cycles dans une structure (Lemieux & Major 2006).
MC-View	Visualisation de plusieurs motifs à la fois dans une même structure. (voir chapitre 3, manuscrit en préparation).
MC-Fold	Prédiction de structure secondaire et tertiaire à partir de la séquence (Parisien & Major, article soumis).
MC-Seq	Prédiction par grammaire stochastique des séquences supportant une structure (St-Onge et al. 2007)
MC-Cons	Prédiction de structure secondaire consensus à partir d'un alignement de multiple séquences
MC-Sym	Modélisation de structures par satisfaction de contraintes (Major et al. 1991)

2.9 Autres approches

L'analyse structurale d'ARN est encore un domaine en développement, et plusieurs méthodes sont proposées de façon parallèle, pour ne pas dire concurrente, pour essayer de comprendre les règles de repliement de l'ARN et la prédiction de structure. Certaines de ces méthodes donnent des résultats très intéressants. Voici quelques unes des méthodes de recherche qui illustrent différents concepts de *motifs* faciles à manipuler de façon automatique :

COMPADRES est un algorithme qui effectue la recherche de motifs similaires en recherchant des segments de squelette phosphodiester (« *RNA worm* ») pouvant se superposer au fragment recherché (Wadley & Pyle 2004). Le serveur web DIAL raffine cette idée en proposant l'alignement avec possibilité de saut (gaps) (Ferrè et al. 2007). Murray et al. (2005) utilisent les angles de torsion des nucléotides.

Le site web RNA-As-Graphs présente les structures d'ARN comme deux types de graphes simplifiées, ou les hélices sont des points et les autres éléments des arêtes, ou vice-versa. Cette méthode offre des possibilités similaires aux cycles quant aux recherches de motifs et permet le regroupement de structures similaires à travers leurs divers niveaux de généralisation de leurs graphes simplifiés. Harrison et al. (2003) travaillent sur une méthode ressemblant à l'algorithme de recherché d'isomorphismes utilisé par MC-Search mais en utilisant la distance entre les résidus.

ARTS (Alignment of RNA tertiary structures) est un algorithme qui examine des doublets de paires de bases entre deux structures et recherche des groupes de doublets structurellement superposables (Dror et al. 2005). Finalement, Sykes & Levitt (2005) utilisent une variante où les membres des doublets ne présentent pas nécéssairement une relation entre eux mais sont simplement à proximité. En construisant une librairie minimale de trente doublets à partir des structures observées, ils parviennent à reconstruire la plupart des molécules d'ARN.

Ces différentes méthodes proposent comme motifs des fragments d'ARN bien définis algorithmiquement, et rendent donc possible la recherche à grande échelle. La plupart essayent aussi de construire des structures tertiaires grâce aux fragments.

Une toute nouvelle approche mise au point au laboratoire par Marc Parisien (article soumis) dans le logiciel *MC-Fold* utilise une version modifiée du concept de cycle pour reconstruire des structures à partir de la séquence en calculant les probabilités d'adjacence de cycles d'après leur arêtes communes et la fréquence observée dans les structures résolues à ce jour.

2.10 Limitations de l'étude de structures

Il est important de mentionner certaines limitations inhérentes à l'étude de structures moléculaires :

Le premier groupe de limitations, plus ou moins hors de notre contrôle, vient des données utilisées. D'abord, la résolution des structures cristallographiques doit être prise en compte dans l'analyse. Par exemple, certaines structures sont déterminées à une résolution de 7 à 10 Ångströms, au-delà de la longueur d'un lien covalent entre deux atomes, ce qui engendre une incertitude significative sur les données. Cet obstacle devient de moins en moins important au fur et à mesure que les méthodes s'améliorent, la structure complète des ribosomes étant aujourd'hui disponible à des résolutions de 2.4 à 3.0 Å (Ban et al. 2000).

En regardant la liste des espèces pour lesquelles la structure du ribosome est disponible, on s'aperçoit que celles vivant dans des environnements extrêmes sont surreprésentés, leur génome étant plus stable et donc plus facile à cristalliser. Il pourrait être justifié de se demander quel genre de biais cela apporte sur les données, comme par exemple une prévalence de paires GC ou de motifs thermodynamiquement plus stables. Heureusement jusqu'à présent les conclusions qui ont étés tirées de l'étude de ces structures

semblent être généralisables, et plusieurs autres structures de taille moyenne sont disponibles (ribonucléases, ribozymes, ...)

Le deuxième groupe de limitations vient des méthodes d'analyse. L'annotation automatique implique la nécessité de discrétiser des paramètres qui sont en fait continus, et il est donc possible d'avoir deux groupes de nucléotides pratiquement identiques visuellement mais dont l'annotation diffère (effet de seuil). Cet état des choses doit être pris en compte dans l'analyse des données où des étapes supplémentaires seront ajoutées pour recouper les résultats qui auraient dû recevoir les mêmes annotations. Le plus souvent, c'est une inspection visuelle, ce qui malheureusement porte un coup dur aux projets d'annotation automatique.

Finalement, il faut aussi garder en tête que toutes les méthodes de résolution de structures tridimensionnelles ne donnent en final qu'une image statique d'un objet qui est somme toute extrêmement souple et mobile, et cette image est prise avec la molécule hors de son milieu cellulaire naturel et donc dans des conditions légèrement différentes. Pour contrer ce point, l'expérience et le bon jugement lors de la formulation des conclusions sont encore les meilleurs atouts.

2.11 Étude en/hors contexte

Une dernière limitation de l'analyse structurale vient du fait que beaucoup d'études sont réalisées sur des petites molécules synthétiques et isolées. Il apparait fort plausible que certains motifs ne se forment que lorsque dans une solution avec des conditions (pH, sels, etc.) précises et/ou lorsque mis en présence d'autres éléments avec lesquels ils peuvent interagir, et il est donc possible qu'une quantité de motifs d'interaction n'ont pas encore étés découverts. Seules les méthodes expérimentales *in vivo* peuvent éliminer ce problème.

3. Motifs d'ARN

L'ARN est en majorité composé d'hélices double-brin de type A formées d'appariements Watson-Crick canoniques. Ces régions sont maintenant considérées comme « génériques » et bien que certaines protéines se lient aux hélices canoniques, la recherche de motifs se concentre sur les régions d'appariements non-Watson-Crick (Leontis & Westhof 2003).

3.1 Boucles à quatre nucléotides

Déjà avec les alignements de séquence (Woese et al. 1990) et les alignements de structures secondaires (Cannone et al. 2002 ; Lescoute et al. 2005) il était possible de détecter les motifs les plus clairs et les plus abondants : les tétraboucles, c'est-à-dire les boucles de quatre nucléotides. Des études phylogéniques ont révélé que se sont les terminaisons les plus fréquentes des doubles hélices, avec les pentaboucles et les triboucles en suite (Woese et al. 1990). Et donc, avant même la venue des premières structures de ribosomes le type de boucle à la fin de plusieurs hélices était connu. Woese et al. (1990) et Antao et al. (1991) ont aussi noté que les boucles cGNRAg, cUNCGg et, en quantité moindre, gCUUGc étaient particulièrement communes (les lettres minuscules indiquent la paire de base fermant la boucle). La communauté scientifique s'est alors intéressée à déterminer leur structure (Heus & Pardi 1991 ; Jucker & Pardi 1995; Ennifar et al. 2000).

Des études ont montré que ces séquences forment en fait des boucles d'une stabilité supérieure à la plupart des autres séquences (Antao et al. 1991), et qu'ainsi elles participent à l'initiation du repliement de la structure (Tinoco & Bustamante 1999). Chacune des trois adopte une forme caractéristique et distincte, et les trois groupes remplissent des rôles différents dans une structure, tout en étant parfois interchangeables. En effet, lorsqu'aucune interaction particulière n'est requise à l'endroit occupé par la boucle il apparait parfois dans les alignements que l'une ou l'autre peut venir fermer l'hélice de façon équivalente (Woese et al. 1990).

3.2 Autres types de motifs

Plusieurs autres types de motifs d'ARN sont connus. Certains sont vus comme étant plus structuraux, comme le « U-turn » (Jucker & Pardi, 1995), un groupe de trois nucléotides qui présente un changement brusque de direction du squelette phosphaté. Cet élément se retrouve dans le motif GNRA, dans le motif T-loop, ou seul.

Il existe aussi des motifs qui sont fonctionnels bien qu'ils soient aussi décrits de façon structurelle. Le motif « sarcine/ricine » est un site du ribosome vulnérable à l'action des toxines α–sarcine et ricine (Szewczak & Moore 1995). Il existe sept sites de structure semblable à ce motif dans l'ARN ribosomal.

Finalement, on dénombre des motifs qui sont caractérisés par leurs interactions, comme le « A-minor » (Nissen et al. 2001), une adénine qui s'insère dans le sillon mineur d'une hélice voisine. Ce type d'interaction tertiaire est une des plus communes dans l'ARN.

3.3 Motif YNMG

La structure de la tetraboucle cUNCGg à été résolue par RMN (Allain & Varani 1995) puis par cristallographie (Ennifar et al. 2000) et les interactions stabilisant le motif ont alors été décrites (voir **figure 11**). La publication de la structure du ribosome (Nissen et al. 2000) à permis l'observation d'instances supplémentaires de la tetraboucle.

Proctor et al. (2002) ont effectué une étude exhaustive de tetraboucles de séquences diverses et ont ainsi observé des boucles avec un consensus de séquence YNMG qui adoptaient une structure similaire au motif UNCG. Le motif à par la suite été identifié dans une région critique du génome du virus *coxsackievirus* B3 (Du et al. 2003), puis dans divers autres homologues (Du et al. 2004 ; Ohlenschläger et al. 2004 ; Ihle et al. 2005 ; Melchers et al. 2006 ; Headey et al. 2007) en plus d'être la structure adoptée suivant une mutation pathogène de la télomérase humaine (Theimer et al. 2007)

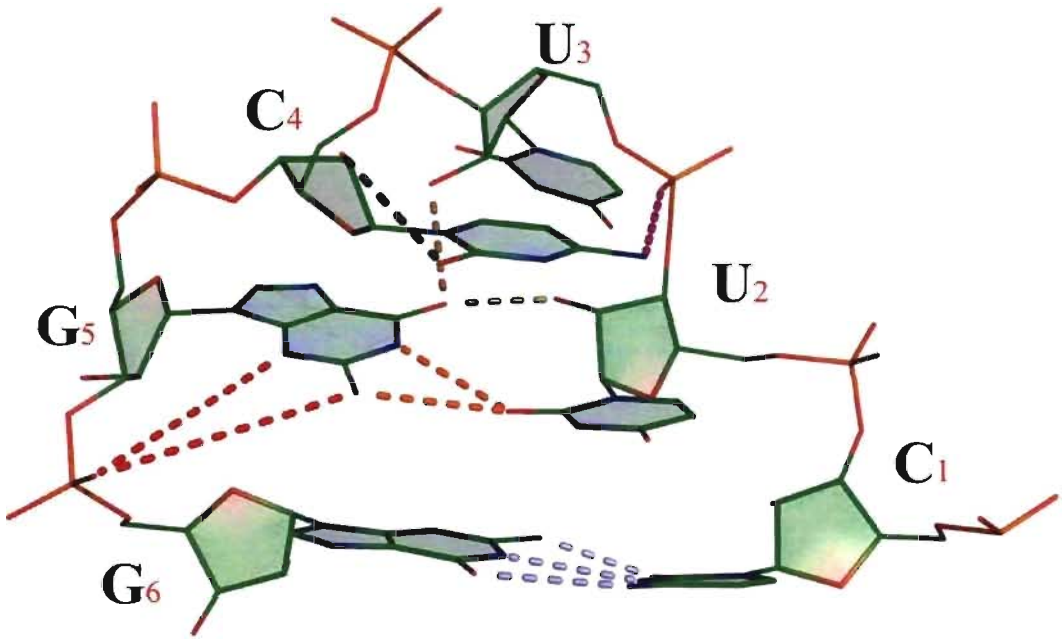


Figure 11 : Le motif YNMG, tel que décrit par Ennifar et al. (2000). Plusieurs liens viennent conférer à ce motif sa grande stabilité : Les nucléotides U_2 et G_5 sont liés par deux appariements, un W/W non-canonique (orange) et une relation entre le groupement O2' de U_2 et l'oxygène O6 de G_5 (vert). Ce dernier atome forme aussi un pont hydrogène (brun) avec le groupement O2' de U_3. Une relation interne entre la base et le ribose (noir) vient restreindre C_4. La boucle est fermée par une paire C-G canonique (bleu pâle). Une molécule d'eau vient compléter une interaction entre G_5 et le groupement phosphate de G_6 (rouge). Des empilements sont aussi présents entre C_1-U_2 et C_4-U_2 (non-illustrés).

Dans tout ces cas, le consensus de séquence YNMG n'est pas toujours respecté (Ihle et al. 2005 ; Melchers et al. 2006), ce qui laisse déjà envisager que d'autres possibilités de séquences ou de structures existent. Notre étude, présentée au chapitre suivant, confirme ce fait en proposant plusieurs nouvelles instances du motif YNMG avec des insertions et autres déviations en séquence et en structure.

4. Objectifs

Tout au long de ces travaux de maîtrise, les objectifs ont étés les suivants :

Objectifs généraux :

- Comprendre la structure de l'ARN et les règles régissant sa formation, pour améliorer la modélisation *de novo*.

- Étudier les motifs d'ARN, pour en dériver une relation entre la séquence et la structure de façon à franchir le fossé qui sépare les deux.

Objectifs spécifiques :

- Approfondir la notion de cycles d'interaction, et étudier la distribution des cycles dans les structures.

- Utiliser et enrichir la suite d'outils d'analyse structurelle du laboratoire (les outils MC-*).

- La partie principale à porté sur la caractérisation du motif d'ARN « YNMG », la recherche et la description de nouvelles instances, et la formulation d'hypothèses sur leurs rôle.

5. Travaux préliminaires

En premier lieu, étant issu du domaine de l'informatique j'ai du acquérir des notions générales de biologie structurale sur les protéines et les acides nucléiques. Cette partie à été principalement achevée à l'aide de cours du programme de bio-informatique. Les principes propres aux structures d'ARN ont étés vus par la lecture d'articles et la réalisation, au début de la période de recherche, de l'étude d'un motif d'ARN simple, les bourgeons (« bulge ») de un seul nucléotide. Le but de cet exercice était de s'initier aux méthodes et outils du laboratoire, de cataloguer les occurrences d'un motif simple et de tirer des conclusions quant à sa distribution dans les structures d'ARN en général.

Les quelques mois suivants ont étés mis à profit pour étudier et utiliser la librairie de manipulation de structure d'ARN déjà développée au laboratoire, *mccore*, ainsi qu'aux outils qui l'utilisent (tels que vus précédemment). J'ai entre autres fait des changements au logiciel *MC-RMSD* pour le rendre plus convivial, mis à niveau les logiciels *MC-Cycle* et *MC-Annotate* pour la dernière version de la librairie *mccore*, et créé un logiciel de classification des instances de motifs, que j'ai appelé simplement *MC-Kit*.

5.1 MC-Kit

Les méthodes disponibles au laboratoire, et principalement la recherche à grande échelle d'instances de sous-structures (graphes) dans l'ensemble de la banque de données PDB, génère une quantité énorme de résultats, souvent des milliers de fichiers. Les évaluer manuellement et/ou visuellement de façon individuelle est extrêmement fastidieux. C'est peu après avoir généré quelques premières centaines d'instances que j'ai réalisé ce point et débuté l'outil *MC-Kit*. De principe assez simple mais contenant maintenant beaucoup de code source, il permet de lire et d'annoter un groupe de fichiers et ensuite de faire différentes opérations avec les résultats d'annotation. Quelques exemples déjà disponibles sont :

- Visualiser toutes les annotations possibles.
- Trouver le consensus d'un ensemble (de séquence, de conformation ribose, de relations, etc.).
- Regrouper tout les doublons selon un ou plusieurs critères d'annotation donné (mêmes séquences, mêmes conformations, mêmes paires entres les résidus X et Y, etc.).
- Sortir toutes les instances distinctes selon un ou plusieurs critères d'annotation.

De plus, certaines opérations de regroupement de données sont possibles comme de regrouper ensemble les résultats d'une opération pour ensuite faire une autre manipulation sur ce groupe seulement, fusionner des groupes, etc.

Ce logiciel à été utilisé durant toutes mes études pour accélérer les travaux.

5.2 Étude sur les cycles

Dans le but de bien comprendre le concept de cycle, et de possiblement trouver une correspondance dans la position d'un cycle et l'identité des cycles avoisinant, j'ai procédé à une étude de ceux-ci pour un certain nombre de structures homologues disponibles dans la PDB. Les ARNt et ARNr sont les molécules avec le plus grand nombre d'instances et j'ai concentré mes recherches sur ces types de structures. J'ai d'abord modifié le programme d'énumération de cycles de façon à obtenir des résultats plus adaptés à mes besoins, c'est-à-dire une liste des cycles présentant pour chacun tous leurs voisins. J'ai ensuite complété la décomposition en cycles de toutes les structures similaires et effectuée une comparaison des voisins pour chaque instance. Des correspondances ont été observées, mais l'étude doit encore être approfondie. Ces travaux ont fait l'objet d'un rapport écrit et d'une présentation orale dans le cadre du cours BIN6001.

5.3 Tables de modélisation

En parallèle à la caractérisation d'un motif de tetraboucle, j'ai cherché toutes les boucles de 1 à 8 nucléotides dans l'ensemble des structures d'ARN de la banque de données PDB dans le but de déceler d'autres motifs de boucles avec insertion et d'étudier les correspondances entre différentes tailles de boucles. J'ai ensuite classifié les résultats selon une nomenclature mise au point par Lisi & Major (2007).

Ce qui nécessitait jusqu'à présent quelques semaines de travail pour chaque taille de boucle à été complété en à peine trois pour l'ensemble des résultats grâce à *MC-Kit.* J'ai donc produit ce que nous avons nommé des « tables de modélisation » pour les différents types de boucles d'ARN, qui énoncent la correspondance entre une séquence de boucle et ses relations d'empilement et d'appariements internes possibles tels qu'observés dans l'ensemble de la PDB. La page suivante présente la table résultante pour les diboucles.

Tableau IV (page suivante) : Table de modélisation pour les diboucles. La table est organisée par index de la paire fermant la boucle (axes vertical et horizontal, nucléotides 1 et 4 pour une diboucle). Les seize cases internes présentent les instances observées pour chaque cas : Les faces d'interaction de la paire de fermeture, la séquence de la diboucle (deux nucléotides) et une représentation symbolique de sa structure : D'abord les interactions entre les nucléotides du squelette, suivi des interactions internes. Ex : SLS-S2-4 indique un empilement (stack) entre les nucleotides 1-2 et 3-4, ainsi qu'une relation d'appariement (pairing) entre les nucléotides 2-4. L : Adjacence simple (link). Aucune instance de diboucle avec des séquences cNNu n'a été observée.

	A			C			G			U		
A	H/C8	GU	LLL	S/H	AA	LLS	S/S	AU GG	LLL-S2-4	S/W	GC UC	LLS
							S/W	AU	SLL			
							W/W	AG	LLL-S1-3			
							W/S	AU	LLL-S2-4			
							W/S	AU	SLL-S2-4			
C	W/H	AC	LSS	Bs/H	GA	LLL	Bs/W	UA	LLL-S1-3			
	W/W	AU	LSL	W/H	GA	LLL	S/S	AU	LLL-S2-4			
	W/W	UC	SLS-P1-3	W/H	AA	LLS	W/H	GG	LSS			
				W/H	AA	LSS	W/W	CG	SLL			
							W/W	AA	LLL-S1-3			
G	S/H	CA	LLL	S/S	AU	LLL	Bs/H	CC UA	LLS	S/W	GA	LLL
	S/H	AA UA UG	LLS	S/W	AA	LSS	S/H	UA	LLS	W/W	CA	LLL
	S/H	AA AG AU CA CG GA GG UA UG	LSS	S/W	CA	LLL	S/H	AA	LSS			
	S/H	AG	SLS	W/W	GU	LLL	W/H	CC UU	LLL			
	S/H	AA	LLL-S1-3	W/W	GU	LLS	W/H	CC UA UU	LLS			
	S/H	UA	LLL-S2-4				W/H	GC UU	LSL			
	S/W	AA CG	SLL				W/H	CC	LSS			
	S/W	AA CA	LSS									
	S/W	AA	LLL-S1-3									
	S/W	CA	LSS-P1-3									
	W/H	CA	LLL									
	W/H	CA	LLL-S1-3									
U	H/H	CG	LSL	Bs/H	AA CA	LSS	Bs/W	AC UC	LLL	S/W	GA	LLL
	S/C8	AG	LSS	S/H	AA CA	LSS	Bs/W	UC	LLL-S1-3			
	W/C8	GC	LSL				S/S	AU	LLL			
	W/H	GA	LSL				W/W	AC UC	LLL			
	W/W	UG	LLL				W/W	UC	LLL-S1-3			
							W/W	UC	LLL-S1-3 P2-4			

6. Composition du mémoire

Ce mémoire est présenté selon un format par articles. Le premier article (chapitre 2) expose mes travaux de caractérisation d'un motif d'ARN. Le troisième chapitre contient un second article sous forme de note d'application sur un logiciel que j'ai réalisé lors de travaux avec une étudiante au doctorat, Emmanuelle Permal, sur des virus à génome d'ARN. Le motif qui m'intéresse se retrouve dans toute une famille de virus à génome d'ARN.

Le dernier chapitre contient une brève conclusion, et en annexe se trouve un troisième manuscrit d'article découlant des travaux de E. Permal et pour lequel je suis deuxième auteur, qui discute des motifs d'ARN dans les virus.

Chaque chapitre débute avec un paragraphe d'introduction plus complet.

Chapitre 2: Article – Analysis of the YNMG RNA fold

Ce chapitre présente, sous forme d'article, une analyse approfondie d'un motif précis d'ARN, dit « YNMG ». C'est un motif intéressant parce qu'il démontre bien plusieurs des caractéristiques des structures d'ARN et des motifs : petit et simple, répété à travers différentes structures. Sa conformation exacte varie selon les structures environnantes et démontre une capacité de changement pour s'adapter aux contraintes locales. Le motif YNMG, qui est retrouvé dans de nombreuses structures, est présentement l'objet d'études et de publications, principalement de par sa présence dans un site critique de virus ayant un impact économique significatif.

L'article sera soumis à la revue *RNA* (www.rnajournal.org) sous le format *Application Note*, qui est limité à deux pages.

Analysis of the YNMG RNA fold

Louis-Philippe Lavoie and François Major

Institute for Research in Immunology and Cancer
Department of Computer Science and Operations Research
Université de Montréal
PO Box 6128, Downtown station
Montréal, Québec H3C 3J7
CANADA

Short title: Analysis of the YNMG RNA fold

Keywords: RNA; YNMG fold; tetraloop;

Corresponding author: François Major [redacted])

Four-nucleotides loops (tetraloops) occur frequently in biologically active RNAs. They are unusually stable and often fulfill important structural or biological functions. One of the three major classes of tetraloops, the UNCG motif, has recently been found with an extended sequence consensus, YNMG, and as a critical element of entero- and rhinoviruses. Structural characterization reveals that, while the global fold is conserved, variations are observed in the specific interactions. In this study we report on an exhaustive search and classification of putative YNMG motifs with significant sequence and structure deviation from the accepted consensus. This demonstrates the inherent flexibility of RNA, and its ease of adaptation to local structural pressure.

INTRODUCTION

It is now well established that RNA molecules organizes structurally into smaller recurring substructures (i.e. *motifs)* that aggregate together with Watson-Crick, helical base-pair regions to form the global fold (Moore 1999; Hendrix *et al.* 2005; Leontis *et al.* 2006). Four-nucleotide RNA loops (tetraloops) are among the most common and most studied structural motifs of biologically active RNAs (Woese *et al.* 1990). The fold of these loops has been observed to fall predominantly in three major classes, all displaying high levels of stability: GNRA, UNCG and CUUG (R is a purine, N can be any of the four nucleotides).

The UUCG tetraloop structure was first solved using NMR (Allain & Varani, 1995) followed by x-ray crystallography (Ennifar *et al.* 2000) of a single hairpin of 57 nucleotides. Other occurrences of this motif have also been observed in the structure of the whole ribosome (Nissen *et al.* 2000). The fold is stabilized by multiple internal hydrogen bonds and base stacking interactions (see **figure 1**). Most notably, a double pairing between the first (L_1) and last (L_4) nucleotides of the loop (a base-base pairing and a ribose-base, L_1O2'-L_4O6 pairing) is present, and the L_4 nucleotide is normally in *syn* configuration. Based on extensive analysis of tetraloop thermodynamic stability, Proctor *et al.* (2002) proposed to extend the UNCG family to a new consensus YNMG (Y is a pyrimidine, M is either C or A), and the internal fold interactions were confirmed in the other YNMG loop sequences by their experiments. In each of the instances analysed in this present study, the interactions are generally conserved in one form or another despite variations in sequence or structure.

NMR structures of wild type hairpins containing the YNMG fold have been described previously (Du *et al.* 2003 & 2004; Ohlenschlager *et al.* 2004; Headey *et al.* 2007). More recently, Ihle *et al.* (2005) have reported that the cGUUAg tetraloop present in the stem-loop D of bovine enterovirus 1 RNA folds as a YNMG motif despite the deviation from the normal consensus sequence (lowercase letters in sequences indicate the closing base pair). In this family of virus, the motif's structure, more than its sequence, is recognized to be a key protein binding element in the interaction with its

viral protease (Ohlenschläger *et al.* 2004; Ihle *et al.* 2005). Melchers *et al.* (2006) added support to the GYYA extension after finding (by *in vivo* SELEX) a viable uGCUAg tetraloop with YNMG fold in the stem-loop D of poliovirus oriL. However, a large number of the existing studies use synthetic hairpins (SELEX), do not show the tetraloop in interaction with its probable target, or both.

In this report, we provide a detailed analysis of putative YNMG structures and sequences found in published RNA structures (Berman *et al.* 2000). Multiple occurrences of this motif were retrieved from a RNA subset of the Protein Databank (PDB) and annotated to determine the characteristics common and specific to each fold. In a subsequent analysis these characteristics were used to extract new putative YNMG fold instances. By using RNA structural analysis software tools, we are able to describe the different instances and their interactions. We show that the YNMG fold, and likely RNA as a whole, is very versatile and able to allow local variations in structure. Single nucleotide insertions are easily integrated in the YNMG motif to further diversify its interaction patterns. This research also demonstrates how automated computer analysis can supplement traditional structural analysis methods.

RESULTS

Search for UNCG/YNMG folds. Nucleotide conformations and interactions matching those of the UUCG tetraloop crystal (Ennifar *et al.* 2000) were searched for in a set of high-resolution crystallographic structures (see Matherials & Methods section for search parameters). We found thirty-three instances from only three distinct structural locations: one in the 23S RNA of *Haloarcula Marismortui* (Klein *et al.* 2004) (30 redundant structures, same location), one in a ribosomal RNA fragment crystal (PDB accession code 1DK1) and one last in the synthetic UUCG loop (PDB accession code 1F7Y, dimer) often used as the reference for this fold (Ennifar *et al.* 2000). The tetraloop in PDB 1DK1 was added synthetically to fix the helical fragment under study. All results have a

cUUCGg sequence; a search with no sequence restriction (nNNNNn) resulted in no additional instances.

Structures from that group diverge by up to 0.85 Å (RMSD criterion), with a maximum distance from the reference fold of 0.58 Å. Even with this high degree of structural similitude, the three location cluster in distinct structural groups, hinting that the exact fold might be context specific. Indeed, while the UNCG/YNMG tetraloop is widely reported as a very common motif in RNA, in effect the first level of search resulted in only two instances from wild type rRNA locations. This provides strong evidence that the exact network of interactions stabilizing the fold varies with each location.

Several additional YNMG instances, mainly from viral RNA (see **table I**), were obtained from a review of recent literature and added to our list to expand our search criteria. The last two cases have cGUUAg & uGCUAg sequences, extensions to the fold family proposed by Ihle *et al.* (2005) and Melchers *et al.* (2006), respectively. All except PDB 2EVY exhibit the G_{L4} *syn* nucleotide, but we did not otherwise observe clear structural consensus with regard to pairings, stackings or sugar puckerings. The specimens from this group diverge by up to 3.4 Å from the reference fold (4.12 Å cluster dispersion). PDBs 1RAW (ATP Aptamer) and 2EVY (poliovirus oriL SLD tetraloop) contain the instances with greatest deviation: Their sets of NMR models by themselves diverge by 3.56 Å (over 10 models) and 1.31 Å (over 18 models), respectively. Without these two sets the RMSD of the NMR models from literature is brought down to 1.47 Å from the reference fold (1.74 Å cluster dispersion).

Using all current YNMG instances, we next performed a more exhaustive search with less stringent search criteria (see Material & Methods), resulting in several new putative YNMG-fold instances. Firstly, all loops with a sequence matching the nYNMGn consensus were retrived, regardless of the fold other than a requirement for a pairing interaction between the first and sixth nucleotides (the closing pair). Following this, the sequence was generalized to remove the identity constraint on the nucleotides and find additional YNMG-fold sequences outside of the consensus. Finally, as much as possible

the adjacency constraints between nucleotides of the loop were removed with only pairing and stacking interactions specified, with and without sequence identity. Because the majority of the proposed YNMG motifs so far are only available as NMR structures, these searches were conducted on the complete set of NMR and high-quality crystallographic RNA structures available in the Protein Databank. This allowed the discovery of several putative "YNMG-like" motifs presenting alterations to the traditional sequence, the structure, or both. From simple single-nucleotide sequence insertions to one specimen with two strands and a 3-nt insertion, all results were accumulated and visually evaluated for similitude to the canonical YNMG fold or one of the proposed specimens from literature. A list of selected specimens is presented in **tables II** and **III** : the former contains loops that form the YNMG motif, and which present an insertion or are involved in interaction (or both). The next table lists 4-nt loops with nYNMGn sequence but that do not fold as a YNMG motif. The unlisted remaining entries were all yYNMGg tetraloops with a classic YNMG fold, and will not be discussed further (see **table S-I** in supplementary materials for a complete list of YNMG instances). On a general note, all loops with a non Watson-Crick closing pair have a non-YNMG fold. Some instances (2J00 1090-1095 & 2J01 542-551) also are non-YNMG despite a sequence matching the consensus and a Watson-Crick C-G or U-G closing pair. All cUNCGg tetraloops, the most thermodynamically stable version of the sequence (Proctor *et al.* 2002), adopt a likely fold candidate. The instances will be discussed in greater details in the following pages.

Analysis of YNMG sites in 23S ribosomal RNA. When searching the 23S ribosomal RNA (LSU), only one four nucleotide loop with YNMG fold is found (out of 21 tetraloops flanked by a Watson-Crick (W/W) pairing) (see **table II**), and it is in fact the location matching the restrictive UUCG tetraloop descriptor and is the instance most similar to the tetraloop of PDB 1F7Y (0.5 Å RMSD). Another YNMG sequence candidate but with non-W/W closing pair, aCAAGa (2838-2842), does not adopt a good

YNMG fold (see **table S-II** for a list of all tetraloops in ribosomal RNA, and figure S-1 for the motifs in 23S rRNA, both in supplementary materials).

In loop 1769-1774 (cUUCGg), the YNMG motif is involved in RNA-RNA tertiary interactions with two nearby adenines through the second loop nucleotide (U1771), in a W/W pairing (A2018) and a stacking (A1885). The rest of the nucleotides of the fold are inward in a compact group, positioning the phosphate backbone for interaction with ribosomal protein S37 which has several arginine residues in hydrogen bonding distance (see **figure 2**). Study of phylogenetic data (Cannone *et al.* 2000), reveals that this loop generally varies within the YNMG sequence consensus but punctual deviations are noted (AUUG, UUUG, UUAG). The two interacting adenines are completely conserved throughout the three domains of life, except in the one instance of UACG sequence, where the one involved in pairing (A1885) becomes a uracil.

Removing the backbone and sequence constraints (see Material & Methods) yields four more candidates with a 1-nucleotide insertion. Although all have folds similar to the reference UUCG, their sequence does not follow the yYNMGg consensus. All candidates are involved in tertiary interaction with rRNA, proteins or both.

Loop 136-142 (cUUCGcg, insertion is underlined) is near a GNRA tetraloop (252-257) in *H. Marismortui* with the second uracil in the loop bulged out of the YNMG fold toward the GNRA (see **figure 3a**). The distances visible in the crystal are not within hydrogen bonding range but two water molecules are present between the two loops. In this instance, the inserted nucleotide is stacked on a cytosine lower in the stem (C130), suggesting that it could serve to hold the tetraloop in the proper orientation for the GNRA loop. Most of the intra-loop interactions found in 1F7Y are conserved or find their equivalent despite the insertion (see **figure 3b**). The fourth nucleotide of the tetraloop (a guanine) is in the typical *syn* configuration and the first and fourth loop nucleotides are stacked. Unfortunately, the conservation of this YNMG loop is very poor and the crystals and alignments show many gaps in this region, with inconsistent sequence numbering. In *Escherichia Coli* (Berk *et al.* 2006), the structurally equivalent position presents a triloop and the corresponding GNRA does not exist. However another tetraloop is found a few

nucleotides before (137-142) with uUUCGa sequence and a near perfect YNMG fold. This loop is interacting with a protein in its minor groove side, with the second loop nucleotide bulging out over the protein (see **figure 3c**). The two instances are similar (1.70 Å RMSD) but at slightly different locations in the ribosomal RNA. In *Thermus Thermophilus* (Selmer *et al.* 2006), a suitable equivalent to the first loop of *H. Marismortui* is a gCUUGc loop at position 153, over 10 Å from its corresponding RNA partner, this time a gCUUGuc loop (271i-271o), a CUUG-like fold with a nucleotide insertion. The additional tetraloop is gGGAAc (139-142a) and here is also involved in a protein interaction, as well as RNA tertiary interaction but the loop is neither a YNMG fold nor a GNRA.

The YNMG fold at position 195-201 of *H. Marismortui* (cGCAAug) is involved in RNA-RNA interactions on both sides of the loop, as well as a presentation of the phosphate backbone to nearby ribosomal protein L15, similar to the loop 1769. Several ions and water molecules are visible in the crystal in this area. Here the insertion is looped out to the surface of the ribosome, with no apparent interactions, and the second loop nucleotide is one again protruding from the YNMG fold, on the minor groove side toward nearby RNA strands and protein L15. Interestingly, the phylogenetic data here reveals that a YNMG motif at this position is likely a rare occurrence, with a majority of uAACNa sequences, but an overall mixed consensus. In *T. Thermophilus* the equivalent loop at this position (225-229) has an aGAAAu sequence but is in fact a triloop with the last A bulged out of the fold, RNA-RNA interactions on both sides of the fold and the backbone oriented toward protein L15, in the same way as in *H. Marismortui*. In *E. Coli* the loop is once again a tetraloop, but with cAACCg sequence and all four nucleotides in a compact fold toward the major groove. In all three crystals the local structural configuration is well conserved.

Loop 670-676 of *H. Marismortui* (gAGUaUc) is in interaction with ribosomal protein L4E on its minor groove side, again with the second loop nucleotide (G672) bulging out slightly at the protein (see **figure 4a**). Several water molecules are visible in the crystal inside and around the loop. The insertion (A674), this time inside the tetraloop

itself, is paired with a cytosine (C36). In *E. Coli*, this hairpin (cGAAuAg, 611-617) is also capped by a loop with the same fold and insertion bulging out, but here the helix has a different twist so that protein is on the 3' side of the loop and around the inserted nucleotide, with the second loop nucleotide packed closer to the fold. Despite this change, the two folds are extremely similar ($RMSD^{(-2)}$ = 0.85 Å). In *T. Thermophilus* the loop is a pentaloop (cGUUGAg, starting at 612), and although the local organisation of the structure is similar to *H. Marismoirtui* the loop has an unrelated fold. The first two instances show a significant RMS deviation from the reference UNCG fold (>2 Å), however it is visually similar to the only other YNMG-fold candidate with a G-C closing pair (loop 1134-1140 of 16S rRNA) (see **figure 4b**) and have similar interactions as the reference loop (see **figure 4c**).

Loop 872-878 (uGAAAgg, *H. Marismortui*) is, despite its sequence and insertion, an almost perfect YNMG fold (see **figure 5a** and **5b**). Protein L2 lies in its major groove and above it, with the insertion acting like a thumb under the protein, also stacking with a nucleotide in the same stem. The second loop nucleotide is bulging away on the minor groove side and involved in a stacking with an adenine and a pairing with a guanine. This configuration is maintained in *T. Thermophilus* (same sequence, 779-785) and *E. Coli* (uGAAAag, 779-785). In fact, the phylogenetic data reveals that this loop and its region are extremely well conserved, with the insertion varying between A, G and U but significant covariation with its interaction partners.

Analysis of YNMG sites in the 16S rRNA. A search of all tetraloops in 16S ribosomal RNA reveals 19 tetraloops in *T. Thermophilus* of which three are sites with YNMG fold (see **table II** or **table S-II** and **figure S-2** in supplementary materials), with sequences cUACGg (positions 342-347 and 1449-1454) and cUUCGg (419-424). The three loops cluster together to within 0.65 A. Another loop with a sequence uUAAGu (1090-1095) does not adopt a YNMG fold. As described by Proctor (2002), the first of these YNMG occurrences (loop 343) is a typical YNMG fold interacting with a GNRA at position 159. The YNMG motif is very well conserved but the GNRA is not, only the two adenines

seem to be relevant. Indeed, they are the part of the loop involved in the interaction (see **figure 6**).

Loop 419-424 is also very similar to the reference UUCG fold (0.57 Å), and is conserved in most species although a small percentage present a GNRA sequence consensus instead. This loop is located at the tip of the *latch* stemloop in the SSU, a region that opens and closes to allow the mRNA strand to reach the decoding center (ref crystalo SSU), which would lead to believe it is a critical area and structurally conserved, though not necessarily as a YNMG fold. Its precise tertiary interactions are not clear in the 16S rRNA crystals available.

Loop 1449-1454 also superposes very well onto the loop from 1F7Y (0.5 Å) but this time the phylogenetic data (Cannone *et al.* 2002) shows a mix of GNRA and CUUG sequences as well, suggesting that there is no requirement for a specific nucleobase fold at this stem. Indeed, it is involved in interaction with a protein but through the stem rather than the loop motif. In *E. Coli* the entire hairpin is longer and the loop is out of reach from the protein. In both cases the second loop nucleotide appears to be alone as the distance to the next closest strand is greater that normal hydrogen bonding ranges.

A fourth loop at position 1134-1140 with a 1-nucleotide guanine insertion between the loop's last nucleotide and the closing base pair (gUUCGgc in *T. Thermophilus*) was found to also adopt a YNMG-like fold (1.7 Å from the reference UUCG fold (PDB 1F7Y)). In *E. Coli* the loop becomes cUCCGgc. The bulging G is involved in a stacking with a lower cysteine (C1128) and a pairing (G1142) in the same helix, a configuration that is also occurring in other instances. Here it has the net effect of forcing an additional bend into the stem, fixing the position of the YNMG motif so that it is presented to the surface of the ribosome, possibly marking this loop as a binding site but phylogenetic data shows variation in the sequence, sometimes outside the YNMG consensus.

When the search parameters are widened further, an additional two-stranded YNMG-like fold is found (C883,U884,U561,C562,A563,G567). This represents the more extreme case of sequence insertions and variations (see **figure 7a**) with deviations of 2.8

Å from the reference UUCG fold. However, $RMSD^{(-2)}$ is 1.65 Å. The fourth loop nucleotide, here an adenosine, is in *syn* and similar pairings and stackings are observed (see **figure 7b**). Study of alignments reveals that the nucleotides composing this motif present a high degree of conservation as well as covariation of the interactions partners, indicating that it is a stable point of the SSU.

Analysis of YNMG sites in non-ribosomal RNA. Of the motif instances from non-ribosomal RNA, only two show the loop in contact with nearby RNA or a protein.

Loop 13-18 (cUUCGg) of PDB 1EKZ shows part of *Drosohphila* Staufen protein in contact with the minor groove of the YNMG motif with L_2 acting as a thumb on one side of the protein. The 36 models in the NMR set diverge by as much as 2.95 Å, but most of the dispersion comes from L_2 moving to stay close to the protein (see **figure 8a**).

Loop 26-31 (cUUCGg) of 1P6V shows SmpB protein again on the minor groove side of a tmRNA fragment stemloop. This instance is very similar to the reference fold (1.43 Å).

As noted earlier, the motif found in the NMR models of the ATP-binding aptamer (PDB accession code 1RAW) represents the set with the greater amount of RMS deviation for a single location, and serves to show the flexibility of the YNMG motif's backbone (see **figure 8b**)

The instances in the structures of stemloop D of coxsackievirus (PDB 1RFR and 1ROQ) and the consensus of enterovirus (1TXS) are all structurally very close to the reference fold (> 1.47 Å dispersion) despite their uYACGg sequences.

The stemloop D of BEV1 (PDB 1Z30) presents a strong sequence deviation (cGUUAg) but is also a genuine YNMG motif (1.4 Å from reference).

PDB 2EVY contains a YNMG motif found in the poliovirus oriL homolog of stemloop D, and presents another strong deviation (uGCUAg). This sequence is not a wild type one but was found to be viable by *in vitro* SELEX experiments (Melchers *et al.* 2006). The study also noted that YNMG motifs with a U-G closing base pair at not as stable as those with a C-G closing pair.

The loop found in the structure of human telomerase RNA (PDB code 1Q75) provides another interesting example of a YNMG motif with an insertion (Theimer *et al.* 2003). The normal sequence is cUCGCug, a pentaloop. A two-base mutation in the loop was found to transform the sequence into cUC***AG***<u>ug</u> (mutation in bold, insertion underlined) and cause the loop to shift to a YNMG motif with U109 bulged out, in a way similar to other examples with an insertion. Further analysis also revealed the pentaloop to be almost as stable as its 4-nt counterpart. From this, the authors suggested that a closer look at pentaloops in available structures might reveal more examples of embedded tetraloop motifs.

DISCUSSION

Nucleotide identity distribution in the sequence. As reported by Proctor & al. (2002), although there seems to be no thermodynamic constraints on the identity of the nucleotide in the second position of the loop (L_2), it was less likely to be a guanine in their observed sequences. Du & al. (2003) pointed out that the second nucleotide is the most flexible of the loop, and does not present intra-loop interactions specific to any of the four types and so has the least restrictions on its identity. However, out of all the tetraloops instances with nYNMGn consensus in high-quality structures (regardless of fold), only one location has a G present at position L_2: 2J01 713-718 (LSU of *T. Thermophilus*). Review of alignments of 23S rRNA (Cannone *et al.* 2002) reveals that this is the only occurrence of this sequence (gUGAGa) for this loop, with GNRA or UMAC being the most prevalent consensus. Indeed, when viewed the structure of the tetraloop in this structure presents an S-shaped backbone and a purine stack on the 3' side in a way similar to the classical GNRA, and although the overall fold is different the two shapes could fulfill similar structural roles. In effect, this means that there is no guanine at position L_2 in any YNMG tetraloops, with exception of loop 670-676, which presents a unique L_3L_4 insertion – gAGUaUc. Possibly the YNMG fold definition should be revised to exclude a guanine in second position, e.g. YHMG, but a sequence consensus is less and less useful to properly describe the fold.

In all tetraloops regardless of fold and sequence a guanine in position L_2 has also the lowest distribution frequency of any position for all nucleotides (data not shown). Proctor *et al.* (2002) attributed the low frequency of G_{L2} in his tetraloop experiments to the particular stability of GNRA loops in PCR, the distribution seems to hold even in resolved wild-type structures. This could be due to fact that those sequences become too sensitive to mutations in any other position, which would result in a significant change of fold (YNMG to GNRA). Ihle *et al.* (2005) reported that the cGUUAg sequence adopts a YNMG fold, but cGUAAg is a GNRA. gUAAGa (PDB 1U9S A182-A187), another tetraloop with a similar nYNMGn sequence (but notice the G-A closing pair), also adopts a GNRA-like fold.

Interestingly, after comparing the alignments of stem loop D sequences for 60 viruses, Du *et al.* (2003) reported that in this particular location the L_2 position is a guanine or an adenine in 23% and 52% of the sequences respectively, despite UUCG being the more stable sequence. Possibly this class of virus could be exploiting a niche in structural recognition of YNMG tetraloops, or the YRMG sequences have a flexibility better suited to binding with the viral protease.

Flexibility of different subsequences and closing base pairs. Recent studies (Proctor *et al.* 2002; Ohlenschlager *et al.* 2004) have shown that optimal thermostability of the YNMG motif is found with a C-G closing base pair, but as seen in the data, candidates are found with several types of closing base pairs. When no insertion is present, only the loops with a canonical (W/W) C-G, U-G or U-A closing pair adopt a fold matching or similar to the accepted YNMG fold. With insertions this is extended to include G-C (W/W). Although there is a clear requirement for strong (canonical) closing base pair, the reduced stability of non-CG pairs could translate into greater flexibility in the loop. Indeed, several of the specimens with insertions present non-CG closing pairs.

Based on phylogenetic evidence, Proctor *et al.* (2002) postulated that yYNAGg tetraloops are more flexible in shape and recognition ability than their yYNCGg counterpart. The list of tetraloops from high-resolution crystal structures contains only

one cYNAGg single-strand loop with a C-G Watson-Crick closing base pair (PDB 2J01 A542-551, sequence cCAAGg,), but its fold is very distorted, possibly by the close proximity of two proteins (L21, L18). There is also only one yYNAGg instance (2J01 A826-A831, sequence uUUAGg), also with a distorted (but different) fold and this particular tetraloop appears to be nested inside the ribosome and involved in several tertiary and protein interactions. If the restriction on the closing base pair is further relaxed, we find more of the other nYNAGn specimens from **table III**, none of which are good candidates for the characteristic YNMG fold. In brief, the only nYNAGn tetraloop found in this study to fold into the YNMG motif is from 1Q75, a pathogenic mutation of human telomerase RNA (first item in **table I**). In contrast, all of the surveyed nUNCGn single-strands tetraloops with a C-G (W/W) adopt a fold compatible with the YNMG motif. While the diversity of the specimens observed here might not allow for an absolute statement, it does appear that tetraloops with the yYNCGg sequence have more constraints to adopt the YNMG fold. This suggests that the type of pairing, as much as the loop nucleotides themselves, plays an important role in the correct folding of loop motifs. nCAAGn sequences seem to be a special case in that they never adopt a YNMG fold.

YNMG fold as a flexible tertiary interaction motif. When looking at the stability of the YNMG motif, it is worthwhile to compare it to a related tetraloop motif, the GNRA. The occurrences of YNMG folds that present sequence insertions, similarly to those in the GNRA backbone (Legault *et al.* 1998; Huppler *et al.* 2002; Lemieux & Major, 2006), supports the notion that the evolutive pressure of RNA motifs in general is not on the sequence but on the structure. The final sequence would therefore be dependant of the structural requirements in each specific context. It is worthwhile to note that in both cases, the folds seem more permissive to insertions in their 3' stacks.

In a majority of the observed instances of the YNMG fold, the motif is involved in at least one, sometimes several, tertiary interaction with nearby RNA, interactions with proteins, or both. Indeed, many of the YNMG motifs reported to date in the literature

were found at the critical stemloop D binding site of a family of virus (Du *et al.* 2003 & 2004; Ohlenschläger *et al.* 2004; Ihle *et al.* 2005; Melchers *et al.* 2006; Headey *et al.* 2007). This present study brings further evidence to support the notion that YNMG-fold interactions are predominantly structure based (Ihle *et al.* 2005).

Conclusion

Starting from a large-scale search of YNMG motifs in currently available RNA structures, we have extracted all instances. From there, we have analysed their fold and interactions to show that this motif is very flexible in sequence and structure, beyond the current yYNMGg fold consensus although the overall shape and characteristics are maintained. The instances found include not only the straight, CG-closed 4-nucleotide tetraloop but also representatives of the fold with sequence insertions and other types of W/W closings (CG, GC, AU, UG) and one multi-stranded instance where the fold and its typical interactions are preserved. We show that this flexibility allows the fold to adopt diverse structural and functional roles in the molecule such as RNA-RNA hairpin stabilization and different forms of protein binding (major groove, minor groove, backbone).

The canonical fold as originally described in literature has few representatives in the current set of high-resolution crystallographic biological structures. Instead, deviations are observed in the set of specimens of the YNMG fold, and it is likely that new instances will be discovered in the future with still further variations of the definition. The more stable characteristics of the fold are a L_4 *syn* nucleotide, pairings between L_1-L_4 and two-nucleotide stacks on both 5' and 3' sides with L_2 and L_3 on opposing of the loop. Instances have been found where one or more of these characteristics are absent while the overall fold is still preserved, and all other parameters are found to vary with little consensus. The GNRA motif has also been shown to be highly accommodating in its sequence and structure (Lemieux & Major, 2006), and the CUUG fold also exists with sequence insertions (unpublished observation). This suggests that RNA as a whole is extremely malleable and under selective pressure can easily

modify its local structure or sequence to preserve the global structure; this inherent flexibility should be integrated in future searches of RNA motifs. Devices such as isostericity matrices (Lescoute *et al.* 2005) provide further hints of this dynamic complexity, but more work is required. To fully understand the rules governing RNA whole-structure formation and to ultimately achieve structure prediction from sequence, we must first understand the local constraints, or lack thereof, permitting such variations while still preserving global structure and function. RNA structural analysis currently yields descriptions of fixed constructs, and the present nomenclature is insufficient to capture the full flexibility of RNA. As new motifs are discovered, and their variations analysed more thoroughly, it will become increasingly awkward to describe them in a way which can group the related conformations into one concise and meaningful concept.

MATERIAL AND METHODS

Dataset composition. All datafiles come from the Protein Databank (Berman *et al.* 2000) as of June 2007. Crystals containing RNA with resolution of 3 Å or lower are considered of high-quality for the purposes of this study. Although there was no filtering of redundant crystals in the initial search, this was taken into account when compiling the data. For the searches which included NMR data, the quality threshold was raised to 3.6 Å for crystallographic data.

Structure search and annotation. By manipulating the structure as a graph of nucleotides connected by basic interactions (backbone adjacency, pairing, stacking) (Gendron *et al.* 2001), it is possible to selectively consider only specific aspects of a motif and efficiently search for structural motif instances with or without the sequence constraint. *MC-Search*, a motif search program, takes as input a graph descriptor declaring required constraints on nucleotides identity and relationships, and searches for graph isomorphisms in a submitted set of 3D RNA structures. The results are PDB files containing the motifs, which are then annotated again as graphs with the same method. The nomenclature used is that of Leontis & Westhof (2001) for the base pairings, and Major & Thibault (2006) for the base stackings.

Closing pair definition. To recover as many specimens as possible, in the searches throughout this article the closing base pair of the loops was not restricted to a canonical Watson-Crick pairing but instead opened to pairing between any face (Hoogstein, sugar, Watson; Leontis & Westhof 2001) of the nucleotides.

RMS deviation computation. Unless otherwise specified in the text, all RMSD data is computed on all atoms after optimal superposition of the instances. Also, specifically for the YNMG motif it should be noted that the second loop nucleotide (L_2) has few movement restrictions and is normally involved in tertiary interaction, making its exact position dependant on the local context of the instance. It is therefore often of lower relevance in the RMS deviation sum when comparing two instances for structural similarity, and in these cases it is worthwhile to exclude it from the RMSD computation. This has been done where appropriate in this study and is noted as "$RMSD^{(-2)}$".

Reference motif. At several point in the text, instances are evaluated by their RMS deviation from an "ideal" UUCG tetraloop. Because it was the first to be crystallized and thoroughly caracterized, the motif found in the crystal structure with PDB ID 1F7Y (first model) is used as reference.

ACKNOWLEDGMENTS

We thank Emmanuelle Permal for assistance with manuscript preparation. FM is a CIHR investigator and a member of the Centre Robert-Cedergren of the Université de Montréal. LPL holds a CIHR scholarship in bioinformatics (Université de Montréal, programme biT).

REFERENCES

Allain F, Varani G. 1995. Structure of the P1 helix from group I self-splicing introns. *J Mol Biol 250*:333-353.

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

Berk V, Zhang W, Pai R, Cate J. 2006. Structural basis for mRNA and tRNA positioning on the ribosome. *Proc Natl Acad Sci U S A 103*:15830-15834.

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

Cannone J, Subramanian S, Schnare M, Collett J, D'Souza L, Du Y, Feng B, Lin N, Madabusi L, M□ller K, Pande N, Shang Z, Yu N, Gutell R. 2002. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics 3:2.

Du Z, Yu J, Andino R, James T. 2003. Extending the family of UNCG-like tetraloop motifs: NMR structure of a CACG tetraloop from coxsackievirus B3. *Biochemistry 42*:4373-4383.

Du Z, Yu J, Ulyanov N, Andino R, James T. 2004. Solution structure of a consensus stem-loop D RNA domain that plays important roles in regulating translation and replication in enteroviruses and rhinoviruses. *Biochemistry 43*:11959-11972.

Ennifar E, Nikulin A, Tishchenko S, Serganov A, Nevskaya N, Garber M, Ehresmann B, Ehresmann C, Nikonov S, Dumas P. 2000. The crystal structure of UUCG tetraloop. *J Mol Biol 304*:35-42.

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

Headey S, Huang H, Claridge J, Soares G, Dutta K, Schwalbe M, Yang D, Pascal S. 2007. NMR structure of stem-loop D from human rhinovirus-14. *RNA 13*:351-360.

Hendrix D, Brenner S, Holbrook S. 2005. RNA structural motifs: building blocks of a modular biomolecule. *Q Rev Biophys 38*:221-243.

Huang H, Nagaswamy U, Fox G. 2005. The application of cluster analysis in the intercomparison of loop structures in RNA. *RNA 11*:412-423.

Huppler A, Nikstad L, Allmann A, Brow D, Butcher S. 2002. Metal binding and base ionization in the U6 RNA intramolecular stem-loop structure. *Nat Struct Biol 9*:431-435.

Ihle Y, Ohlenschälger O, Häfner S, Duchardt E, Zacharias M, Seitz S, Zell R, Ramachandran R, Görlach M. 2005. A novel cGUUAg tetraloop structure with a

conserved yYNMGg-type backbone conformation from cloverleaf 1 of bovine enterovirus 1 RNA. *Nucleic Acids Res 33*:2003-2011.

Klein D, Moore P, Steitz T. 2004. The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *J Mol Biol 340*:141-177.

Legault P, Li J, Mogridge J, Kay L, Greenblatt J. 1998. NMR structure of the bacteriophage lambda N peptide/boxB RNA complex: recognition of a GNRA fold by an arginine-rich motif. *Cell 93*:289-299.

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

Leontis N, Lescoute A, Westhof E. 2006. The building blocks and motifs of RNA architecture. *Curr Opin Struct Biol 16*:279-287.

Leontis N, Westhof E. 2001. Geometric nomenclature and classification of RNA base pairs. *RNA 7*:499-512.

Leontis N, Westhof E. 2003. Analysis of RNA motifs. *Curr Opin Struct Biol 13*:300-308.

Lescoute A, Leontis N, Massire C, Westhof E. 2005. Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. *Nucleic Acids Res 33*:2395-2409.

Major F, Thibault P. 2006. RNA Tertiary Structure Prediction (Chapter 15). In: Lengauer T, ed. *Bioinformatics: From Genomes to Therapies*. Weinheim, Germany (2007): Wiley-VCH.

Melchers W, Zoll J, Tessari M, Bakhmutov D, Gmyl A, Agol V, Heus H. 2006. A GCUA tetranucleotide loop found in the poliovirus oriL by in vivo SELEX (un)expectedly forms a YNMG-like structure: Extending the YNMG family with GYYA. *RNA 12*:1671-1682.

Moore P. 1999. Structural motifs in RNA. *Annu Rev Biochem 68*:287-300.

Ohlenschläger O, Wöhnert J, Bucci E, Seitz S, Häfner S, Ramachandran R, Zell R, Görlach M. 2004. The structure of the stemloop D subdomain of coxsackievirus B3 cloverleaf RNA and its interaction with the proteinase 3C. *Structure 12*:237-248.

Proctor D, Schaak J, Bevilacqua J, Falzone C, Bevilacqua P. 2002. Isolation and characterization of a family of stable RNA tetraloops with the motif YNMG that participate in tertiary interactions. *Biochemistry 41*:12062-12075.

Schlünzen F, Zarivach R, Harms J, Bashan A, Tocilj A, Albrecht R, Yonath A, Franceschi F. 2001. Structural basis for the interaction of antibiotics with the peptidyl transferase centre in eubacteria. *Nature 413*:814-821.

Selmer M, Dunham C, Murphy Ft, Weixlbaumer A, Petry S, Kelley A, Weir J, Ramakrishnan V. 2006. Structure of the 70S ribosome complexed with mRNA and tRNA. *Science 313*:1935-1942.

Theimer C, Finger L, Feigon J. 2003. YNMG tetraloop formation by a dyskeratosis congenita mutation in human telomerase RNA. *RNA 9*:1446-1455.

Woese C, Winker S, Gutell R. 1990. Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc Natl Acad Sci U S A 87*:8467-8471.

TABLES

Table I : *Known YNMG motifs from literature. 7 RMN structure sets for a total of 113 models were analysed. All instances show a Watson-Crick closing base pair under the tetraloop.*

PDB	Models	Location	Sequnce	Description
1Q75	20	5-11	cUCAGg	DKC mutant P2b telomerase RNA
1RAW	10	21-25	cUUCGg	ATP binding aptamer
1RFR	20	13-18	uCACGg	SLD of coxsackievirus B3
1ROQ	10	5-10	uCACGg	Tetraloop of SLD of cocxsackievirus B3
1TXS	20	18-23	uUACGg	Consensus SLD of enterovirus and rhinovirus
1Z30	15	7-12	cGUUAg	Wt SLD of bovine enterovirus 1
2EVY	18	5-10	uGCUAg	Mutant hairpin of SLD of poliovirus oriL

Table II: *Partial list of YNMG motifs. These are the most interesting specimens, sometimes with insertion and/or involved in interaction, discussed in the text. All instances have a W/W closing pair.. a) Unless otherwise noted, the rRNA loops instances and their numbering are from* H. Marismortui. *for 23S and* T. Thermophilus *for 16S. b) Sequence of the instance. Insertions are underlined. When different sequences exist in crystals of other species, all are listed. EC* : E. Coli. *HM:* H. Marismortui. *TT:* T. Thermophilus. *c) RMSD distance to the reference UUCG motif, in Ångströms. For RMN model sets, the minimal and maximal distances are given. When the difference is significative, RMSD$^{(-2)}$ is indicated in parenthesis. b) Type of interaction. P-: with protein. RNA-: with nearby RNA.. GNRA- : with GNRA loop. –m: in minor groove. –M: in major groove. –B: Interaction with backbone. e) Summary of the interactions of the insertion. f) Despite similar location numbering, these two instances are not structurally equivalent and should be considered distinct locations. g) For these instances, neighboring RNA strands or proteins are present in the crystal but the interaction cannot be clearly established. h) This instance is formed of two distinct parts of the same rRNA strand. The detailed location is C883,U884,U561,C562,A563,G567.*

Location[a]	Sequence[b]	RMSD[c]	Interaction[d]	Insertion[e]
Hairpin bound to Drosophila Staufen ds RNA-binding domain (PDB 1EKZ)				
13-18	cUUCGg	0.87-2.58	P-m	-
tRNA domain of tmRNA in complex with smpB (PDB 1P6V)				
26-31	cUUCGg	1.43	P-m	-
Stemloop that binds RBD12 of mouse nucleolin (PDB 1QWA)				
8-13	uCCCGa	1.39-1.68	-	-
23S ribosomal RNA				
0'1769-'0'1774	cUUCGg	0.51	P-B, RNA-m	-
'0'136-'0'142[f]	cUUCGcg (HM)	2.02 (1.65)	GNRA-m	Interactions in same stem
B137-B142[f]	uUUCGa (EC)	1.20	P-	
'0'195-'0'201	cGCAAug	2.31 (2.16)	P-B, RNA-m, RNA-M	Bulging out to surface
0'670-'0'676	gAGUaUc (HM)	2.22	P-m	Tertiary RNA interaction
	cGAAuAg (EC)	2.05	P-(3' side)	Protein
'0'872-'0'878	uGAAAgg (HM,TT)	1.63	P-M, RNA-m	Protein and RNA
	uGAAAag (EC)	1.73		
16S ribosomal RNA				
A342-A347	cUACGg	0.70	GNRA-m	-
A419-A424	cUUCGg	0.64	[g]	-
A1449-A1454	cUACGg	0.50	[g]	-
A1134-A1140	gUUCGgc (TT)	1.71	[g]	Interactions in same stem
	gUCCGgc (EC)	1.94		
A883-A567	cU-UCAcugg[h]	2.8 (1.65)	P-M, RNA-m	-

Table III : *Tetraloops with nYNMGn sequence but not a YNMG fold.*

PDB	Location	Sequence	Closing pair	Description
1IE2	9-14	uCCCGa	W/W	In vitro sequence recognized by RBD12 of hamster nucleolin
2J00	A1090-A1095	uUAAGu	W/W	16S ribosomal RNA of Thermus Thermophilus
2J01	A542-A551	cCAAGg	W/W	23S ribosomal RNA of Thermus Thermophilus
2J01	A713-A718	gUGAGa	S/H	23S ribosomal RNA of Thermus Thermophilus
2J01	A826-A831	uUUAGg	W/W	23S ribosomal RNA of Thermus Thermophilus
2J01	B40-B45	uUCCGa	S/H	23S ribosomal RNA of Thermus Thermophilus
1S72	9'39-'9'44	uCCCGa	S/H	23S ribosomal RNA of Haloarcula Marismortui
1S72	0'2838-'0'2843	aCAAGa	S/W	23S ribosomal RNA of Haloarcula Marismortui
1U9S	A182-A187	gUAAGa	S/H	Specificity domain of A-type ribonuclease P
1NBS	119-124	uUUAGa	W/H	Specificity domain of B-type ribonuclease P
1HS2	4-9	uUAAGu	W/W	Hairpin loop found in *E. Coli*
1R4H	3-8	gCAAGc	W/W	Structure of the IIIc domain of GB Virus B IRES Element

Table S-I: *Complete list of YNMG motifs found in this study, including those with sequence insertions. Entries deviating from the normal nYNMGn sequence are in italics.*

PDB	Location	Sequence
1A3M	A11-B16	cUUCGg
1AUD	B28-B35	cUUCGg
1B36	A16-A21	cUUCGg
1BGZ	10-15	cUUCGg
1BYJ	A11-A16	cUUCGg
1C0O	A5-A10	cUUCGg
1D6K	B283-B293	cUUCGg
1DK1	B32-B37	cUUCGg
1DK1	B8-B13	cUUCGg
1EBQ	12-17	cUUCGg
1EBR	12-17	cUUCGg
1EBS	12-17	cUUCGg
1EKZ	13-18	cUUCGg
1F6X	A12-A17	cUUCGg
1F6Z	A12-A17	cUUCGg
1F78	A12-A17	cUUCGg
1F79	A12-A17	cUUCGg
1F7F	A12-A17	cUUCGg
1F7G	A12-A17	cUUCGg
1F7H	A12-A17	cUUCGg
1F7I	A12-A17	cUUCGg
1F7Y	B32-B37	cUUCGg
1F7Y	B8-B13	cUUCGg
1F7Y	B8-B13	cUUCGg
1FJG	*A1134-A1140*	*gUUCGgc*
1FMN	16-21	cUUCGg
1FYO	A11-A16	cUUCGg
1FYP	A11-A16	cUUCGg
1G70	A54-A64	cUUCGg

PDB	Location	Sequence
1HLX	8-13	cUUCGg
1I6U	C16-C21	cUUCGg
1I6U	D16-D21	cUUCGg
1IKD	7-12	cUUCGg
1JO7	A14-A19	cUUCGg
1K2G	A5-A10	cUUCGg
1KP7	A14-A19	cUUCGg
1KUQ	B32-B37	cUUCGg
1KUQ	B8-B13	cUUCGg
1KUQ	B8-B13	cUUCGg
1L1C	C13-C18	cUACGg
1M5L	A17-A22	cUUCGg
1M82	A10-A15	cUUCGg
1MFJ	A9-A14	cUACGg
1MFY	A14-A19	cUUCGg
1NBS	B152-B157	cUUCGg
1NKW	'0'1549-'0'1554	cCUCGg
1NKW	'0'1708-'0'1713	cUUCGg
1OSW	A8-A13	cUACGg
1P5M	A28-A33	cUUCGg
1P6V	26-31	cUUCGg
1PBR	11-16	cUUCGg
1Q75	5-11	cUCAGg
1QWA	A8-A13	uCCCGa
1RAW	21-25	cUUCGg
1RFR	13-18	uCACGg
1ROQ	5-10	uCACGg
1S72	'0'136-'0'142	cUUCGcg
1S72	0'1769-'0'1774	cUUCGg

PDB	Location	Sequence
1S72	*'0'195-'0'201*	*cGCAAug*
1S72	*0'670-'0'676*	*gAGUaUc*
1S72	*'0'872-'0'878*	*uGAAAgg*
1TLR	10-15	cUUCGg
1TXS	18-23	uUACGg
1ULL	A14-A19	cUUCGg
1UN6	E15-E64	cUACGg
1UN6	F15-F64	cUACGg
1XSG	A12-A17	cUUCGg
1XSH	A12-A17	cUUCGg
1XST	A12-A17	cUUCGg
1XSU	A12-A17	cUUCGg
1YKV	B220-B225	cUUCGg
1Z30	*7-12*	*cGUUAg*
1Z31	A270-A286	cUUCGg
1ZHO	B15-B20	cUUCGg
1ZHO	H15-H20	cUUCGg
2AU4	A20-A25	cUUCGg
2AW4	B1533-B1538	cUACGg
2AW4	B1691-B1696	cUUCGg
2AW7	A1028-A1033	cUUCGg
2AW7	A207-A212	cUUCGg
2EUY	A15-A20	gUUCGc
2EVY	*5-10*	*uGCUAg*
2J00	A1445-A1457	cUACGg
2J00	A342-A347	cUACGg
2J00	A419-A424	cUUCGg
2J01	A1691-A1696	cUUCGg

Table S-II: *All 4-nt loops in the 16S and 23S rRNA of H. Marismortui. For the sake of completeness, any type of base-base pairing is accepted between the closing pairs, discarding only the loops with backbone only or ribose only closing pair interactions. nYNMGn sequences are underlined. Entries in bold are those forming a YNMG fold.*

Location	Sequence	Closing pair
23S ribosomal RNA		
0'165-'0'170	aAACAu	Bs/H
0'252-'0'257	cUCACg	W/W
0'455-'0'460	aGUGAa	W/H
0'458-'0'463	gAACAa	S/S
0'468-'0'473	uGUGAa	W/W
0'506-'0'511	gAAAUa	S/H
0'576-'0'581	cGCGAg	W/W
0'690-'0'695	gGAAAc	W/W
0'733-'0'738	uUCAAg	W/W
0'804-'0'809	cGAAAg	W/W
0'838-'0'843	cCUACa	S/W
0'919-'0'924	uCGAAg	W/W
0'1197-'0'1202	gUAACa	S/H
0'1326-'0'1331	uGAAAa	W/W
0'1499-'0'1504	uUAAUa	W/H
0'1628-'0'1633	gGAAAc	W/W
0'1769-'0'1774	**cUUCGg**	W/W
0'1835-'0'1840	uAGUAa	H/H
0'1862-'0'1867	cGCAAg	W/W
0'1917-'0'1922	gUACAa	S/H
0'2069-'0'2074	uGCGGa	W/W
0'2248-'0'2253	cGGGAg	W/W
0'2301-'0'2306	aAAGAu	W/W
0'2365-'0'2370	gCAAAa	S/S
0'2411-'0'2416	cGAAAg	W/W
0'2611-'0'2616	gAGCUg	C8/H
0'2629-'0'2634	cGUGAg	W/W
0'2695-'0'2700	cGAGAg	W/W
0'2737-'0'2742	cGAGAg	W/W
0'2838-'0'2843	aCAAGa	S/W
0'2876-'0'2881	gGUAAc	W/W
9'89-'9'94	cGCGAg	W/W
16S ribosomal RNA		
A158-A163	gGAAAc	W/W
A190.B-A190.G	cCUUGg	W/W
A260-A265	gUAAUg	Bs/H
A296-A301	uGAGAg	W/W
A342-A347	**cUACGg**	W/W
A379-A384	cGCAAg	W/W
A419-A424	**cUUCGg**	W/W
Location	**Sequence**	**Closing pair**
A522-A527	cAGCCg	W/W
A691-A696	gUGAAa	S/H
A726-A731	cGAAGg	W/W
A862-A867	cUAACg	W/W
A897-A902	cGCAAg	W/W
A1012-A1017	uGAAAg	W/W
A1076-A1081	cGUGAg	W/W
A1090-A1095	uUAAGu	W/W
A1165-A1171	cGAAAg	W/W
A1265-A1270	gGCAAc	W/W
A1449-A1454	**cUACGg**	W/W
A1515-A1520	cGGAAg	W/W

FIGURES

Figure 12 : Reference cUUCGg motf (1F7Y). This instance of the fold is stabilized by a number of intra-loop interactions: G-C Watson-crick closing pair (light blue), U-C (O2 to N1/N2) (orange), U-C (O2' to O6) (green and brown), U-C (O2P to N4) (magenta), one water-coordinated interaction between G4 and G5 (red) and one intra-base, O2 to O2' on the cytosine. The second loop nucleotide, here an uridine, is not held by any base interactions and therefore free for tertiary contacts.

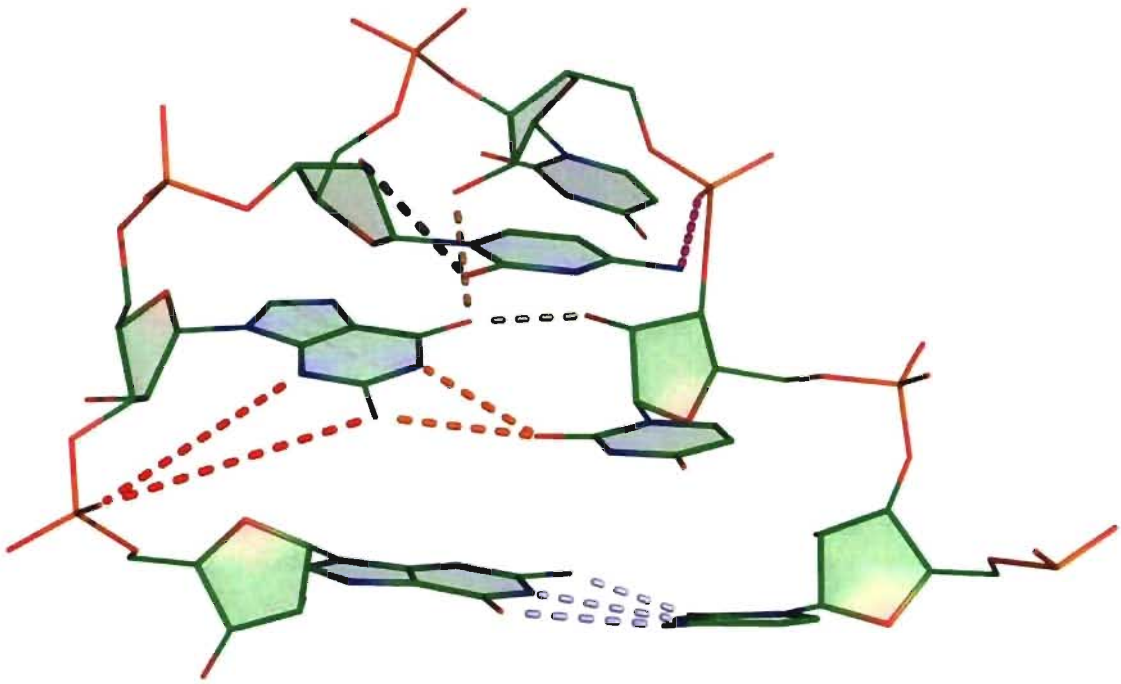


Figure 13 : 23S *H. Marismortui* loop 1769. The 4-nucleotide YNMG motif with it's closing C-G base pair (all in blue). In this instance, the second loop nucleotide of the motif, here an uridine, is interacting with two adenines (in magenta). The RNA backbone (orange) is exposed to ribosomal protein S37 (dark pink) which has several charged arginines (red) readily positioned for interaction. Relevant distances are shown with dotted black lines.

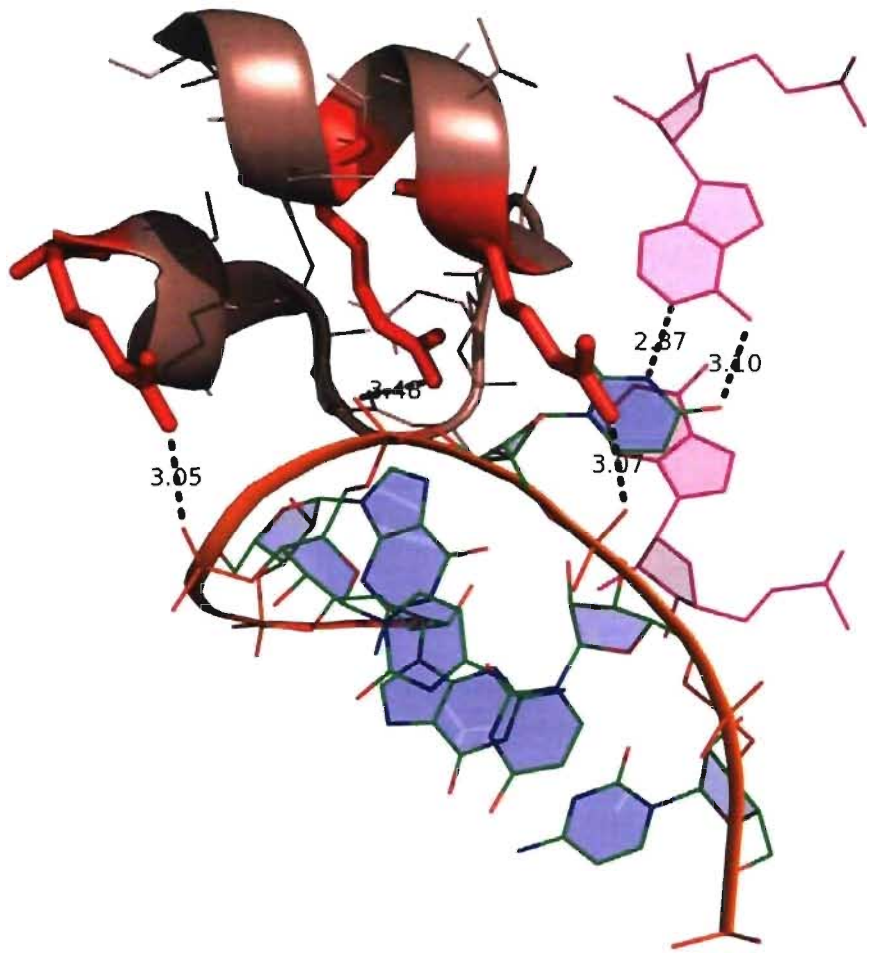


Figure 14 : 23S *H. Marismortui* loop 136. A) This YNMG motif (green) with sequence cUUCGcg is near a GNRA tetraloop (blue). The cytosine insertion (not shown in this figure), is bulged out and stacks with a lower cytosine. Water molecules are present in the crystal between the two hairpins, possibly mediating the interaction. B) The important intra-loop interactions typical of the YNMG motif, and the fourth loop nucleotide in *syn*, are conserved despite the insertion. C) The motif from E. Coli (green with phosphate backbone in orange) is with a protein (dark pink)

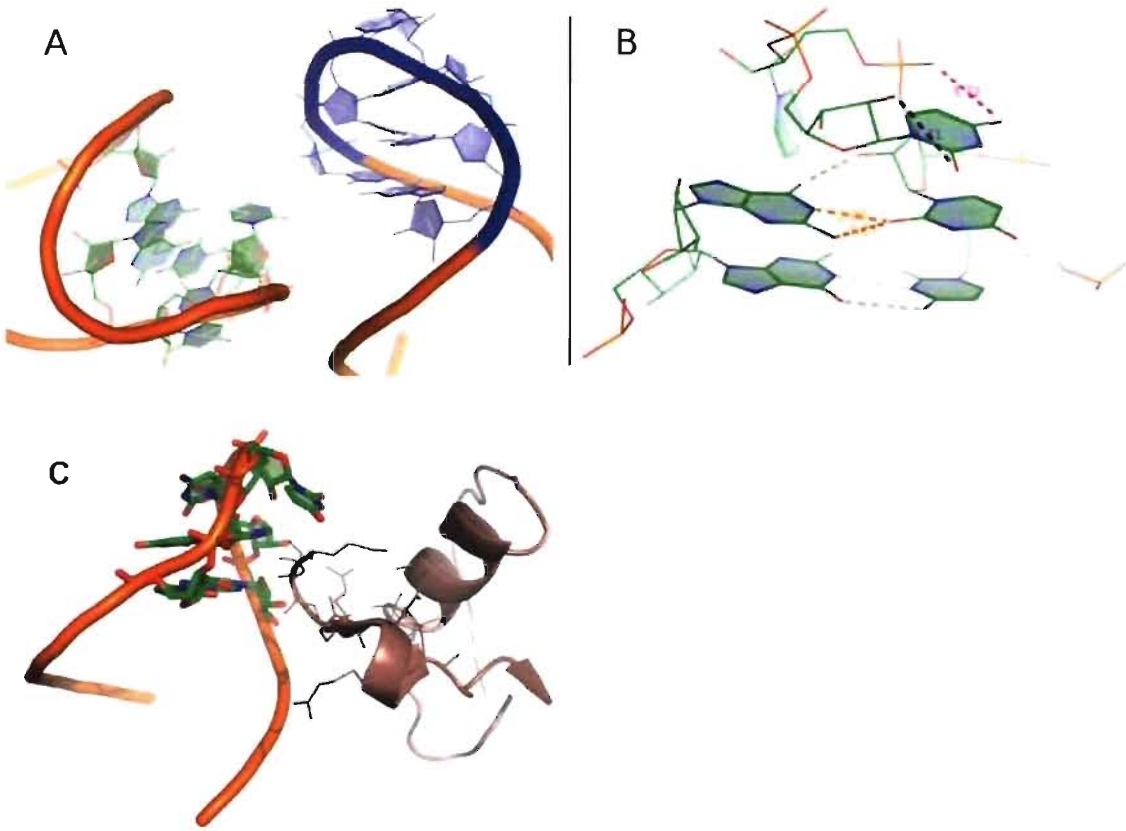


Figure 15 : 23S *H. Marismortui* loop 670 (gAGUaUc). A) Stereo image showing the interactions. In this instance, the insertion (orange) is inside the tetraloop (red) rather than at its usual position after the loop just before the closing base pair (G670 in magenta). The protein (in green) is on the minor groove side, and the insertion is paired with C36 (yellow). B) Superposition of the two instances of loop 670 from *H. Marismortui* (sequence gAGUaUc, in red) and *T. Themophilus* (cGAAuAg, in green) with the other YNMG-fold candidate with a G-C closing pair, loop 1134 of 16S rRNA (gUYCGgc, in blue). Despite the sequence difference, all three loops are structurally similar. C) The internal interactions stabilizing the YNMG fold find their equivalent in loop 670, as seen here in *T. Thermophilus*. The bond colors are as in figure 1.

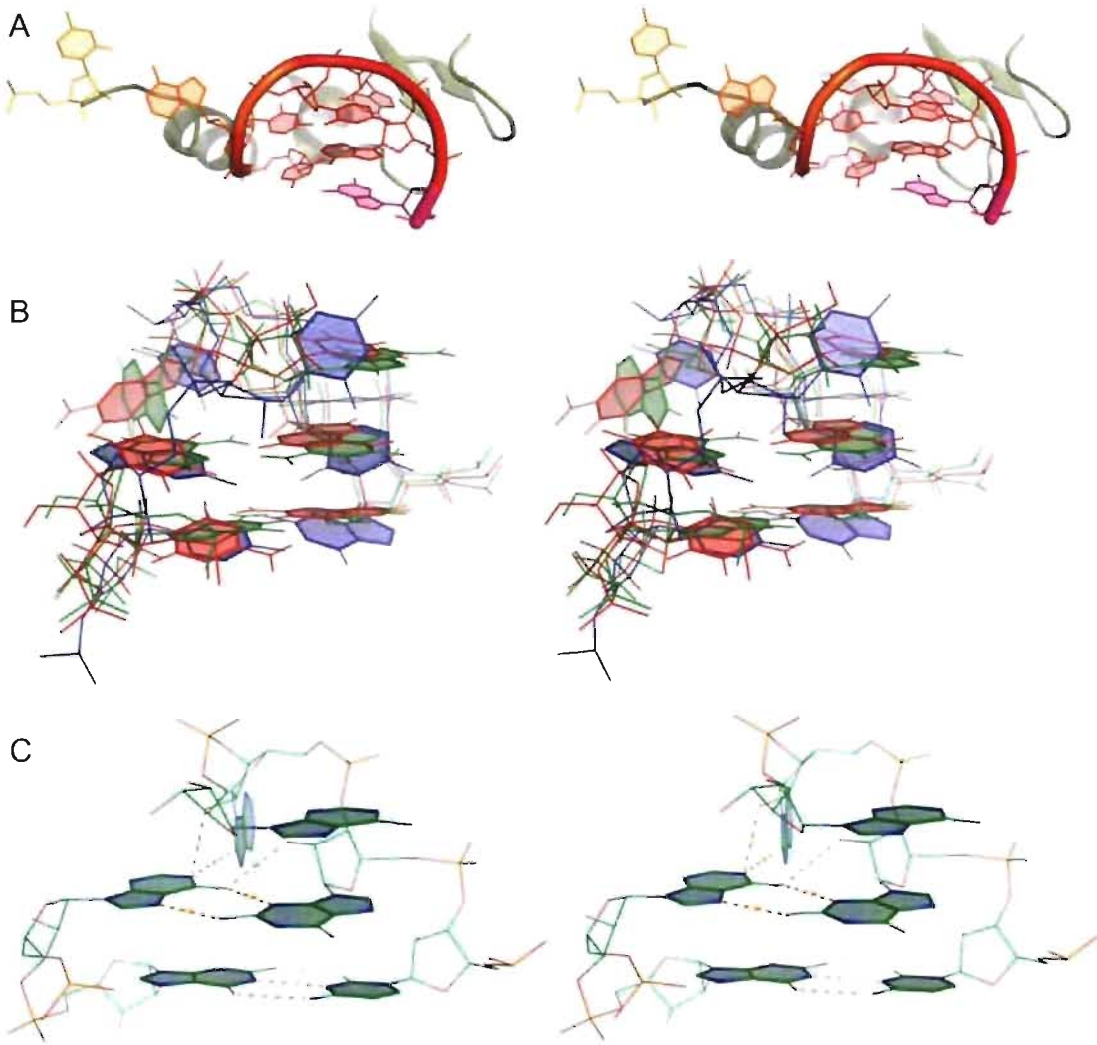


Figure 16 : 23S *H. Marismortui* loop 872 (uGAAAgg). This YNMG motif is mimicking as a GNRA sequence, but is still a convincing YNMG fold candidate. A) The internal stackings and L_1-L_2 pairings are conserved. B) Superposition of loop 872 (in green) with the reference motif (PDB 1F7Y, in red). Only the position of the second loop nucleotide shows a significant difference in position, as expected for the YNMG fold.

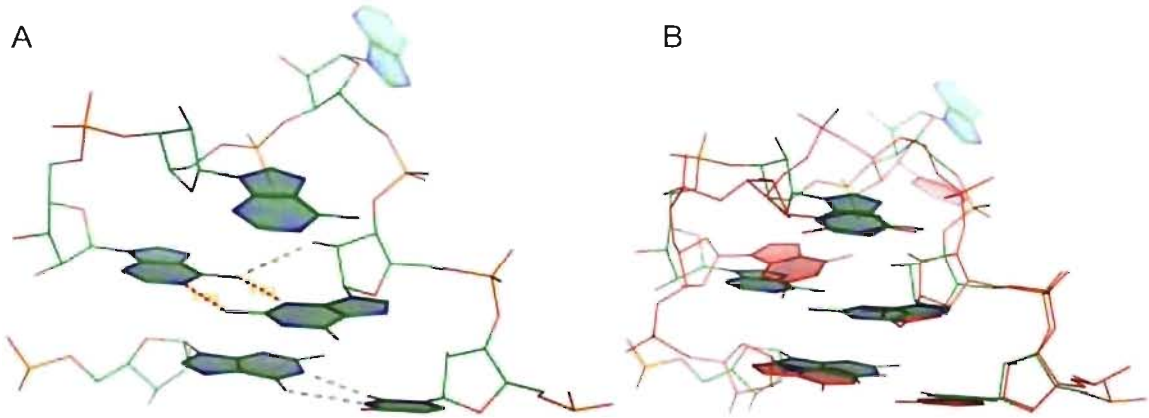


Figure 17 : 16S *T. Thermophilus* loop 343 in interaction with a GNRA motif (in blue). Only the two adenines of the GNRA are used and show a higher degree of conservation than the rest of the loop. They form pairings with the closing pair (C-G, in yellow and red respectively) of the YNMG.

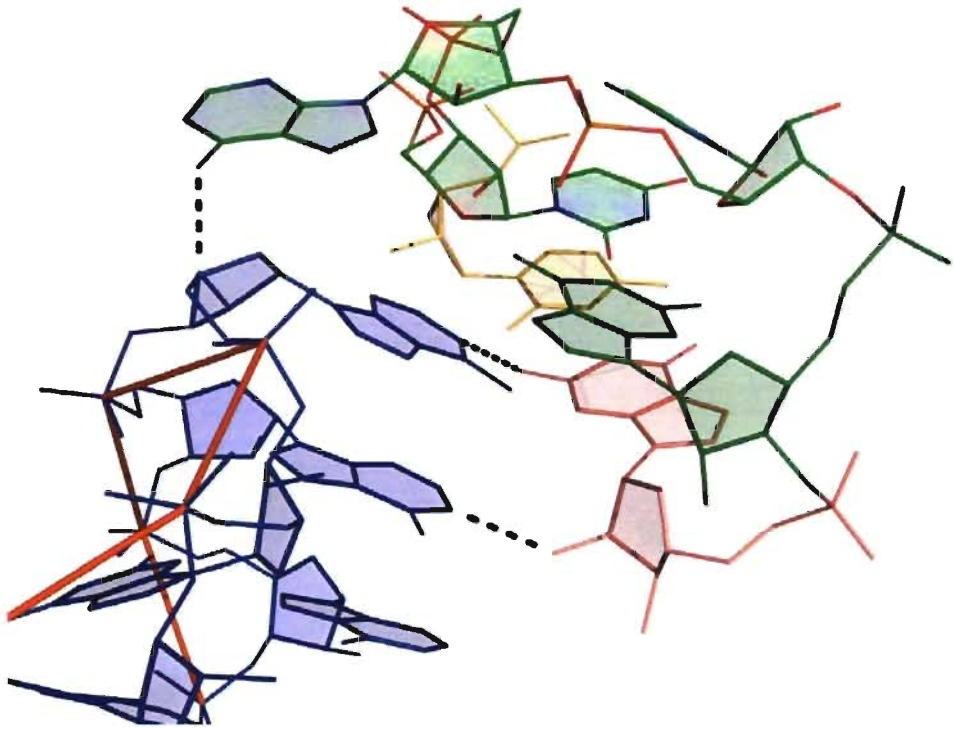


Figure 18 : 16S *T. Thermophilus* loop 883. A) This instance (in red) is embedded deep inside the core of the ribosome, interacting with nearby proteins (in green) and RNA strands (orange). B) The motif is stabilized by internal interactions and has its fourth nucleotide in *syn*, as in 1F7Y. Hydrogen bonds are indicated in dotted lines with colors as in figure 1 to show equivalence.

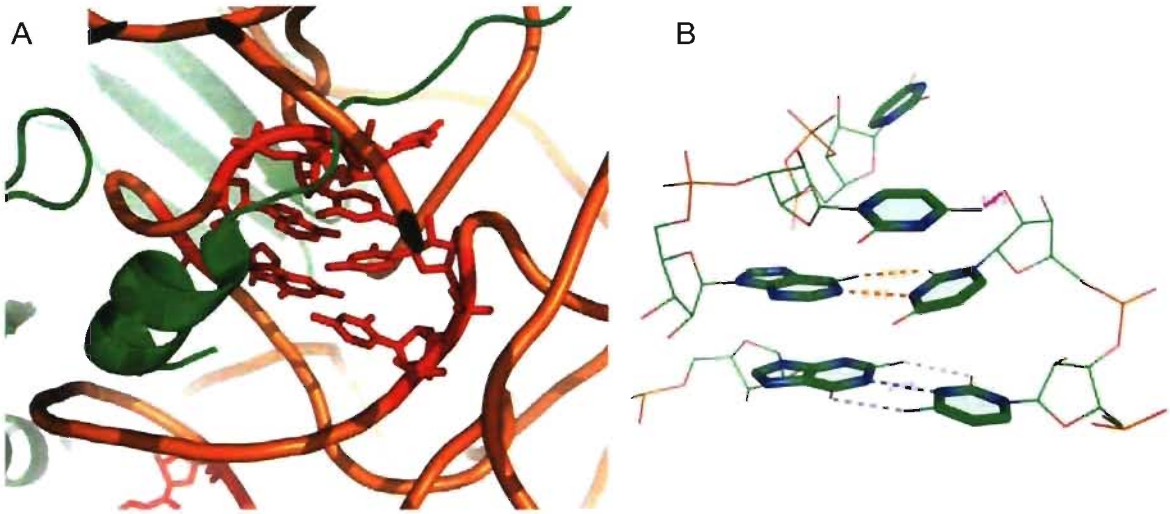


Figure 19 : YNMG motifs in non-ribosomal RNA. In this figure, loop nucleotide are colored in the same way: closing pair (orange), L_1 (green), L_2 (red), L_3 blue and L_4 (yellow). A) Loop 13-18 of is part of an RNA hairpin bound to *Drosohphila* Staufen protein. Over the 36 RMN models of this structure, the second loop nucleotide (L_2) moves to stay in contact with it's protein partner. B) In PDB 1RAW, which shows an ATP-binding aptamer in complex with ATP, both middle loop nucleotides (L_2 and L_3) show large movement amplitude.

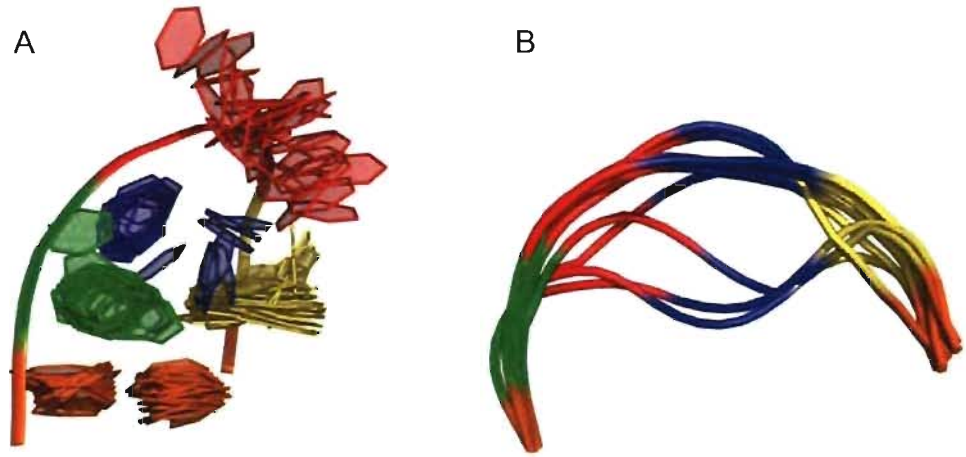


Figure S-1 : 23S rRNA of H. Marismortui showing the YNMG folds. The only genuine tetraloop is shown in blue. Three YNMG pentaloops forming a YNMG fold are in red. Proteins chains are in green, and the RNA phosphodiester backbone in orange.

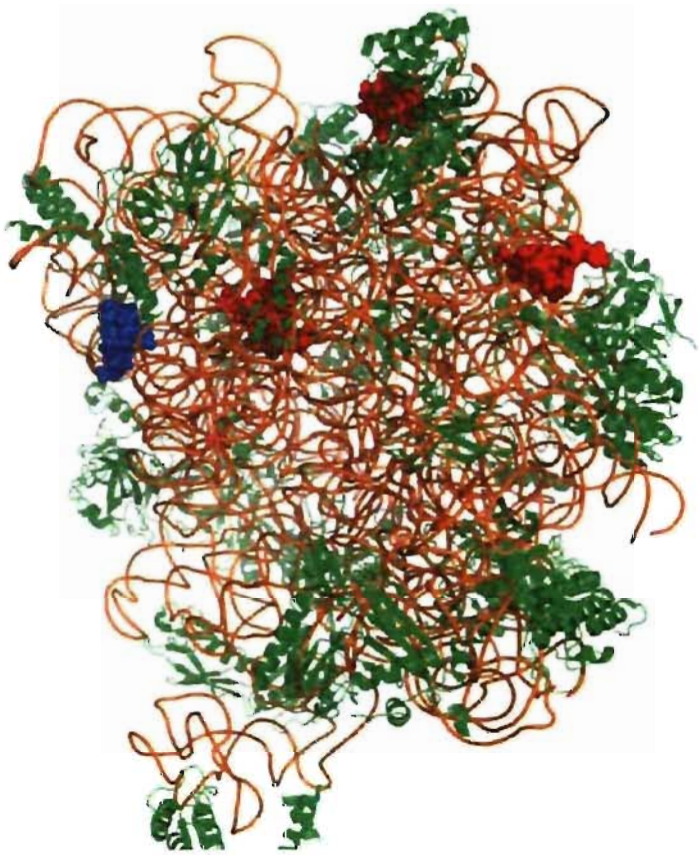
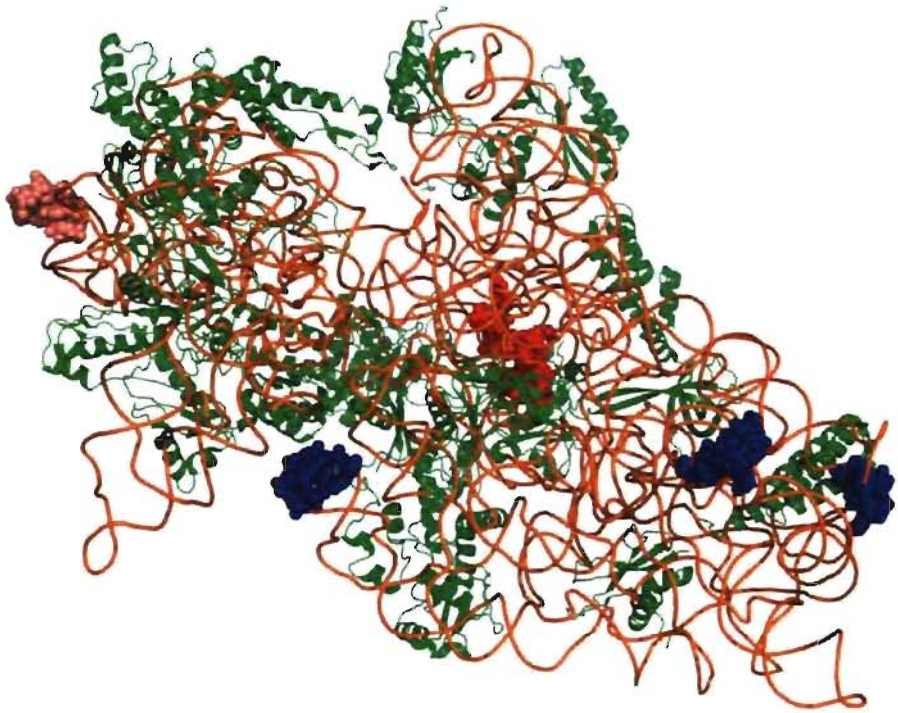


Figure S-2 : 16S rRNA of H. Marismortui showing the YNMG folds. The SSU contains three tetraloop YNMG folds (in blue) and two YNMG-like (red and pink). The rRNA phospodiester backbone is in orange and protein chains in green.



Chapitre 3: Article – MC-View: An online tool for RNA motif visualisation

MC-View est un logiciel pour faciliter la visualisation de motifs dans les structures d'ARN, en les localisant et permettant de voir la structure entière avec les motifs annotés et colorés pour un repérage facile. Le programme repose sur la librairie *mccore*, développée au laboratoire. *MC-View* à été réalisé lors d'une collaboration avec Emmanuelle Permal, une étudiante au doctorat aussi sous la direction de François Major, pour des travaux d'études de motifs dans les virus à génome d'ARN. L'idée du logiciel est d'elle, et j'en ai fait l'implantation ainsi qu'une partie de la production de résultats pour ses analyses, dont les résultats sont dans un autre manuscrit placé en annexe de ce mémoire. Romain Rivière à réalisé la première version du code pour produire les structures secondaires dans un fichier PDF à l'aide de l'algorithme de naview (Bruccoleri & Heinrich, 1988). J'ai repris ce code et l'ai mis à niveau et intégré dans *MC-View* de façon à obtenir aussi une vue annotée et colorée de la structure secondaire.

Ce chapitre présente un article court suivant le format de note d'application (« application note ») de deux pages, qui sera soumis à la revue *Bioinformatics*

Structural bioinformatics

MC-View: An online tool for RNA motif visualisation

Louis-Philippe Lavoie, Emmanuelle Permal, Romain Rivière and François Major*

Institute for Research in Immunology and Cancer, Computer Science and Operations Research De-partment, Université de Montréal, Montréal, QC, Canada, H3C 3J7

ABSTRACT

Summary: MC-View is an online tool that allows for searching and visualisation of structural motifs in submitted RNA structures. It uses predefined structural motifs such as GNRA tetraloop or sarcin-ricin motif and searches for homologous structures in a PDB file. As result, it outputs two files that allow study of RNA motifs in their whole molecule context: a pdf file of the RNA graph of the molecule and a PyMol script with all motifs annotations.

Availability: MC-View is available via http://major.iric.ca/mcview.

Contact: [redacted]

1 INTRODUCTION

RNA structural motifs are the building blocks of RNA molecules (Hendrix et al., 2005; Leontis et al., 2006). A large number has been characterized so far and their study becomes more refined each year. Different methods of discovery and analysis have been proposed, but one common thread is that the study of RNA motifs is the key to understanding the elusive nature and mechanisms of RNA as a whole. While fast, fully automated tools begin to appear, visualisation is still one of the simplest and yet most effective re-course to gain deeper insight into the structures. To this end, a tool to rapidly produce complete renderings of studied structures should be in the arsenal of every structural biologist.

Here we present an online tool, MC-View that allows quickly finding and visualizing complex RNA structural motifs in any supplied structure. Using a flexible graph annotation to describe nucleotides and their interactions (see **figure 1a**), the input molecule is searched for all instances of the selected motifs. The result is a fully annotated secondary structure diagram (see **figure 1b**) and a PyMOL script overlaying the initial three-dimensional structure with the studied motifs, allowing visualization of the motif (see **figure 1c, d**). Putative protein-RNA hydrogen bonds (h-bonds) are also identified.

2 METHODS

Structure annotation. Automated annotation of tertiary RNA structure files has been introduced independently by the groups of Westhof and Major, in RNAView (Yang et al., 2003) and MC-Annotate (Lemieux & Major, 2002), respectively. To compute complete structural annotations, MC-View uses MC-Annotate which is itself built upon the MC-Core open source library (http://mccore.sourceforge.net). Structures (in the PDB file format) are annotated and transformed into a graph where nucleotides are vertices and relationships (pairing, stacking, and adjacency) are edges. Additionally, since MC-Core can identify putative h-bonds between RNA nucleotides and amino acids, these bonds are presented in MC-View's re-sults.

Secondary structure extraction. The Naview algorithm (Bruccoleri & Heinrich, 1988) is used to compute nucleotide coordinates in the secondary structure diagrams. MC-View then considers all types of base pairings to draw the final secondary structure, including non-canonical/non-Watson-Crick and tertiary interactions.

Identification of structural motifs. The structural motifs are also translated into graph representation, and then searched for in the complete structure using the Ullman isomorphism algorithm (Ullmann, 1976). The motif descriptor syntax supports a wide range of relationship and ribose conformations, and is easily tailored to any motif that can be described in terms of either sequence or structure.

Input options. The user is required to submit a PDB-formatted structure file and can select from a list of predefined motifs to visualize. A custom motif can be added to the selection, as well as a custom color definition and coordinate file for the secondary structure. It is also possible to ignore types of relationships (pairings, stacking, adjacency or non-standard pairings between a base and the backbone). Several options are presented for further customizations of the output. A complete description of the syntax used to describe the search motifs is available on the MC-View web page.

3 RESULTS

MC-View uses graph theory to manipulate structures algorithmically, so that even the largest ribosomes can be processed in a matter of seconds. All files from the Protein Data Bank are pre-annotated and kept locally on the server (updated weekly), speeding up the computation even more when these structures are used. The output of MC-View is in two parts:

Secondary structure diagram. First, a secondary structure diagram is output as a Portable Document Format (PDF) file (see **figure 1b**). This format was selected because it allows for rich annotation of the diagram, and the vector graphics can be rendered in any display size without loss of quality or easily modified with any software capable of editing PDF vector data.

The diagram can include tertiary interactions as found in the original PDB structure submitted. Additionally if the option is selected, each nucleotide in the diagram is annotated with the full list of its relations to other nucleotides, using the nomenclature established by Leontis & Westhof (2001) for base pairs and by Major & Thibault (2007) for stacking.

PyMOL script. The second part of the output is a PyMOL script containing all the necessary commands to annotate the original PDB structure with an extensive set of sub-motifs and interesting features, including the structural motifs submitted as input:

- Nucleobases, backbone, ribose, phospates, nucleic acids, proteins, ions/others and protein/RNA h-bond interactions are all grouped into distinct selections.

- A large number of pre-defined structural motifs, inspired from the SCOR database (Klosterman et al., 2002), are also defined as PyMoL selections.

Application. MC-View is an easy to use tool to rapidly extract and visualize structurally important regions of RNA. Secondary structure diagrams can be studied for differences with tertiary structures to reach better understanding and predictions of RNA structures. The inclusion of protein-RNA hydrogen bonds in the PyMOL display also positions MC-View as a useful starting tool for this area of research.

4 FUTURE DEVELOPMENTS

MC-View is part of the MC-Tools suite for RNA structural analy-sis, which is actively developed in the Theoretical and Computer Science Biology Lab. Future plans include a better batch process-ing of both input structures and motif descriptors, and integration with other tools for structural annotation being developed in the lab.

ACKNOWLEDGMENTS

FM is a CIHR investigator and a member of the Institute for Research in Immunology and Cancer and of the Centre Robert-Cedergren. LPL holds a CIHR scholarship to support higher educa-tion in bioinformatics (biT program). This work is supported by CIHR grant MT-14604 to FM.

Conflict of Interest : none declared.

REFERENCES

Bruccoleri, R. and Heinrich, G. (1988) An improved algorithm for nucleic acid secon-dary structure display., Comput Appl Biosci, 4, 167-173.

Hendrix, D., Brenner, S. and Holbrook, S. (2005) RNA structural motifs: building blocks of a modular biomolecule., Q Rev Biophys, 38, 221-243.

Klosterman, P., Tamura, M., Holbrook, S. and Brenner, S. (2002) SCOR: a Structural Classification of RNA database., Nucleic Acids Res, 30, 392-394.

Lemieux, S. and Major, F. (2002) RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire, Nucleic Acids Res, 30, 4250-4263.

Leontis, N., Lescoute A., and Westhof, E. (2006) The building blocks and motifs of RNA architecture., Cur Op Struct Biol, 16, 279-287.

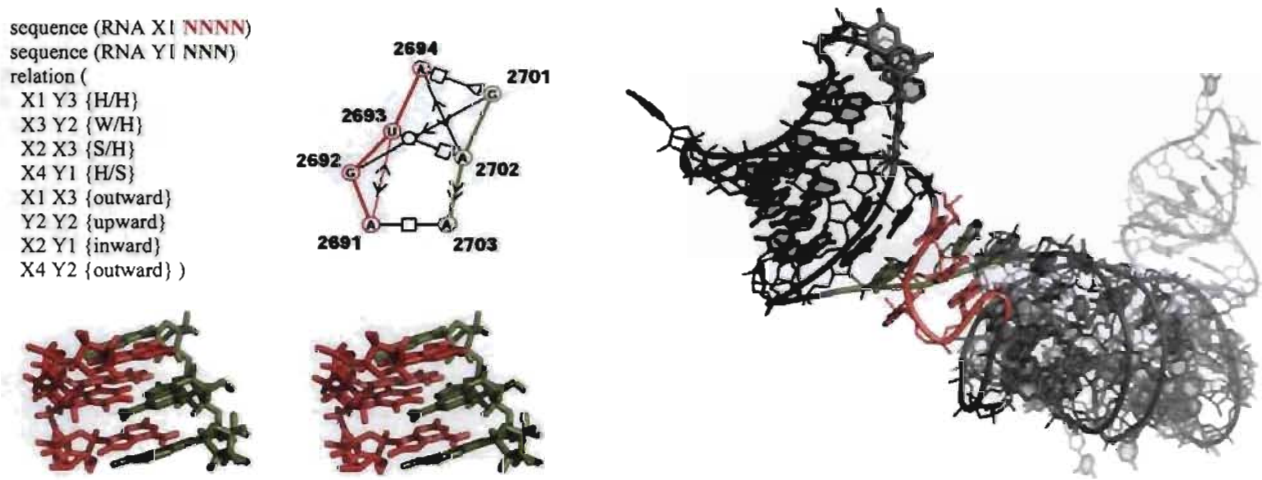
Leontis, N. and Westhof, E. (2001) Geometric nomenclature and classification of RNA base pairs, RNA, 7, 499-512.

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

Ullmann, J. (1976) An Algorithm for Subgraph Isomorphism: ACM Press New York, NY, USA. pp 31-42.

Yang H, Jossinet F, Leontis N, Chen L, Westbrook J, Berman H, Westhof E. 2003. Tools for the automatic identification and classification of RNA base pairs. Nu-cleic Acids Res 31450-3460.

Figure 1. The sarcin-ricin motif. A) Input graph descriptor. This motif is composed of two RNA strands, X and Y of 4 (red) and 3 (green) nucleo-tides respectively. The base pairing interactions are described using the base nomenclature of Leontis & Westhof (2001). The stacking interactions are described using the Major & Thibault nomenclature (2007). B) The corresponding sarcin-ricin RNA graph. C) Stereo view of an occurrence of the sarcin-ricin motif found in 23S rRNA of H. Marismortui. The descrip-tor and the motif found can be described by isomorphic graphs, thus the match. D) The motif inside the 5S rRNA of H. Marismortui.



Conclusion

"It is not because things are difficult that we do not dare; It is because we do not dare that things are difficult."

- Seneca, dans *Epistualea Morales*

Dans ce mémoire j'ai exposé une partie de mes travaux de maîtrise sur l'analyse de structures d'ARN.

À partir de simples recherches pour approfondir mes connaissances, puis une période de travail avec les outils logiciels du laboratoire, j'ai terminé avec la création de mes propres outils pour faciliter l'analyse en général et réalisé l'analyse d'un des motifs d'ARN parmi les plus fréquents. J'ai aussi fait des collaborations avec d'autres membres du laboratoire, participé à l'écriture de leurs manuscrits et présenté mes principaux résultats dans deux manuscrits qui seront soumis prochainement pour publication. Les outils et méthodes développées lors des analyses sont facilement transposables à d'autres études de structure.

Malgré toutes les études sur les boucles YNMG et la sous-classe UNCG, analysant leur thermostabilité, leur rôle, distribution, conservation, etc. la découverte de nouvelles instances avec des insertions en séquence et des déviations importantes dans la séquence démontre bien que l'ARN est très variable. Cette flexibilité est un des aspects qu'il nous faudra comprendre pour bien saisir la nature de ces molécules dans son ensemble.

Problèmes rencontrés

Outre les problèmes reliés aux données et à l'annotation tels qu'exposés dans l'introduction de ce mémoire, j'ai eu aussi à faire face à certains autres imprévus qui sont somme toute inévitables dans un contexte de recherche et développement. Certains des outils logiciels utilisés au laboratoire sont encore en évolution et des instabilités, voire même des incompatibilités apparaissaient de temps à autre et m'ont obligé à recréer le jeu complet de résultat à plus d'une reprise.

Perspectives

Des travaux récents d'autres membres du laboratoire (Parisien & Major, article soumis) viennent de démontrer qu'il est possible de prédire la structure secondaire et même tertiaire de petites molécules d'ARN en se basant sur les fragments déjà observés dans les structures connues actuellement. Cette approche apporte une confirmation supplémentaire à la modularité des structures d'ARN, et confirme la pertinence de l'approche par cycles favorisée par notre laboratoire.

Il serait intéressant d'appliquer cette méthode à quelques unes des nombreuses séquences disponibles de tige-boucle D de virus, ou même le domaine I en entier, pour essayer de trouver une structure commune au-delà de la boucle de terminaison YNMG. La tige-boucle contient une boucle interne de pyrimidines qui est aussi hautement conservée et pourrait être importante pour la liaison avec la protéine virale (Zell et al. 2002). La structure secondaire actuellement utilisée est prédite par Mfold (Zuker 2003), qui ne considère pas les appariements non-canoniques mais une structure disponible de la tige-boucle complète (Du et al. 2004) montre des appariements non-canoniques entre les pyrimidines. L'approche de Parisien & Major est maintenant capable de prédire les appariements non-canoniques avec une bonne fiabilité.

Une deuxième avenue à approfondir dans des travaux futurs serait la question d'équivalence de certains fragments d'ARN. Lors d'un projet de structure d'ARN effectué en parallèle à mes travaux pour le cadre d'un cours, j'ai remarqué que pour des structures ayant des régions de configuration similaires, les cycles présents étaient souvent différents. De plus, certains cycles sont plus souvent associés (remplacés) avec d'autres et il serait donc possible de déterminer des équivalences « préférables ». Cette idée s'apparente à l'isostéricité telle que définie pas Lescoute et al (2006) mais serait étendue à des plus gros fragments. Aussi, les techniques de modélisation par cycles actuelles choisissent les cycles basés sur leur fréquence et leur complémentarité dans l'assemblage en cours de construction, mais des règles de sélection additionnelles pourraient accélérer le processus

Bibliographie

Allain F, Varani G. 1995. Structure of the P1 helix from group I self-splicing introns. *J Mol Biol 250*:333-353.

Altona C, Sundaralingam M. 1972. Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. *J Am Chem Soc 94*:8205-8212.

Antao V, Lai S, Tinoco IJ. 1991. A thermodynamic study of unusually stable RNA and DNA hairpins. *Nucleic Acids Res 19*:5901-5905.

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

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

Bernstein B, Meissner A, Lander E. 2007. The mammalian epigenome. *Cell 128*:669-681.

Bernstein E, Allis C. 2005. RNA meets chromatin. *Genes Dev 19*:1635-1655.

Bird A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev 16*:6-21.

Bruccoleri R, Heinrich G. 1988. An improved algorithm for nucleic acid secondary structure display. *Comput Appl Biosci 4*:167-173.

Cannone J, Subramanian S, Schnare M, Collett J, D'Souza L, Du Y, Feng B, Lin N, Madabusi L, Müller K, Pande N, Shang Z, Yu N, Gutell R. 2002. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs*BMC Bioinformatics 3*:2.

Crick F. 1958. On protein synthesis. *Symp Soc Exp Biol 12*:138-163.

Crick F. 1970. Central dogma of molecular biology. *Nature 227*:561-563.

Dickerson R, Ng H. 2001. DNA structure from A to B. *Proc Natl Acad Sci U S A 98*:6986-6988.

DeLano W. 2002. The PyMOL Molecular Graphics System. *DeLano Scientific, Palo Alto, CA, USA.*

Dror O, Nussinov R, Wolfson H. 2005. ARTS: alignment of RNA tertiary structures. *Bioinformatics 21 Suppl 2*:ii47-53.

Du Z, Yu J, Andino R, James T. 2003. Extending the family of UNCG-like tetraloop motifs: NMR structure of a CACG tetraloop from coxsackievirus B3. *Biochemistry 42*:4373-4383.

Du Z, Yu J, Ulyanov N, Andino R, James T. 2004. Solution structure of a consensus stem-loop D RNA domain that plays important roles in regulating translation and replication in enteroviruses and rhinoviruses. *Biochemistry 43*:11959-11972.

Ennifar E, Nikulin A, Tishchenko S, Serganov A, Nevskaya N, Garber M, Ehresmann B, Ehresmann C, Nikonov S, Dumas P. 2000. The crystal structure of UUCG tetraloop. *J Mol Biol 304*:35-42.

Ferrè F, Ponty Y, Lorenz W, Clote P. 2007. DIAL: a web server for the pairwise alignment of two RNA three-dimensional structures using nucleotide, dihedral angle and base-pairing similarities. *Nucleic Acids Res 35*:W659-668.

Fire A, Xu S, Montgomery M, Kostas S, Driver S, Mello C. 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature 391*:806-811.

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

Gilbert W. 1986. Origin of life: The RNA world. Nature. pp 618.

Ginalski K, Grishin N, Godzik A, Rychlewski L. 2005. Practical lessons from protein structure prediction. *Nucleic Acids Res 33*:1874-1891.

Greenberg S. 2003. Respiratory consequences of rhinovirus infection. *Arch Intern Med 163*:278-284.

Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. 1983. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell 35*:849-857.

Harms J, Schluenzen F, Zarivach R, Bashan A, Gat S, Agmon I, Bartels H, Franceschi F, Yonath A. 2001. High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell 107*:679-688.

Harrison A, South D, Willett P, Artymiuk P. 2003. Representation, searching and discovery of patterns of bases in complex RNA structures. *J Comput Aided Mol Des 17*:537-549.

Headey S, Huang H, Claridge J, Soares G, Dutta K, Schwalbe M, Yang D, Pascal S. 2007. NMR structure of stem-loop D from human rhinovirus-14. *RNA 13*:351-360.

Hendrix D, Brenner S, Holbrook S. 2005. RNA structural motifs: building blocks of a modular biomolecule. *Q Rev Biophys 38*:221-243.

Hershkovitz E, Sapiro G, Tannenbaum A, Williams L. 2006. Statistical analysis of RNA backbone. *IEEE/ACM Trans Comput Biol Bioinform 3*:33-46.

Heus H, Pardi A. 1991. Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science 253*:191-194.

Hoffmann B, Mitchell G, Gendron P, Major F, Andersen A, Collins R, Legault P. 2003. NMR structure of the active conformation of the Varkud satellite ribozyme cleavage site. *Proc Natl Acad Sci U S A 100*:7003-7008.

Horton JD. 1987. A Polynomial-Time Algorithm to Find the Shortest Cycle Basis of a Graph. *SIAM Journal on Computing 16*:358-366.

Huang X, Ali H. 2007. High sensitivity RNA pseudoknot prediction. *Nucleic Acids Res 35*:656-663.

IHGSC. 2004. Finishing the euchromatic sequence of the human genome. *Nature 431*:931-945.

Ihle Y, Ohlenschlï¿½ger O, Hï¿½fner S, Duchardt E, Zacharias M, Seitz S, Zell R, Ramachandran R, Gï¿½rlach M. 2005. A novel cGUUAg tetraloop structure with a conserved yYNMGg-type backbone conformation from cloverleaf 1 of bovine enterovirus 1*Nucleic Acids Res 33*:2003-2011.

Janowski B, Younger S, Hardy D, Ram R, Huffman K, Corey D. 2007. Activating gene expression in mammalian cells with promoter-targeted duplex RNAs. *Nat Chem Biol 3*:166-173.

Jones P, Baylin S. 2007. The epigenomics of cancer. *Cell 128*:683-692.

Jucker F, Pardi A. 1995. GNRA tetraloops make a U-turn. *RNA 1*:219-222.

Kabsch W. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A 32*:922-923.

Krasilnikov A, Xiao Y, Pan T, Mondragï¿½n A. 2004. Basis for structural diversity in homologous RNAs*Science 306*:104-107.

Kruger K, Grabowski P, Zaug A, Sands J, Gottschling D, Cech T. 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell 31*:147-157.

Lemieux S, Major F. 2002. RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. *Nucleic Acids Res 30*:4250-4263.

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

Leontis, N., Westhof, E. 2001. Geometric nomenclature and classification of RNA base pairs. *RNA 7*:499-512.

Leontis N, Westhof E. 2003. Analysis of RNA motifs. *Curr Opin Struct Biol 13*:300-308.

Leontis N, Lescoute A, Westhof E. 2006. The building blocks and motifs of RNA architecture. *Curr Opin Struct Biol 16*:279-287.

Lescoute A, Leontis N, Massire C, Westhof E. 2005. Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments. *Nucleic Acids Res 33*:2395-2409.

Li E. 2002. Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet 3*:662-673.

Li L, Okino S, Zhao H, Pookot D, Place R, Urakami S, Enokida H, Dahiya R. 2006. Small dsRNAs induce transcriptional activation in human cells. *Proc Natl Acad Sci U S A 103*:17337-17342.

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

Liu J, Gough J, Rost B. 2006. Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLoS Genet 2*:e29.

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.

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

Mattick J. 2001. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep 2*:986-991.

Melchers W, Zoll J, Tessari M, Bakhmutov D, Gmyl A, Agol V, Heus H. 2006. A GCUA tetranucleotide loop found in the poliovirus oriL by in vivo SELEX (un)expectedly forms a YNMG-like structure: Extending the YNMG family with GYYA. *RNA 12*:1671-1682.

Moore P. 1999. Structural motifs in RNA. *Annu Rev Biochem 68*:287-300.

Murray L, Arendall Wr, Richardson D, Richardson J. 2003. RNA backbone is rotameric. *Proc Natl Acad Sci U S A 100*:13904-13909.

Murray L, Richardson J, Arendall W, Richardson D. 2005. RNA backbone rotamers--finding your way in seven dimensions. *Biochem Soc Trans 33*:485-487.

Napoli C, Lemieux C, Jorgensen R. 1990. Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. *Plant Cell 2*:279-289.

Ng K, Pullirsch D, Leeb M, Wutz A. 2007. Xist and the order of silencing. *EMBO Rep 8*:34-39.

Nissen P, Ippolito J, Ban N, Moore P, Steitz T. 2001. RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc Natl Acad Sci U S A 98*:4899-4903.

Ohlenschläger O, Wöhnert J, Bucci E, Seitz S, Häfner S, Ramachandran R, Zell R, Görlach M. 2004. The structure of the stemloop D subdomain of coxsackievirus B3 cloverleaf RNA and its interaction with the proteinase 3C. *Structure 12*:237-248.

Olivier C, Poirier G, Gendron P, Boisgontier A, Major F, Chartrand P. 2005. Identification of a conserved RNA motif essential for She2p recognition and mRNA localization to the yeast bud. *Mol Cell Biol 25*:4752-4766.

Parisien M, Major F. 2005. A new catalog of protein beta-sheets. *Proteins 61*:545-558.

Pennisi E. 2007. Genetics. Working the (gene count) numbers: finally, a firm answer? *Science 316*:1113.

Proctor D, Schaak J, Bevilacqua J, Falzone C, Bevilacqua P. 2002. Isolation and characterization of a family of stable RNA tetraloops with the motif YNMG that participate in tertiary interactions. *Biochemistry 41*:12062-12075.

Prusiner S. 1982. Novel proteinaceous infectious particles cause scrapie. *Science 216*:136-144.

Qiu S, Adema C, Lane T. 2005. A computational study of off-target effects of RNA interference. *Nucleic Acids Res 33*:1834-1847.

Recht M, Douthwaite S, Puglisi J. 1999. Basis for prokaryotic specificity of action of aminoglycoside antibiotics. *EMBO J 18*:3133-3138.

Romano N, Macino G. 1992. Quelling: transient inactivation of gene expression in Neurospora crassa by transformation with homologous sequences. *Mol Microbiol 6*:3343-3353.

Schneider B, Morávek Z, Berman H. 2004. RNA conformational classes. *Nucleic Acids Res 32*:1666-1677.

Shapiro B, Yingling Y, Kasprzak W, Bindewald E. 2007. Bridging the gap in RNA structure prediction. *Curr Opin Struct Biol 17*:157-165.

St-Onge K, Thibault P, Hamel S, Major F. 2007. Modeling RNA tertiary structure motifs by graph-grammars. *Nucleic Acids Res 35*:1726-1736.

Sykes M, Levitt M. 2005. Describing RNA structure by libraries of clustered nucleotide doublets. *J Mol Biol 351*:26-38.

Szewczak A, Moore P. 1995. The sarcin/ricin loop, a modular RNA. *J Mol Biol 247*:81-98.

Theimer C, Finger L, Feigon J. 2003. YNMG tetraloop formation by a dyskeratosis congenita mutation in human telomerase RNA. *RNA 9*:1446-1455.

Tinoco IJ, Bustamante C. 1999. How RNA folds. *J Mol Biol 293*:271-281.

Torres-Larios A, Swinger K, Krasilnikov A, Pan T, Mondragón A. 2005. Crystal structure of the RNA component of bacterial ribonuclease P. *Nature 437*:584-587.

Ullmann, J. (1976) An Algorithm for Subgraph Isomorphism: *ACM Press* New York, NY, USA. pp 31-42.

Ussery D, Hallin P. 2004. Genome update: Length distributions of sequenced prokaryotic genomes. *Microbiology 150*:513-516.

Vila-Sanjurjo A, Ridgeway W, Seymaner V, Zhang W, Santoso S, Yu K, Cate J. 2003. X-ray crystal structures of the WT and a hyper-accurate ribosome from Escherichia coli. *Proc Natl Acad Sci U S A 100*:8682-8687.

Vismara P. 1997. Union of all the Minimum Cycle Bases of a Graph. Electronic Journal of Combinatorics.

Wadley L, Pyle A. 2004. The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. *Nucleic Acids Res 32*:6650-6659.

Westhof E, Massire C. 2004. Structural biology. Evolution of RNA architecture. *Science 306*:62-63.

Wimberly B, Brodersen D, Clemons WJ, Morgan-Warren R, Carter A, Vonrhein C, Hartsch T, Ramakrishnan V. 2000. Structure of the 30S ribosomal subunit. *Nature 407*:327-339.

Woese C, Winker S, Gutell R. 1990. Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc Natl Acad Sci U S A 87*:8467-8471.

Xu X, Ji Y, Stormo G. 2007. RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. *Bioinformatics 23*:1883-1891.

Zell R, Sidigi K, Bucci E, Stelzner A, Görlach M. 2002. Determinants of the recognition of enteroviral cloverleaf RNA by coxsackievirus B3 proteinase 3C. *RNA 8*:188-201.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res 31*:3406-3415.

Index des termes clés

Annexe 1 : Article – On structural motifs in viral RNA

Est joint en annexe un article du Dr. Emmanuelle Permal rapportant les résultats d'une étude sur la recherche de motifs dans l'ARN génomique de virus. J'ai contribué à ces travaux de deux façons. D'abord, le logiciel *MC-View* de visualisation de motifs est né suite aux besoins de cette étude, et j'ai participé à la génération des résultats présentés ici. J'ai aussi contribué à la rédaction du texte du présent article.

L'article sera soumis à *Nature Structural & Molecular Biology*

On structural motifs in viral RNA

Emmanuelle Permal[1], Louis-Philippe Lavoie[1] and François Major[1]

Institute for Research in Immunology and Cancer, Université de Montreal, PO Box 6128, Downtown station, Montreal, Québec H3C 3J7, Canada.
Correspondence should be addressed to F.M[redacted]

All viruses express themselves through an RNA stage, as genomic RNA or mRNA, and possess structured RNA elements that contain RNA motifs. Some of them are specific to a class of element, such as bulges, and some are not. Our study reports interesting features of RNA motifs in viral RNA and especially bulge motifs that are often involved in protein binding.

INTRODUCTION

Viruses are obligate intracellular parasites that infect organisms from all domain of life. They have different shapes, different type of genome (DNA or RNA), different sizes (from 20 to 400 nm) and different hosts. However, a virus always needs to highjack the cell machinery (either eukaryote or prokaryote) to express itself. To do so, it has to pass through a RNA stage where its genome will be expressed as a RNA molecule; in the case of DNA viruses it will happen after transcription. When this stage is reached, the virus is a RNA molecule that is able to fold into a tertiary structure that can be analyzed and can give some useful information of therapeutic interest. Despite many studies on viral RNA elements such as the Trans Activation Response RNA element (TAR) of Human Immunodeficiency Virus 1(Aboul-ela et al., 1995), the stem-loop D (SLD) of Internal Ribosome Entry Site (IRES) of several *Picornaviridae* (Ohlenschlager et al., 2004; Headey et al., 2007), or the entire IRES (Martinez-Salas & Fernandez-Miragall, 2004), no work has been done to identify all of their structural features at large.

RNA tertiary structure possesses motifs that can de described as RNA graphs where nodes are nucleotides and edges are relations between them: bases pairing, bases stacking and bases adjacency (Major, 2007). All types of pairings can be considered in three-dimensional structure motif description: canonical (Watson-Crick and Wobble), non-canonical and base-backbone links. The last type of pairing happens when the oxygen atoms O1P, O2P or O2' from phosphate group and ribose, respectively, of a specific nucleotide makes a hydrogen bond with a partner nucleotide. RNA motifs are widespread in tertiary structure of RNA molecules and constitute structural building blocks (Hendrix et al., 2005) essential to the folding.

MATERIAL AND METHOD

To study structural features of viral RNA tertiary structures, we have developed a new tool, called *MC-View* (Lavoie et al – Unpublished), which takes as input a PDB file and RNA motifs described as RNA graphs and outputs all motifs found in the structure as a PyMOL (Delano, 2002) script and an annotated secondary structural graph representation. The color attributed to each RNA motif allows seeing motifs in their context in the RNA molecule, their mobility (in files that contain structure more than one solved model), and protein-interactions (when files contain protein structures bound to RNA). It has been adapted from the *MC-Search* (Hoffmann et al., 2003; Olivier et al., 2005) program that searches for one specific motif into several PDB files. To explore all viruses elements, since there is no complete virus 3D structure yet (except the Bluetongue virus dsRNA; PDB file 1H1K), we built a structural database of all the viral RNA 3D fragments available in the Protein DataBank (PDB last accessed on February 13th 2007)(Berman et al., 2000). It is a subset of the PDB of nearly 80 viral RNA elements such as pseudoknots, TAR hairpin, IRES hairpins, and much more **(Supplementary Table 1)**. To search for RNA motifs in this database we collected structural information on well-known motifs from SCOR database (Klosterman et al., 2002), and work from our lab on structural motifs **(Supplementary Table 2)**. Using *MC-View* with our two databases, we have been able to analyze all PDB files from our subset containing a selection of 17 RNA motifs in a visual way (the PyMOL sessions are available at the following address: http://www.bioinfo.iric.ca/MotifViralRNA/); the

coloration of a molecule seen in PyMOL gives two pieces of information: the localization of RNA motifs from the motif database (coloured regions) and the localization of potentially new RNA motifs (uncoloured regions) in the RNA molecule **(Fig. 1).** Using a combination of two methods: annotation and motif research, we analyzed the results in all the uncoloured regions of RNA elements submitted to *MC-View* **(Supplementary Methods)**.

RESULTS

The Bulge Motif

We identified new RNA motifs specific to their RNA element. The bulge RNA motifs, which did not belong to our recurrent RNA motif dataset, were often in uncolored regions. A bulge motif is defined here as an internal loop containing a number of nucleotides, from zero to n, between two canonical base pairs (Watson-crick or Wobble) that do not fold into a perfect A-form helix with canonical interactions; the number of nucleotides can be odd or even and usually creates a bend in the RNA molecule. The results, in Table 1, show that many bulges are specific to a RNA element. We observe that the bulges with only one nucleotide are common since they can be found in ribosomal RNA and elsewhere (See PDB 1ETF, 1FQZ and 2XIY in Table 1). On the contrary, when the number of nucleotide is higher some bulges are unique; this is the case for seven of eighteen bulges searched with *MC-Search* in PDB (See PDB 1ARJ, 1BIV, 1AJU, 1ETF, 1KP7, 1N66, 1RFR, 1XJR in Table 1). Bulges often participate in protein binding and thus adopt essential geometries for recognition; however our description of all of these characterized motifs only contains the nucleotide sequence information.

RNA-protein motifs

MC-View allows us to search in PDB database for the prot-RNA motif that consists in a pairing between an amino acid and a nucleotide. This motif was found in several RNA elements, but two were overlapping bulge motifs supporting the role of bulges as binding motifs for proteins and one was within a GNRA tetraloop. Since the

prot-RNA motif only defines a base pair, we gave a closer look to the pattern that we would describe and search in the PDB; these descriptions are shown in Table 2 (See 1BIV, 1ETF for bulges that interact with protein and 1A1T for the GNRA loop). Those three motifs are specific to their RNA element meaning that we did not find any occurrence in any RNA molecules with the relations that we described. This result shows that the bulges and their pairing to amino acids can define new types of complex motifs. In our study, they were unique but it would not be surprising to find some repetitive amino acid-RNA motifs. **Figure 2** shows an example of the motif formed by nucleotides G9, G11 and G14 respectively paired to Arg77, Arg73 and Arg70 in the Bovine Immunodeficiency TAR-TAT peptide complex (in dark red) and the bulge motif that contains G9 and G11 (in cyan). Described as a GNGNNG sequence bound to three independent arginines, this specific arrangement was only retrieved in BIV TAR-TAT PDB file confirming the importance of bulges in ligand recognition.

New interesting motifs

In the Hepatitis B virus epsilon stem-loop (PDB 2IXY), we noticed a singular combination of a triloop with a single cytosine nucleotide bulge called pseudo-triloop (see Table 2) (Flodell et al., 2006). Interestingly, described in *MC-Search* with two canonical base pairs it is only common to a well known RNA element: the Iron Responsive Element (IRE) that regulates the iron metabolism through binding to Iron Regulatory Protein (IRP). Despite the fact that the two triloop sequences are different, the recognition process of epsilon stem-loop by the viral reverse transcriptase and IRE by IRP may be similar and highlights new research directions (**Figure 3**).

In all the files analyzed we found many interesting features; one of them was in the loop of one model 14 of the NMR structural file of the C4 promoter of Influenza A virus (PDB ID 1MFY). We observed some structural variations in helical parts of the promoter with a strong conservation of the YNMG tetraloop motif; but the major change was a kink-turn motif annotated by *MC-View* in the loop. We hypothesize that it might be a type of K-loop motif defined by Nolivos and colleagues as a kink-turn within a loop (Nolivos et al., 2005). In *Haloarcula marismortui*, kink-turns from rRNA interact with nine of the 31 proteins of the large ribosome subunit (Klein et al., 2001); the formation

of this K-loop in a transitory stage could therefore serve as a nucleation site, like kink-turns in ribosome, and help the promoter in binding the polymerase (**Figure 4**).

The diloop motif

The diloop motif that we defined as a loop with two consecutive nucleotides closed by any type of base pair, canonical or not and including backbone-base interactions, was often present in the viral RNA element. There is no study specially reporting on the diloop motif, but as we analyzed our results it was found in almost all structures and sometimes more than once in the same PDB file, pointing out that it is an omnipresent motif. Using *MC-Search*, we found approximately 8000 occurrences of diloop in the PDB, including all NMR models and all database redundancy. In the 16s rRNA of *Haloarcula marismortui* (PDB 1S72), 58 diloops were observed. The diloop motif is, thus, a widespread motif that would need further investigation.

CONCLUSION

In summary, we determined that the bulges found in TAR from HIV-1, HIV-2 and BIV, RRE from HIV-1, stem-loop from domain IIID in HCV IRES, Y domain from poliovirus, SLD from Coxsackievirus B3 were unique in the Protein DataBank confirming the important role of bulge motif in ligand recognition and suggesting that a specific study on this motif would be interesting. We also found a potential K-loop that might also help in the protein binding process and confirm the strong resemblance between the pseudo-triloop of different essential RNA elements.

This study is the first one to address the annotation of RNA structures by their three-dimensional motif. Although it is limited to a specific type of structure, viral RNAs, it could be made on a larger dataset of RNA structures and lead to amazing discoveries on the use and the repartition of motifs in RNAs.

Note: Supplementary information is available on the Nature Structural & Molecular Biology website.

ACKNOWLEDGMENTS

We thank Véronique Lisi and Caroline Louis-Jeune for sharing their expertise on Kink-Turn and C-motif and Romain Rivière for his contribution with pdb2pdf. This work was supported by a grant from the Canadian Institute of Health Research (CIHR) (MT-14604) to FM. FM is a CIHR investigator, a member of the Institute for Research in Immunology and Cancer and a member of the Centre Robert-Cedergren of the Université de Montréal. LPL holds a CIHR scholarship to support higher education in bioinformatics (Université de Montréal, biT program).

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Figure 1 A conserved structure of HCV IRES (1KP7) annotated with *MC-View*. Colored in Green, a tetraloop; in grey, helical Watson-crick base pair; in orange, the regions of the structure that do not match any motif description from the *MC-View* RNA motif database.

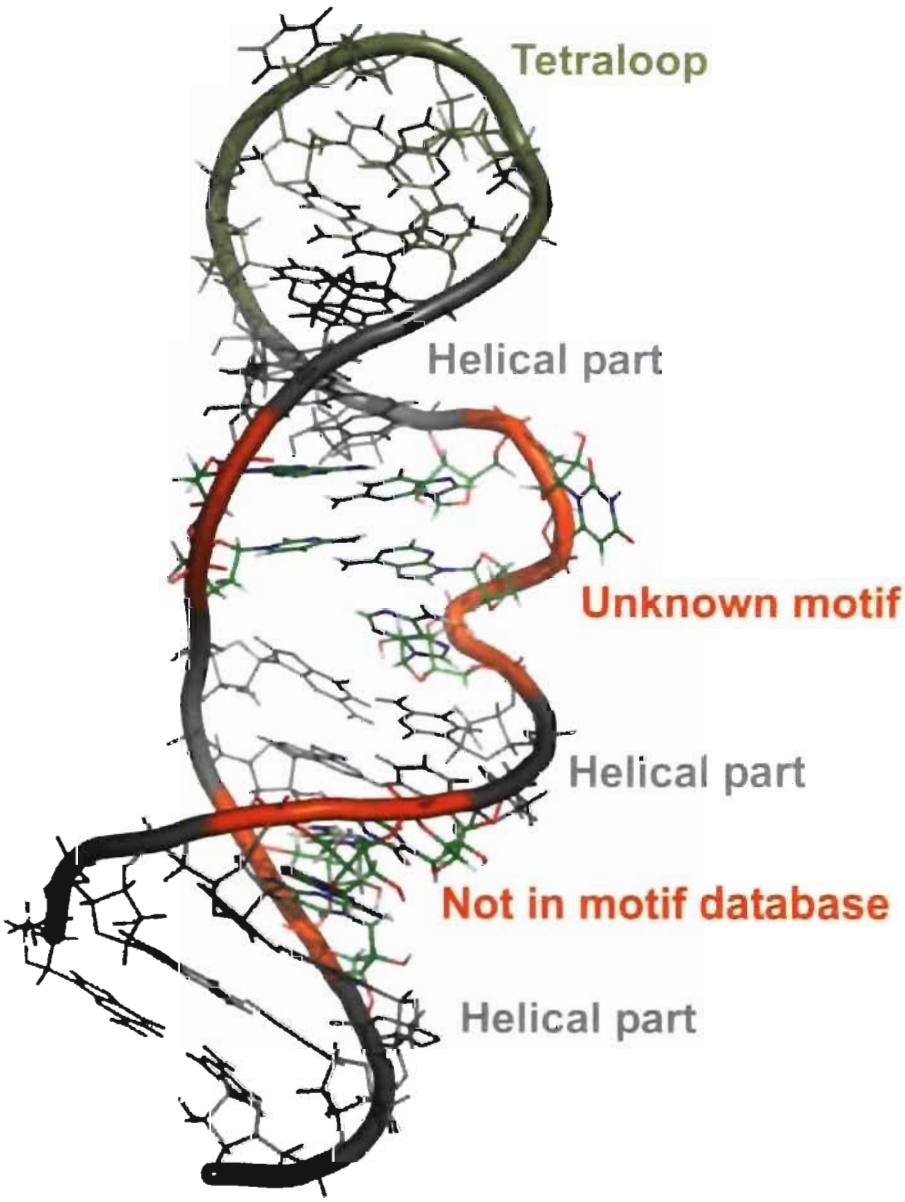


Figure 2 The Bovine Immunodeficiency Virus Trans-Activation Response RNA element (BIV TAR) in complex with TAT protein annotated with *MC-View*. The close-up shows the bulged part of BIV TAR that contains two motifs searched with *MC-Search*; the double-bulge motif and the triple G-Arg motif. In Green, a tetraloop; in grey, helical Watson-crick basepair; in cyan, the double-bulge motif; in red, the triple G-Arg motif.

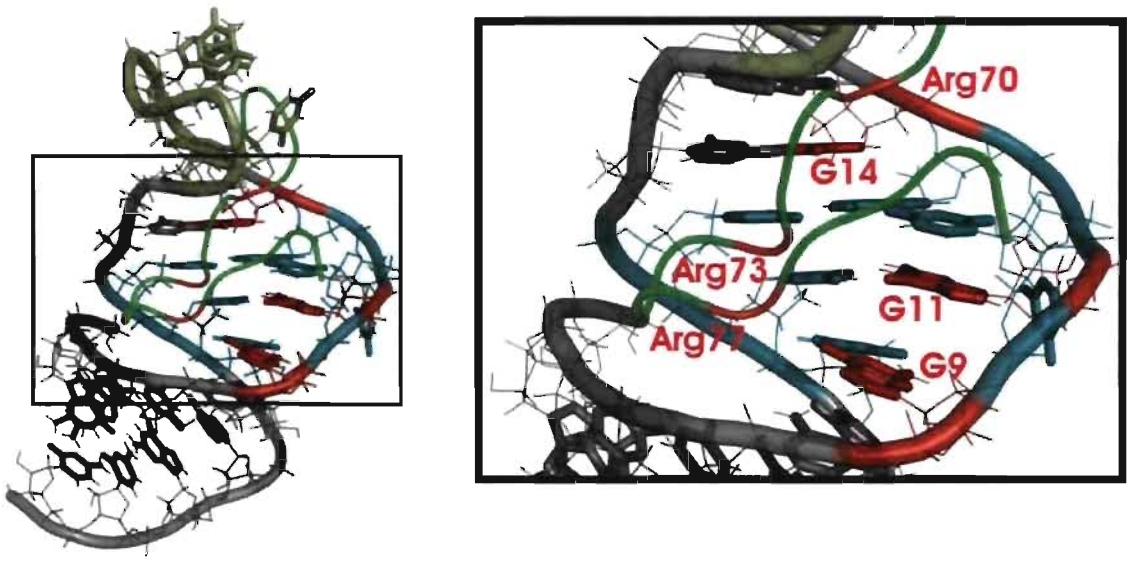


Figure 3 Hepatitis B virus pseudo-triloop (1MFY) In Blue the triloop; in cyan and blue the bulged out cytosine, in Grey helical Watson-crick basepair; in cyan, a wobble base pair

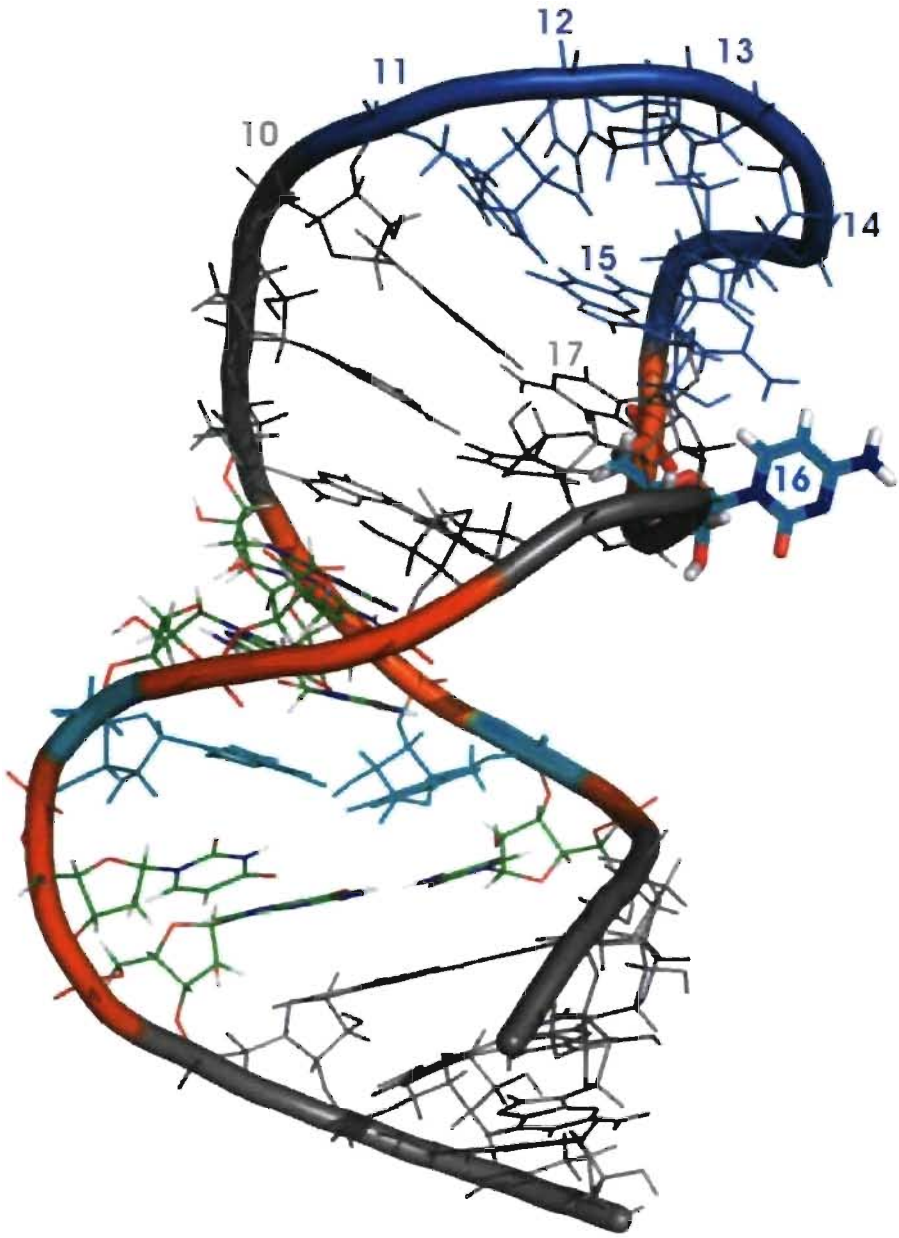


Figure 4 The K-loop motif in 2IXY. Two models of the C4 promoter influenza A are superimposed. In Green, a tetraloop; in grey, helical Watson-crick basepair; in cyan, a wobble base pair; in pink, the k-loop motif.

Table 1 Viral bulges frequencies. Lower case letters stand for nucleotides involved in Watson-crick or wobble base pairs. Specific means that there is no other occurrence of the motif in the PDB. Non-Specific motifs were found in other RNA molecules. H.M. is for Haloarcula marismortui; E.C. for Escherichia coli; and SECIS for selenocysteine insertion sequence.

RNA element	PDB	Bulge sequence	Motif searched with *MC-Search*	Observation
HIV-1 TAR RNA	1ARJ	5' aUCUg 3' 5' cu 3'	5' nUCUn 3' 5' nn 3'	Specific to HIV-1 TAR RNA
HIV-2 TAR RNA	1AJU	5' aUUg 3' 5' cu 3'	5' nUUn 3' 5' nn 3'	Specific to HIV-2 TAR RNA (but found in synthetic aptamers)
BIV TAR RNA	1BIV	5' aUgUg 3' 5' ccu 3'	5' nUnUn 3' 5' nnn 3'	Specific to BIV TAR RNA
HIV-1 RRE	1ETF	5' gc 3' 5' gAc 3'	5' nn 3' 5'nAn 3'	Non-specific, in 16S rRNA binding site for s8, in SARS s2m
HIV-1 RRE	1ETF	5' gGGc 3' 5' gGUAc 3'	5' nGGn 3' 5' nGUAn 3'	Specific to HIV-1 RRE
Domain IIID of HCV IRES	1FQZ	5' gu 3' 5' gUc 3'	5' nn 3' 5' nUn 3'	Non-specific, five occurrences in *H.M.* ribosomal RNA (but different closing base-pair).
Pseudo 5'-splice site RSV	1S2F	5' gUg 3' 5' cc 3'		
HBV Stem-loop	2IXY	5' cu 3' 5' gUg 3'		
Influenza A virus promoter	1JO7	5' cAAg 3' 5' cUg 3'	5' nAAn 3' 5' nUn 3'	Non-specific, 1 occurrence in *E.C.* ribosomal RNA (but different closing base-pair).
Domain IIIB of HCV IRES	1KP7	5'gCu 3' 5' aCc 3'	5' nCn 3' 5' nCn 3'	Non-specific, 2 occurrences in *E.C* ribosomal RNA (but different closing base-pair).
Domain IIIB of HCV IRES	1KP7	5' cAAUGc 3' 5' gACg 3'	5' nAAUGn 3' 5' nACn 3'	Specific to Domain IIIB of HCV IRES

RNA element	PDB	Bulge sequence	Motif searched with *MC-Search*	Observation
Influenza A complementary viral promoter	1M82	5' cAg 3' 5' cUUg 3'	5' nAn 3' 5' nUUn 3'	Non-specific, in ribozyme and SECIS mRNA hairpin
Y domain of poliovirus (SYNTH)	1N66	5' cCUc 3' 5' gUUg 3'	5' nCUn 3' 5' nUUn 3'	Specific to poliovirus Y domain
SLD of Coxsackievirus B3	1RFR	5' nUCUn 3' 5' nUUUn 3'	5' nYYYn 3' 5' nYYYn 3'	Specific to SLD
SARS CoV s2m	1XJR	5' cAc 3' 5' gAGg 3'	5' nAn 3' 5' nAGn 3'	Non-specific, in synthetic HIV-1 DIS(MAL) genomic RNA
SARS CoV s2m	1XJR	5' cCGAg 3' 5' cAg 3'	5' nCGAn 3' 5' nAn 3'	Specific to SARS Cov s2m
HBV Stem-loop	2IXY	5' gc 3' 5' gCg 3'	5' nn 3' 5' nCn 3'	Non-specific, two occurrences in *H.M.* ribosomal RNA and one in IRE (but different closing base-pair).

Table 2 Unique viral motifs. Using *MC-Search* we searched for interesting motifs that were not bulges observed in our dataset. Some were unique in the PDB; they are recorded in this table.

RNA element	PDB	Motif observed	Motif searched with *MC-Search*	Observation
HIV-1 SL3 STEM-LOOP RNA	1A1T	5' cGGAGg 3' Pairing: A4-Arg	5' nNNANn 3' Pairing: A-Arg	No other occurrence of this motif.
BIV TAR RNA	1BIV	5' GUGUAG 3' Pairing: G1-Arg1, G3-Arg2, G6-Arg3	5' GNGNNG 3' Pairing: G1-Arg, G3-Arg, G6-Arg	Unique in PDB
HIV-1 RRE	1ETF	5' UGGG 3' Pairing: U1-Arg1, G2-Arg1, G3-Gln1, G4-Gln1	5' NNNN 3' Pairing: N1-Arg1, N2-Arg1, N3-Gln1, N4-Gln1	Unique to HIV-1 RRE
HBV Stem-loop	2IXY	5'gcUGUgCg 3'	5' nnNNNnCn 3'	In IRE with sequence acAGUgCu

Aboul-ela F, Karn J, Varani G. 1995. The structure of the human immunodeficiency virus type-1 TAR RNA reveals principles of RNA recognition by Tat protein. *J Mol Biol 253*:313-332.

Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H, Westbrook J. 2000. The Protein Data Bank and the challenge of structural genomics. *Nat Struct Biol 7 Suppl*:957-959.

Delano WL. 2002. The PyMOL Molecular Graphics System. *DeLano Scientific Palo Alto*:USA.

Flodell S, Petersen M, Girard F, Zdunek J, Kidd-Ljunggren K, Schleucher J, Wijmenga S. 2006. Solution structure of the apical stem-loop of the human hepatitis B virus encapsidation signal. *Nucleic Acids Res 34*:4449-4457.

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

Headey SJ, Huang H, Claridge JK, Soares GA, Dutta K, Schwalbe M, Yang D, Pascal SM. 2007. NMR structure of stem-loop D from human rhinovirus-14. *Rna 13*:351-360.

Hendrix DK, Brenner SE, Holbrook SR. 2005. RNA structural motifs: building blocks of a modular biomolecule. *Q Rev Biophys 38*:221-243.

Hoffmann B, Mitchell GT, Gendron P, Major F, Andersen AA, Collins RA, Legault P. 2003. NMR structure of the active conformation of the Varkud satellite ribozyme cleavage site. *Proc Natl Acad Sci U S A 100*:7003-7008.

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

Klosterman PS, Tamura M, Holbrook SR, Brenner SE. 2002. SCOR: a Structural Classification of RNA database. *Nucleic Acids Res 30*:392-394.

Major F. 2007. *RNA tertiary structure prediction. In Lengauer,T. (ed.), Bioinformatics: From Genomes to Therapies Wiley-VCH, Weinheim, Germany, Vol. I.*

Martinez-Salas E, Fernandez-Miragall O. 2004. Picornavirus IRES: structure function relationship. *Curr Pharm Des 10*:3757-3767.

Nolivos S, Carpousis AJ, Clouet-d'Orval B. 2005. The K-loop, a general feature of the Pyrococcus C/D guide RNAs, is an RNA structural motif related to the K-turn. *Nucleic Acids Res 33*:6507-6514.

Ohlenschlager O, Wohnert J, Bucci E, Seitz S, Hafner S, Ramachandran R, Zell R, Gorlach M. 2004. The structure of the stemloop D subdomain of coxsackievirus B3 cloverleaf RNA and its interaction with the proteinase 3C. *Structure 12*:237-248.

Olivier C, Poirier G, Gendron P, Boisgontier A, Major F, Chartrand P. 2005. Identification of a conserved RNA motif essential for She2p recognition and mRNA localization to the yeast bud. *Mol Cell Biol 25*:4752-4766.

Ullman JR. 1976. An algorithm for subgraph isomorphism. *J Assoc Comput Mach 23*:31-42.

Supplementary information

Supplementary Table 1

Content of the database that we used for our study of RNA motifs in viral RNA. Grey, files that do not contain any motif from supplementary table 2 and that were not suitable to further analysis. Italic, files that contain pseudoknots.

PDB ID	RNA MOLECULE DESCRIPTION
1A1T	SL3 PSI-RNA HIV-1
1A34	SATELLITE TOBACOMOSAIC VIRUS
1A60	TYMV PKNOT
1ANR	HIV-1 TAR RNA
1ARJ	HIV-1 TAR RNA ARG BOUND
1BIV	BIV TAR-TAT
1BMV	ICOSAHEDRAL VIRUS
1CGM	CUCUMBER GREEN MOTTLE MOSAIC VIRUS
1CWP	COWPEA CHLOROTIC MOTTLE VIRUS
1CX0	HDV RIBOZYME
1DDL	DESMODIUM YELLOW MOTTLE TYMOVIRUS
1DRZ	HDV\U1A SPLICEOSOMAL PROTEIN
1ESH	BROME MOSAIC VIRUS (+) STRAND RNA
1ETF	HIV-1 RRE
1F5U	KISSING DIMER MOLONEY MURINE LEUKEMIA VIRUS
1F8V	OF PARIACOTO VIRUS
1H1K	BLUETONGUE VIRUS
1H2C	EBOLA VIRUS
1H2D	EBOLA VIRUS
1HVU	*HIV-1 PKNOT*
1I4B	BROME MOSAIC VIRUS (+) STRAND RNA
1I4C	BROME MOSAIC VIRUS (+) STRAND RNA

1I46	BROME MOSAIC VIRUS (+) STRAND RNA
1IK1	HRV-14 SLD
1J07	INFLUENZA A
1JZC	BROME MOSAIC VIRUS GENOMIC (+)-RNA
1KAJ	*MOUSE MAMMARY TUMOR VIRUS*
1KNZ	ROTAVIRUS MRNA 3' CONSENSUS
1KP7	HCV IRES EIF3 BINDING SITE
1KPD	*MOUSE MAMMARY TUMOR VIRUS*
1L2X	*VIRAL RNA PSEUDOKNOT*
1LAJ	TOMATO ASPERMY
1M82	COMPLEMENTARY RNA PROMOTER OF INFLUENZA A VIRUS
1MFY	INFLUENZA A VIRUS C4
1N1H	REOVIRUS
1N38	REOVIRUS
1N66	Y-DOMAIN OF POLIOVIRUS
1PGL	BEAN POD MOTTLE VIRUS
1R4H	IIIC DOMAIN OF GB VIRUS B
1RFR	STEMLOOP-D OF COXSACKIEVIRAL RNA
1RMV	RIBGRASS MOSAIC VIRUS
1RNK	*MOUSE MAMMARY TUMOR VIRUS*
1ROQ	COXSACKIEVIRUS B3
1S2F	ROUS SARCOMA VIRUS
1S34	ROUS SARCOMA VIRUS
1SJ3	HDV RIBOZYME
1SJ4	HDV RIBOZYME
1SJF	HDV RIBOZYME
1UON	REOVIRUS
1VBX	HEPATITIS DELTA VIRUS
1VBY	HEPATITIS DELTA VIRUS
1VBZ	HEPATITIS DELTA VIRUS
1VC0	HEPATITIS DELTA VIRUS
1VC5	HEPATITIS DELTA VIRUS
1VC6	HEPATITIS DELTA VIRUS
1VC7	HEPATITIS DELTA VIRUS

1VTM	TOBACCO MOSAIC VIRUS
1WNE	FOOT AND MOUTH DISEASE VIRUS
1XJR	SARS CoV s2m
1XOK	ALFALFA MOSAIC VIRUS
1YOQ	TWORT PHAGE
1Z30	BOVINE ENTEROVIRUS 1 SLD
2AGN	HCV IRES
2AZO	FLOCK HOUSE VIRUS
2AZ2	FLOCK HOUSE VIRUS
2BBV	BLACK BEETLE VIRUS
2C4Q	MS2-RNA HAIRPIN
2EVY	POLIOVIRUS 5'NTR CLOVERLEAF STEM LOOP D
2FZ2	TURNIP YELLOW MOSAIC VIRUS
2GIC	VESICULAR STOMATITIS VIRUS
2GTT	RABIES VIRUS NUCLEOPROTEIN-RNA
2HIX	ROUS SARCOMA VIRUS
2IXY	HEPATITIS B VIRUS
2IXZ	HEPATITIS B VIRUS
2NOQ	CRICKET PARALYSIS VIRUS IRES
437D	BEET WESTERN YELLOW VIRUS

Supplementary Table 2

Description of the RNA motifs searched in all our dataset described in Supplementary Table 1.

MOTIF	DEFINITION
PENTALOOP	Five nucleotide loop closed by a canonical or wobble base-pair
TETRALOOP	Four nucleotide loop closed by a canonical or wobble base-pair
TRILOOP	Three nucleotide loop closed by a canonical or wobble base-pair
DILOOP	Two nucleotide loop closed by any base-pair
WOBBLE	Wobble base-pair
SARCIN-RICIN	Sarcin-ricin motif
A-MINOR	Adenine nucleotide insert in the minor groove
CMOTIF-1	C-motif : description 1 – an internal loop motif
CMOTIF-2	C-motif : description 2- – an internal loop motif
BASE TRIPLE	Two consecutive nucleotides paired together into a dinucleotide platform with one of them paired with an other nucleotide
KINK-TURN1	Kink turn motif: description 1 – an internal loop motif
KINK-TURN2	Kink turn motif: description 2 – an internal loop motif
gnra	Tetraloop with gnra sequence
GNRA	Tetraloop with the gnra fold
YNMG	Tetraloop with ynmg sequence
U-TURN	Loop with a sharp bend in backbone
PROT-RNA	Pairing between an amino acid and a nucleotide
HELIX	Two consecutive watson-crick base-pairs

Supplementary Methods

We analyzed by visual inspection all output from *MC-View* with PyMOL. Two things maintained our attention: the uncolored regions and in NMR files with more than one model the RNA motif flexibility. We focused our analysis on uncolored region and annotated them with *MC-Annotate(Gendron et al., 2001)* to discover all interaction (base-pairing, base-stacking and base-adjacency) in these regions. We then search for the newly annotated region with *MC-Search(Hoffmann et al., 2003; Olivier et al., 2005)*, a tool that allows to search for RNA motifs, described as RNA Graph, in PDB format files by using the Ullman algorithm for graph isomorphism(Ullman, 1976). We looked for each newly described motifs into the whole the Protein DataBank (last accessed on march 2007 with 3464 files containing RNA). We then looked at the number of occurrences found with *MC-Search* and noticed if the motif described was specific (not found elsewhere) or non-specific (found in other type of RNA molecule) to the RNA element. Results are resumed in Table 1 and Table 2.