Université de Montréal

Approche plurielle à l'étude de la structure tertiaire de l'ARN chez les virus

Par

Emmanuelle Permal

Département d'Informatique et de Recherche Opérationnelle

Faculté des Arts et des Sciences

Thèse présentée à la Faculté des études supérieures

en vue de l'obtention du grade de Philosophiæ Doctor (Ph. D.)

en bio-informatique

Avril 2007

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée :

Approche plurielle à l'étude de la structure tertiaire de l'ARN chez les virus

Présentée par :

Emmanuelle Permal

a été évaluée par un jury composée des personnes suivantes :

Sylvie Hamel

Président-rapporteur

François Major

Directeur de recherche

Pascale Legault

Membre du jury

.....

Représentant du doyen

Daniel Gautheret

Examinateur externe

RÉSUMÉ

L'étude de la structure des macromolécules d'ARN constitue une étape indispensable à la compréhension de l'expression de l'information génétique et à l'exploration du rôle pluriel de l'ARN. Afin d'apporter des éléments de réponse sur les virus et leur habilité à détourner la machinerie cellulaire de leur organisme hôte, nous nous sommes intéressés à leur structure tridimensionnelle. L'étude de leur génome et de leur structure constituant un enjeu important dans l'identification de nouvelles cibles thérapeutiques et l'éradication des maladies qu'ils provoquent. Dans cette thèse, l'emphase a été faite sur les motifs ARN, qui sont des blocs de construction récurrents de la structure tridimensionnelle des molécules d'ARN et sur leur utilisation chez les ARN viraux. Ces motifs sont à la base de notre hypothèse qui est qu'ils confèrent la fonction aux molécules d'ARN. Les identifier et les localiser au sein des ARN viraux est, de ce fait, une information utile à la compréhension du mode de fonctionnement des virus.Trois approches ont été abordées durant ma thèse : la construction d'éléments viraux en utilisant la modélisation 3D d'ARN par homologie, la découverte du caractère ubiquitaire d'un motif ARN viral et enfin l'annotation de nombreux éléments viraux avec leur composition en motif ARN. Le nombre de structures tridimensionnelles disponibles dans les bases de données a explosé depuis la résolution de la structure cristalline de la sous-unité du ribosome en 2000. Ceci a mis à notre disposition près de 3000 structures d'ARN que nous avons utilisés lors de notre recherche. Afin d'atteindre nos objectifs, des outils bioinformatiques permettant l'analyse de la structure de l'ARN (annotation, modélisation moléculaire, recherche de motifs, entre autres) ont été utilisés et de nombreux procédés ont été développés. Le premier aspect de ce travail concerne la propriété de conservation de la structure tridimensionnelle au sein des virus; afin de prédire la structure tertiaire des ARN, un pipeline de modélisation par homologie a été développé et a fourni quatre nouveaux modèles structuraux de la tige boucle D virale, un élément régulateur très conservé chez les *Picornaviridae*. Le deuxième aspect révèle l'organisation particulière d'un motif ARN; le motif quartet a été caractérisé et recherché dans la base de données protéiques PDB (Protein DataBank) amenant à la découverte qu'il s'agissait d'un motif structural ubiquitaire stabilisateur. Le troisième aspect concerne la redondance des motifs ARN; une étude globale caractérisant les

éléments viraux a été réalisée, soulignant les caractères uniques des motifs bulge et fournissant le premier catalogue d'éléments d'ARN viraux annotés avec leurs motifs intrinsèques. Lors de notre étude sur les éléments viraux, il est apparu qu'ils constituaient un excellent choix pour l'étude de la structure de l'ARN, car possédant de nombreux motifs d'intérêt. Nous avons aussi pu observer la redondance de certains motifs ainsi que le caractère unique d'autres. Ce sont ces derniers, représentant l'identité de chaque élément étudié, qui pourraient être utilisés dans la recherche de nouvelles cibles thérapeutiques.

Mots clés : ARN, modélisation, motif d'ARN, recherche de motif, structure, virus

ABSTRACT

The study of RNA macromolecule structure is essential to understand genetic expression and to explore the plural role of RNA. To give some clues on the virus's ability to recruit the cellular machinery of the host, we investigated the 3D structure of some elements of their genome. Their genomic and structural analysis is crucial for the identification of novel drug target and for the eradication of the corresponding disease. In this thesis, the focus is made on RNA motifs, that are recurrent building block of the RNA 3D structure, and their use in viral RNA. Our hypothesis of work is that RNA motifs confer the function to the RNA molecule. Therefore, their identification and localization is useful for the understanding of how viruses work. We present three approaches in this thesis: (i) viral element building using 3D RNA homology modeling; (ii) the characterization of an ubiquitous viral RNA motif; (iii) the annotation of numerous viral elements given their RNA motif composition. The number of available 3D structure has increased in databases since the ribosomal RNA subunit crystal structure was solved in 2000. We used more than 3000 RNA structures during our research. To achieve our goals, we developed many processi combined with bioinformatics tools allowing RNA structure analysis (annotation, molecular modeling, motif search...). The first aspect of this work concerns the property of tridimensional structure conservation between different viruses. To predict the tertiary structure of the RNA, a homology modeling pipeline was developed and gave four new 3D structural models of the viral stem-loop D, a highly conserved element wihin picornaviridae. The second aspect reveals the peculiar organization of an RNA motif: the GC quartet. It was characterized and searched in the PDB database (Protein DataBank) leading to its description as a ubiquitous stabilizing structural motif. The third aspect concerns the redundancy of the RNA motif. A broad study characterizing the viral elements was realized, highlighting the unique feature of bulge motifs and providing the first catalogue of viral RNA elements annotated with their intrinsic motifs. During this study, it appears clearly that viral elements constitute an excellent model choice to analyze the RNA structure as they possessed numerous interesting motifs. In addition, we could observe the redundancy of some motifs and, in the contrary, the specificity of others.

Those last ones, that represent the identity of each element studied, could be used in the design of new drugs as targets.

Key words: Bioinformatics, Modeling, motif detection, structure, virus, RNA, RNA motifs

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES 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 Hoogsteen d'une base

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

O1P, O2P : Atome d'oxygène sur le groupement phosphate d'un nucléotide

P : Purine

RMN : Résonance Magnétique Nucléaire

S : Face Sucre d'un nucléotide

U : Uridine

W : Face Watson-Crick d'un nucléotide

wc : Appariements de type Watson-Crick

Y : Pyrimidine

A ma grand-mère Evane pour la motivation, mon mari William pour le soutien et mon petit Jo pour les gros câlins

« Rien ne sert de courir ; il faut partir à point »

Jean de La Fontaine – Le lièvre et la tortue.

"To know that we know what we know, and to know that we do not know what we do not know, that is true knowledge."

Copernicus

REMERCIEMENTS

Je remercie mon directeur François Major pour la confiance qu'il m'a accordée durant ma thèse. Les IRSC pour le soutien financier et surtout pour le congé maternité qui m'a permis de pouvoir poursuivre mon doctorat.

Un grand merci aussi à tous les membres, passés et présents, du Laboratoire de Biologie Informatique et Théorique pour les discussions fructueuses ou non, les éclats de rires et les bons moments. Biensur, il y a des chouchous parmi tous. Je pense particulièrement à Anita et Ratiba qui sont devenues des amies et aussi comme moi des mamans. Il y a aussi Chabane et sa femme qui m'ont aidé à leur manière quand mon petit Jo est arrivé; Karine S pour avoir été présente au bon moment; Karine C pour la danse du vendredi; Louis-Philippe pour son aide et son soutien dans la dernière ligne droite; Marc pour l'apprentissage des sacres; Martin pour avoir toujours répondu à mes questions; Philippe, qui est reparti voir son Estrie, pour plein de petits coups de pouce et pour son efficacité. Tamàs pour les moments de folies au lab.

Je tiens aussi à remercier Elaine Meunier pour beaucoup trop de choses, Nicolas Lartillot pour l'éveil à la réalité, Valentin Guignon pour les crêpes qui m'ont maintenue en forme, l'IRIC pour la superbe vue sur l'oratoire St-Joseph et mon superbe environnement de travail.

Et surtout, je tiens à remercier mon mari William et mon fils Jonah pour leur patience et pour tout simplement être là, à mes côtés.

AVANT-PROPOS

Les molécules d'ARN se retrouvent en général sous la forme de simples brins se repliant sur eux-mêmes en une structure tridimensionnelle. Cette dernière est maintenue grâce à la formation de ponts hydrogènes (ponts H) entre des paires de nucléotides, les empilements de bases et les interactions ioniques. Notre hypothèse de travail est que l'on retrouve dans les structures tridimensionnelles des ARN, des arrangements spatiaux de nucléotides appelés motifs qui leur confèrent leurs diverses fonctions.

Les motifs d'ARN sont les blocs de construction qui permettent le repliement des molécules d'ARN. Ils sont largement étudiés grâce à différents types d'approches biochimiques, biophysiques et bioinformatiques ainsi qu'aux différents niveaux d'organisation de la structure de l'ARN, primaire, secondaire comme tridimensionnelle. Dans cette thèse, ces motifs sont décrits et manipulés sous forme de graphe d'ARN, où les nœuds du graphe sont les nucléotides et les arêtes sont les interactions entre les nucléotides, grâce à des outils bioinformatiques spécialisés dans l'étude de la structure tridimensionnelle des molécules d'ARN.

Durant cette thèse, nous nous sommes plus particulièrement intéressés à la structure tridimensionnelle des ARN viraux en utilisant les motifs dont elle est composée. C'est afin de tenter de percer les secrets des virus et d'apporter des éléments de réponse sur d'éventuelles cibles thérapeutiques que nous avons identifiées et caractérisées de nouveaux motifs au sein de certaines structures 3D d'ARN viraux. Ceci nous a permis de développer une méthode de modélisation de la structure tridimensionnelle par homologie de motifs, de définir un nouveau motif d'ARN (le motif quartet) et d'annoter une base de données de molécules d'ARN viraux avec des motifs d'ARN récurrents.

CHAPITRE 1 INTRODUCTION

1.L'acide ribonucléique (ARN)

L'Acide Ribonucléique (ARN) est un polymère de nucléotides, au même titre que l'Acide Désoxyribonucléique (ADN), qui joue plusieurs rôles au sein des cellules et des virus. Il diffère de l'ADN par son sucre qui est un ribose contrairement à un désoxyribose et par la présence de l'uridine qui remplace la thymidine dans le jeu de nucléotides monomères (Adénosine, Cytidine, Guanosine et Uridine). Alors que l'ADN est essentiellement une molécule de stockage de l'information génétique, l'ARN est aussi une molécule transitoire qui permet, entre autres et sous ses différentes formes structurales, la régulation, le traitement et l'expression de cette information.

L'ARN est transcrit à partir de l'ADN grâce à l'ARN polymérase, un enzyme protéique, permettant la production d'ARN messager (ARNm), d'ARN de transfert (ARNt) et d'ARN ribosomal (ARNr) pour ne citer que ceux qui sont majoritairement exprimés et qui servent à la production des protéines dans une cellule. Ces ARN sont impliqués spécifiquement dans un mécanisme appelé traduction conduisant l'information codée par l'ADN vers un produit fonctionnel qui est la protéine (selon le dogme central de la Biologie Moléculaire : Information (ADN) → Expression (ARN) → Fonction (Protéine) formulé par Francis Crick en 1958) (voir figure 1).

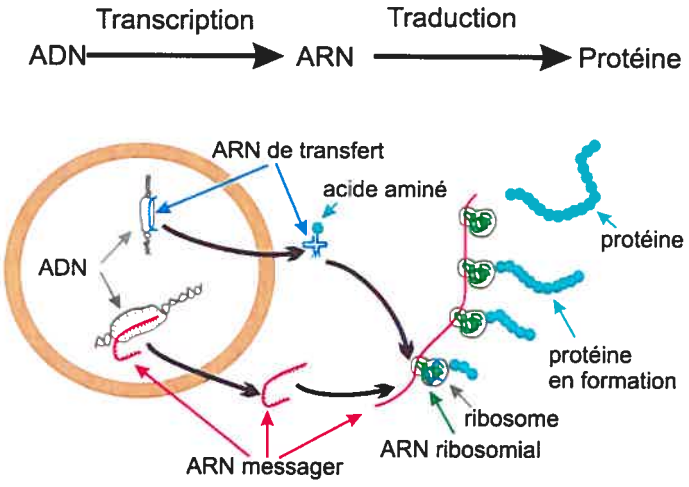


Figure 1 Dogme central de la biologie moléculaire

L'expression de l'information génétique se fait de l'ADN vers la protéine par l'intermédiaire de l'ARNm qui après avoir été transcrit de l'ADN, est traduit en protéine dans le cytoplasme par l'ARNt et les ribosomes. Sur ce schéma, l'ARNt est représenté en bleu, l'ARNm en rouge, l'ARNr en vert et gris et les acides aminés en bleu turquoise.

Les données actuelles démontrent que le monde des ARN est beaucoup plus complexe que le suppose le dogme de la biologie moléculaire. Dans la base de données RFAM (Griffiths-Jones et al., 2003; Griffiths-Jones et al., 2005), on compte, actuellement (février 2007), 574 familles d'Acide Ribonucléique (ARN) alors qu'il y a quelques années on n'en considérait que 3 : les ARN messager (ARNm), les ARN ribosomique (ARNr) et les ARN de transfert (ARNt).

L'ARN joue aussi le rôle de support d'information génétique chez certains virus, d'enzyme comme dans le cas de la sous-unité ARN de la ribonucléase P (RNase P) chez les bactéries (Guerrier-Takada et al., 1983) ou de régulateur de l'expression génétique chez les procaryotes (Gottesman, 2004) et les eucaryotes (Fire et al., 1998).

La découverte en 1993 des microARN (miARN) chez *Caenorhabitis elegans*, des petits ARN essentiels au développement et assurant la régulation de l'expression des ARNm par un mécanisme encore mal caractérisé (Lee et al., 1993), a amorcé une nouvelle ère dans l'étude de la molécule d'ARN. Ils sont depuis abondamment étudiés et ont permis de mettre en avant une nouvelle classe d'ARN qu'on appelle communément les ARN non-codant (ARNnc).

Les différentes familles d'ARN et leurs implications au sein de divers processus nécessaires au développement et à la vie d'un organisme a changé la vision et redéfinit le rôle des ARN (i.e. contrôle du développement par les miARN, de l'information par les ARN guides (ARNg (Blum et al., 1990)) via l'édition, de la toxicité par les ARN transfert-messager (ARNtm (Felden et al., 1996; Williams & Bartel, 1996)), etc.) (voir tableau 1). Alors qu'ils étaient considérés comme un support transitoire de l'information génétique dont la fonction était exprimée sous forme de protéines, les ARN sont

maintenant connus pour leur rôle fonctionnel intrinsèque. Ainsi le dogme central de la biologie moléculaire nouvellement formulé de la façon suivante :

Information (ADN et ARN) → Expression (ARN) → Fonction (ARN et Protéine) semble plus réaliste.

Tableau I Quelques classes d'ARN et leurs fonctions

ARN	*Rôle*	*Type*
ARN messager	Transcrit de l'ADN, il transporte l'information génomique.	codant
ARN de transfert	Élément essentiel à la traduction de l'ARNm en protéine, il transfère un acide aminé au polypeptide en formation.	Non-codant
ARN ribosomal	Composant ARN du ribosome, essentiel à la traduction de l'ARNm en protéine, il interagit avec l'ARNt et l'ARNm pour permettre la formation de la protéine. On le tient responsable de l'activité enzymatique du ribosome.	Non-codant
ARN transfert-messager	Recyclage des protéines coincées en cours de traduction dans le ribosome chez les bactéries.	Non-codant
ARN small nucleolar	Il guide des modifications chimiques de certains ARN	Non-codant
ARN small nuclear	Épissage des ARNm	Non-codant
MicroARN	Inhibition de l'expression d'un ARNm cible.	Non-codant
Sous-Unité ARN de la Ribonucléase P	Maturation de l'extrémité 5' de l'ARNt	Non-codant
ARN guide	Édition des molécules d'ARN.	Non-codant

A.Les ribozymes

Bien avant 1993 et la découverte des petits ARN non codant, les ARN avaient déjà montré leur potentiel enzymatique. Le dogme limitant le rôle de l'ARN à celui de messager avait déjà été ébranlé par la mise en évidence en 1982 de l'activité

enzymatique de l'intron du groupe I de *Tetrahymena thermophila* pour permettre son auto-épissage (Kruger et al., 1982) suivi en 1983 par celle du potentiel catalytique de la sous-unité ARN de la ribonucléase P chez la bactérie *Escherichia coli* (Guerrier-Takada et al., 1983). Ces deux découvertes avaient révélé l'aptitude des molécules d'ARN à avoir des activités enzymatiques qui étaient auparavant l'apanage des protéines. Elles ont donné naissance à un nouveau terme : celui de ribozyme qui définit un ARN ayant des propriétés enzymatiques. En 1989, Thomas R. Cech et Sidney Altman, partagèrent le prix Nobel de chimie pour leurs travaux sur les ribozymes. C'est donc au début des années 80 que l'engouement pour la recherche sur les molécules d'ARN a connu son premier souffle. Il a permis, entre autres, d'augmenter le nombre de famille d'ARN ayant un rôle catalytique, en particulier avec la découverte du ribozyme « Hammerhead » (Forster & Symons, 1987), puis d'amener à découvrir l'existence de plusieurs nouveaux ARN non-codants.

Figure 2 Ribonucléase P bactérienne

Structure de la Ribonucléase P en train de maturer l'extrémité 5' d'un ARNt. La sous-unité ARN de la RNAse P est en rouge et gris, la sous-unité protéique est en bleu et l'ARNt est en vert (Kazantsev et al., 2005). (Licence : http://creativecommons.org/licenses/by-sa/2.5/deed.fr_CA)

2.La structure de l'ARN

Afin d'accomplir son rôle dans différents mécanismes, l'ARN se sert de sa structure, dont l'unité fondamental est le nucléotide. Ce dernier est capable de se lier, de s'apparier et de s'empiler avec un autre nucléotide.

A.Le nucléotide

L'ARN est un polymère de plusieurs nucléotides reliés entre eux par des liaisons phosphodiester entre les groupements phosphates et les sucres (Figure 3). Les nucléotides portent les bases, ou bases azotées et sont divisés en deux classes : les purines (R) et les pyrimidines (Y). La première classe regroupe les nucléotides ayant deux cycles aromatiques, l'adénosine (A) et la guanosine (G) et la deuxième classe ceux n'en ayant qu'un, la cytidine (C) et l'uridine (U). Un nucléotide est pratiquement planaire; cette caractéristique est une des clés qui a permis l'élucidation de la structure de l'ADN par Watson et Crick en 1953 (Watson & Crick, 1953).

Par rapport à l'ADN, l'atome d'oxygène O2' du ribose donne un groupement nucléophile supplémentaire au nucléotide de l'ARN. De plus, le remplacement, dans l'uridine, du méthyle ($-CH_3$) présent dans la thymidine par un atome d'hydrogène diminue l'encombrement stérique. Ces deux caractéristiques donnent aux molécules d'ARN une plus grande réactivité avec le O2' nucléophile et une certaine souplesse structurale avec l'absence du groupement méthyle.

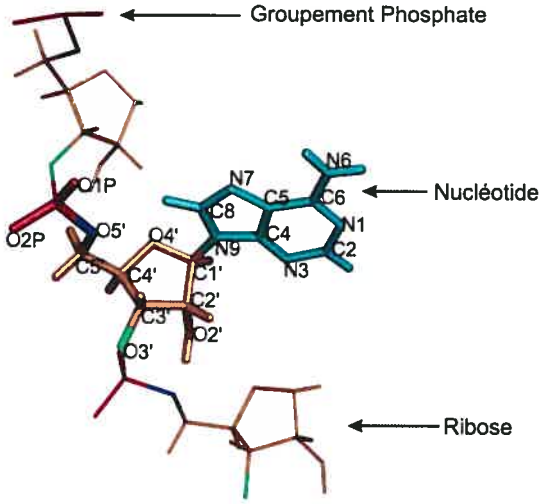


Figure 3 Structure de l'adénosine

L'unité fondamentale de l'ARN, le nucléotide, est composée d'une base azotée (en cyan), d'un ribose (en beige) et d'un groupement phosphate (en rouge). Les atomes O3' (en vert) et O5' (en bleu) du ribose assurent le maintien du squelette de l'ARN en se liant au groupement phosphate.

Les nucléotides peuvent être modifiés après l'étape de transcription, portant leur nombre de 4 à près d'une centaine observée (Limbach et al., 1994). Ils sont très nombreux dans les ARNt (environ 80) mais sont aussi présents dans les autres types d'ARN. Leur rôle est d'ajouter des fonctionnalités à l'ARN comme, par exemple, l'Inosine (dérivée de l'adénosine) ou la Pseudouridine (dérivée de l'uridine) qu'on peut trouver en première position de l'anticodon de l'ARNt lui permettant de s'apparier à plus d'un type de base (Figure 4)

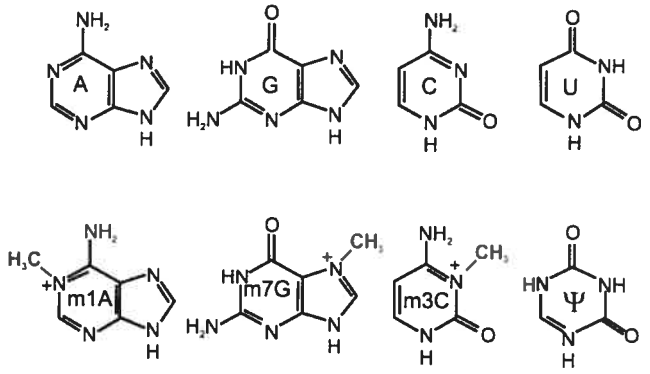


Figure 4 Les bases dans l'ARN

Les bases azotées Adénine, Guanine, Cytidine et Uracile (en haut). Quatre bases modifiées fréquentes dans les ARNt (en bas) où m1A est une méthyl-1 adénine, m7G est une méthyl-7 guanine, m3C est une méthyl-3 cytosine et ψ est une pseudo-uracil. Les modifications faites aux bases azotées sont indiquées en rouge.

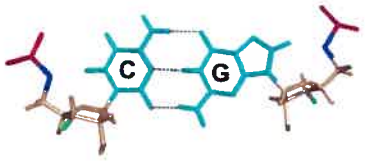
Dans la cellule, les molécules d'ARN ne sont pas linéaires, elles sont structurées. Les forces stabilisantes principales de cette structure sont l'appariement, l'empilement des bases et les interactions électrostatiques avec les ions.

B.Les appariements de nucléotides

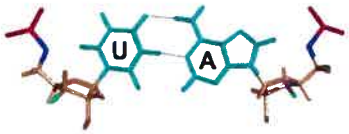
Une paire de base se forme quand deux nucléotides sont liés par un réseau de ponts hydrogène. Il existe des appariements de type canonique (Figure 5) et de type non-canonique. Le premier type comprend ceux dit « Watson-Crick », soit A-U et G-C et

celui dit « Wobble » soit G-U. Ils représentent la majorité des cas observés dans les molécules d'ARN (approximativement 63% - http://www-lbit.iro.umontreal.ca/mcannotate/base_pairs/). Les appariements dit non-canoniques concernent tous les autres appariements potentiels; par exemple une paire de bases U-U. Les appariements entre les nucléotides conduisent l'ARN vers sa structure tridimensionnelle.

A) Appariement CG Watson-Crick



B) Appariement UA Watson-Crick



C) Appariement UG Wobble

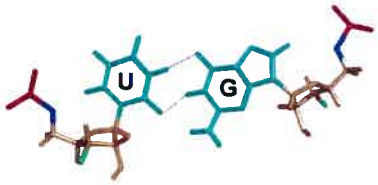
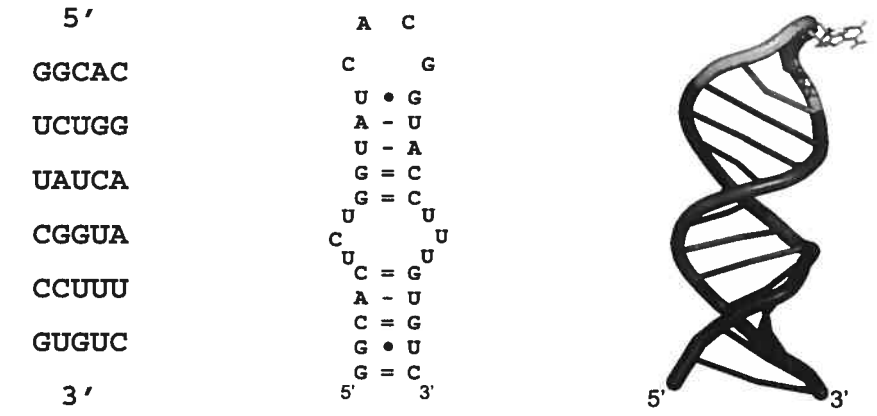


Figure 5 Appariements canoniques.

A) et B) Appariements de type Watson-Crick. C) Appariement de type Wobble. Les codes couleurs sont les mêmes que sur la Figure 3 : la base azotée est en cyan, le ribose est en beige et le groupement phosphate est en rouge. Les ponts hydrogènes sont représentés par des lignes pointillées noires.

3.Étudier la structure de l'ARN

L'étude de la structure des macromolécules d'ARN constitue une étape indispensable à la compréhension de l'expression de l'information génétique. Celle-ci peut se faire par l'analyse de sa structure en 3 niveaux d'organisation (Figure 6):



Structure primaire | Structure secondaire (Deux dimensions – 2D) | Structure tridimensionelle

Figure 6 Les trois niveaux d'organisation de la structure de l'ARN.

Représentation de la structure primaire (à gauche), secondaire (au centre) et tridimensionnelle (à droite) de la tige-boucle D (SLD) du virus de Coxsackie B3 (structure résolue par Ohlenschlager et al.(Ohlenschlager et al., 2004)).

A.La structure primaire :

La structure primaire est la séquence en nucléotides de l'ARN représentée par une combinaison des lettres A, U, C et G. On peut représenter des ensembles de nucléotides par d'autres lettres et selon la nomenclature de l'International Union of Pure and Applied Chemistry (IUPAC: http://www.iupac.org/index_to.html). Par exemple, l'ensemble (A, G) par la lettre R, (C, U) par la lettre Y, (A, C, G, U) par la lettre N.

B.La structure secondaire :

En bioinformatique, elle est définie comme étant composée par l'ensemble des appariements canoniques (Mount, 2004). Typiquement, la structure secondaire est composée d'un arrangement de tiges-boucles (Figure 6 et 7).

En bio-informatique, l'étude de la structure des molécules d'ARN se fait majoritairement au niveau de la structure secondaire. Ceci s'explique par le fait qu'il

existe beaucoup plus de données génomiques, 130 milliards de nucléotides dans la base de données *Entrez nucléotide*, que de données structurales, approximativement 3000 structures d'ARN dans la base de données Protein DataBank (PDB) en février 2007 (Berman et al., 2000b) et que la prédiction de le structure à partir de la séquence primaire est beaucoup plus simple au niveau secondaire que tertiaire. Un grand nombre d'outils ont donc été développés pour prédire la structure secondaire à partir de données génomiques. La structure secondaire d'un ARN peut être prédite selon des critères thermodynamiques (par exemple avec les logiciels MFOLD (Mathews et al., 1999; Zuker, 2003) remplacé en 2005 par UNAFOLD (Markham & Zuker, 2005) et ALIFOLD (Zuker, 2003)), recherchée selon leur topologie décrite par un utilisateur (par exemple avec les logiciels de recherche de motifs comme RNAMotif (Macke et al., 2001), RNAMOT (Laferriere et al., 1994) et ERPIN (Lambert et al., 2004)) ou bien même comparée avec une autre structure (par exemple avec le logiciel RNASTA (Guignon et al.). Cependant, toutes ces méthodes ne permettent pas une étude complète de l'ARN puisque toutes les interactions non canoniques et les empilements, qui contribuent considérablement à la stabilisation de la molécule, ne sont pas pris en compte. C'est pourquoi, il est aussi important d'enrichir l'information structurale par celle que peut procurer l'étude de la structure tridimensionnelle de l'ARN.

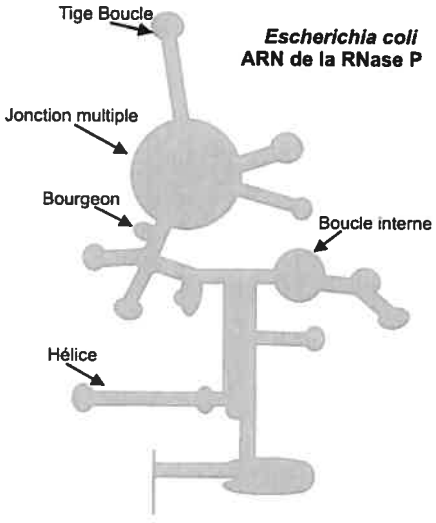


Figure 7 Structure secondaire de la sous-unité ARN de la Ribonucléase P *d'Escherichia coli.*

Les différents motifs retrouvés régulièrement dans les structures secondaires d'ARN sont indiqués par une flèche.

C.La structure tertiaire :

La structure tertiaire d'une molécule d'ARN correspond à sa structure tridimensionnelle définie par ses coordonnées atomiques(Moss, 1996). On fait ainsi référence à l'organisation des atomes en trois dimensions.

Les autres appariements et les interactions d'empilement s'ajoutent à ceux de la structure secondaire pour former la structure tertiaire. Dans la suite du manuscrit, nous y référons comme étant la structure tridimensionnelle.

4.Déterminer la structure tertiaire d'un ARN

C'est en 1968 que la cristallisation des molécules d'ARN a commencé et qu'il a enfin été possible de s'intéresser à leur structure tridimensionnelle de plus près (Kim & Rich, 1968). Plusieurs ARNt furent ensuite cristallisés (Kim et al., 1974; Basavappa & Sigler, 1991), puis le domaine P4-P6 de l'intron de groupe I (Cate et al., 1996), puis d'autres pour aboutir au cristal du ribosome (Ban et al., 2000; Yusupov et al., 2001) qui correspond à la plus longue molécule d'ARN cristallisée à date. Toutes ces structures ont apporté des informations précieuses sur la structure de l'ARN. L'avancement de la connaissance sur la structure tridimensionnelle de l'ARN est tributaire de celle des méthodes pour la déterminer.

A.Les méthodes biophysiques

Les méthodes biophysiques les plus utilisées afin d'obtenir des structures tridimensionnelles d'ARN sont la cristallographie aux rayons X et la spectroscopie de résonance magnétique nucléaire (RMN). Il n'est plus question ici de séquence (comme pour les deux premiers niveaux d'organisation d'un ARN) mais de coordonnées atomiques en trois dimensions. Les structures sont alors représentées sous forme de fichiers, au format PDB, contenant les coordonnées tridimensionnelles de tous les atomes constituant la ou les molécules d'ARN. Les structures sont compilées et disponibles dans la Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB - www.rcsb.org/) (Berman et al., 2000a; Berman et al., 2000b; Berman et al., 2007). La spectrométrie de masse (Thomas & Akoulitchev, 2006) et la

cryomicroscopie électronique (Mueller et al., 2000; Schuler et al., 2006) sont deux méthodes plus récentes qui permettent aussi la détermination de la structure tridimensionnelle d'ARN.

Ce n'est que récemment que les méthodes biophysiques ont permis de produire une grande quantité de structures pour les ARN (3534 fichiers contenant des ARN en juin 2007 – Nombre de fichiers collectés à partir de la base de données PDB au laboratoire depuis janvier 1994). La différence fondamentale entre ces deux méthodes est l'état de départ de la molécule dont on veut connaître la structure : solide sous forme de cristal pour l'une et en phase liquide pour l'autre. Elles apparaissent donc complémentaires pour étudier tous les aspects structuraux des molécules d'ARN.

Cristallographie à rayons X

La cristallographie aux rayons X est la méthode biophysique la plus utilisée pour la détermination de structures de macromolécules biologiques. Elle a permis de résoudre plus de 2344 structures sur 3534 recueillis à partir de la PDB (Berman et al., 2000b), à la date du 1er juin 2007. C'est une méthode physique permettant de déterminer des structures protéiques et d'acides nucléiques. Le principe est de faire croître un cristal ordonné de la molécule d'intérêt, puis de la soumettre à des rayons X afin d'obtenir un patron de diffraction. Ce patron est ensuite analysé pour créer une carte de densité des électrons qui servira à déterminer les coordonnées atomiques (le modèle) de la structure tridimensionnelle de la molécule d'intérêt.

Spectroscopie de résonance magnétique nucléaire

La spectroscopie de résonance magnétique nucléaire (RMN) est la deuxième méthode la plus utilisée pour résoudre la structure d'un ARN (898 structures déterminées sur les 3534 à la date du 1er juin 2007). Elle repose sur le principe que tous les noyaux de tous les atomes ayant un spin non nul possèdent un moment magnétique intrinsèque et un moment angulaire. Les molécules sont soumises à un champ magnétique statique puis exposés à un second champ magnétique oscillatoire. Certains noyaux d'atome de la molécule « résonnent » et d'autre non. La spectroscopie permet d'analyser les résultats du spectre de résonance des atomes et de calculer, grâce à

l'observation de contrainte d'angles, des distances et des torsions entre les atomes, la structure tridimensionnelle de la molécule.

B.Méthodes chimiques et enzymatiques

D'autres méthodes de détermination de la structure existent, notamment les méthodes chimiques et enzymatiques. Par exemple l'emploi d'enzymes, comme la Nucléase S1 (NS1) et la RNase V1 (RV1), localisent les régions simple brin (sites de clivage pour NS1) et double brins (sites de clivage pour RV1). Un autre exemple est celui de l'agent chimique diméthyl sulfate (DMS) qui peut être utilisé pour modifier les Guanines (G) et Cytosines (C) non impliquées dans des appariements de type Watson-Crick. Les positions modifiées peuvent être ensuite clivées par un traitement à l'aniline indiquant ainsi les GC Watson-Crick et les GC non Watson-Crick.

C.La modélisation théorique.

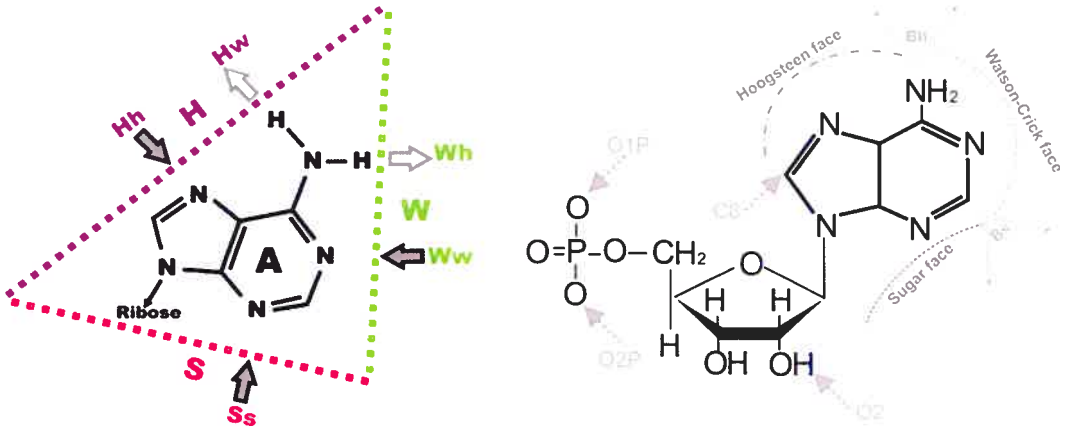
Lorsqu'une structure n'est pas disponible dans une base de données, il est possible de faire un modèle théorique de celle-ci grâce à un système de modélisation en lui soumettant les informations nécessaires à la construction d'une structure comme la séquence et ses interactions mesurées expérimentalement ou déduites de sa séquence. Les systèmes de modélisation des ARN reposent sur des principes différents. Par exemple, Manip permet de manipuler de façon interactive des structures d'ARN en utilisant un champ de force (Massire & Westhof, 1998). Yammp utilise des pseudo-atomes représentant ainsi la structure des nucléotide de façon restreinte (Robert & Harvey, 1993). NAB se base sur le même principe que *MC-Sym*, la construction par nucléotides rigides mais en se servant de la géométrie des base contrairement aux interactions entre nucléotides comme *MC-Sym* (Macke & Case, 1998). Celui présenté dans le chapitre 2 pour la modélisation par homologie est fondé sur la résolution du problème de satisfaction de contraintes où les contraintes sont les relations entre les nucléotides (Major et al., 1991).

Afin d'analyser les structures tridimensionnelles des ARN, il est important de bien les décrire; ceci peut être fait en les définissant dans un espace discret. C'est le but de l'annotation des structures d'ARN.

5.L'annotation des structures d'ARN.

La description des structures et surtout celle des appariements passe à un niveau supérieur de complexité quand il s'agit de structure tridimensionnelle. En effet, une nomenclature des faces de bases nucléotidiques (Figures 8 et 9) introduite par Leontis et Westhof, puis améliorée par Lemieux et Major, permet de nommer de manière non ambiguë tous les types d'appariement (Leontis & Westhof, 2001; Lemieux & Major, 2002). Ces deux nomenclatures sont à la base de cette thèse qui utilise extensivement l'annotation des molécules d'ARN pour définir, manipuler, caractériser et rechercher les motifs d'ARN.

Pour chaque base, trois faces contenant des donneurs et des accepteurs sont définies : la face Watson Crick (symbolisée par un W), celle où se retrouvent les donneurs et les accepteurs impliqués dans les appariements canoniques et décrits par les fameux chercheurs; la face du côté sucre (symbolisée par un S),; et finalement la face du côté opposé à celle du sucre, la face « Hoogsteen » (symbolisée par un H),, nommée ainsi en l'honneur du chercheur qui a observé pour la première fois un appariement impliquant celle-ci. Des lettres en minuscule (w, s, h) décrivent les zones présentes dans les différentes faces; par exemple A Wh symbolise la face Watson d'une Adénosine côté Hoogsteen. L'agencement des faces en contact dans l'appariement défini le type d'appariement. Par exemple, une paire Watson-Crick est définie ainsi selon la nomenclature de Lemieux et Major : G-C Ww/Ww, décrivant ainsi les faces impliquées dans l'appariement, soient les faces Ww du G et Ww du C.



A) B)

Figure 8 Nomenclature des faces d'interaction.

A) Représentation de l'Adénine (A) et de ses faces d'interaction avec la nomenclature de Leontis et Westhof. W = Watson, S= Sucre, H = Hoogsteen, et de Lemieux et Major à W = Watson –Crick (Ww,Wh,Ws), S= Sucre (Ss, Sw), H = Hoogsteen (Hh, Hw,Hs) , Bs = sucre bifurqué, Bh = Hoogsteen bifurqué. B) Représentation de l'Adénosine (A) avec son ribose et son groupement phosphate. Les atomes O1P, O2P du groupement Phosphate et O2' du ribose du squelette de l'ARN sont souvent impliqués dans la formation des ponts hydrogènes.

Les appariements possèdent aussi une orientation qui se définit par la position des riboses par rapport à l'axe formé par les deux nucléotides appariés ou par l'orientation des vecteurs normaux aux plans des bases.

Le vecteur normal au plan d'une base, $\vec{n}_Y$, est déterminé par une rotation avec la règle de la main droite de l'atome N1 vers N6 (Major, 2007). Un vecteur $\vec{\sigma}$ a été défini par Major et Thibault afin de connaître les côtés impliqués dans l'empilement, il représente le vecteur normal des bases; $\vec{\sigma} = \vec{n}_Y$ pour les pyrimidines et $\vec{\sigma} = -\vec{n}_Y$ pour les purines. La définition a été établie de façon à ce que tous les vecteurs normaux $\vec{\sigma}$ pointent dans la même direction dans une double-hélice de type A.

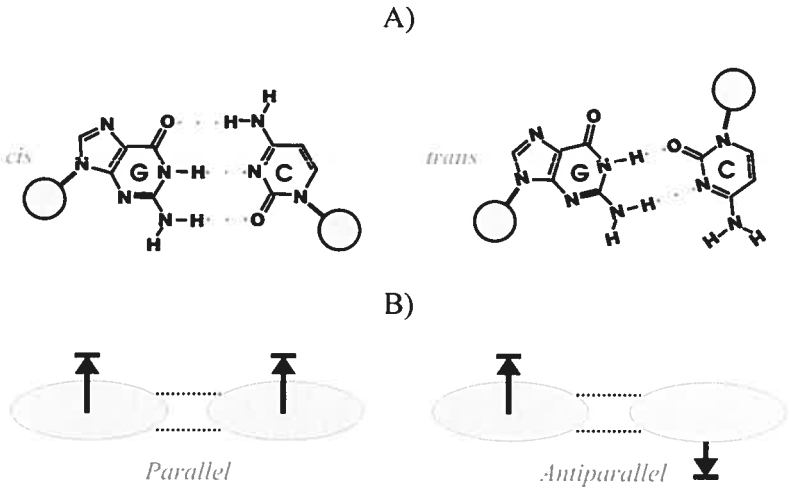


Figure 9 Orientation des paires de base : Exemple de deux paires GC Watson-Crick.

A) À gauche, une paire GC W/W cis et à droite, une paire GC W/W trans. Un appariement peut être cis si les deux riboses sont orientés du même côté ou trans si ils sont orientés vers des directions opposées. B) À gauche, un appariement de type parallèle et à droite un appariement de type antiparallèle. Les flèches représentent la direction des vecteurs $\vec{\sigma}$. Une paire de base sera parallèle si les vecteurs $\vec{\sigma}$ des deux bases pointent dans la même direction et antiparallèle autrement.

La notion d'empilement ou « stacking » est aussi introduite au niveau de la structure tridimensionnelle. C'est une caractéristique chimique qui est due à la présence de noyaux aromatiques dans les nucléotides. Deux bases adjacentes ou non sont capables de former des interactions d'emplilements via leurs cycles aromatiques (2 pour les purines et 1 pour les pyrimidines) et apparaissent quand on les observe comme si elles étaient empilées. Ce lien non covalent est très fort et aide grandement, avec les liaisons hydrogènes, à la stabilisation de la structure tridimensionnelle des acides nucléiques.

Il existe deux types d'empilement entre deux nucléotides : adjacent quand les deux bases empilées sont adjacentes dans la séquence de nucléotide ou non adjacent quand ce n'est pas le cas. Ils peuvent ensuite s'annoter en suivant la nomenclature de

Major et Thibault qui définit quatre classes d'empilement grâce au vecteur $\vec{\sigma}$ (Major, 2007). Les empilements « upward » et « downward » concernent deux nucléotides avec un $\vec{\sigma}$ qui pointe dans la même direction et diffèrent par la position des nucléotides impliqués (Figure 10 a) et b)). Un empilement « outward » est observé quand les $\vec{\sigma}$ pointent dans des directions opposées et un « inward » quand c'est l'un vers l'autre (Figure 10 c) et d)).

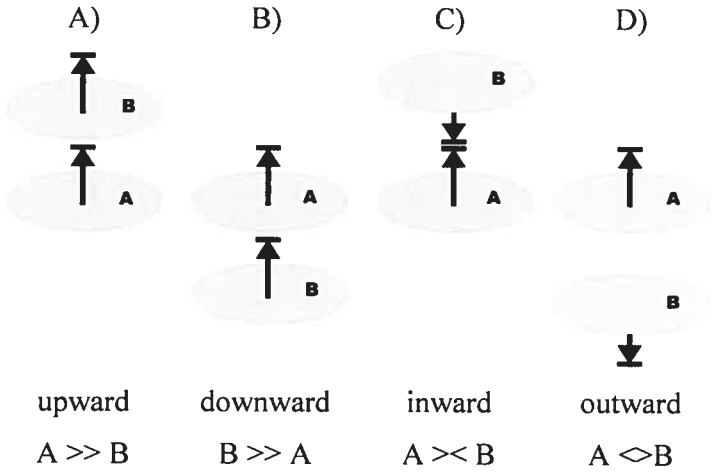


Figure 10 Les quatre types d'empilement des nucléotides.

De gauche à droite : a) « upward » b) « downward » c) « inward » d) « outward ». Les formes ovales représentent un nucléotide, les lettres A et B sont les identifiants des nucléotides. Les flèches indiquent la direction vers laquelle pointe le vecteur $\vec{\sigma}$.

L'énergie des liaisons hydrogènes, des empilements et des interactions électrostatiques contribuent à stabiliser une molécule d'ARN.

6.Les motifs

Il existe dans les molécules d'ARN des sous-ensembles de nucléotides qui adoptent des structures particulières et récurrentes : ce sont des motifs. Ils sont définis de différentes manières et en fonction de la nomenclature utilisée ou bien de l'interprétation de l'arrangement structural observé. Donc, il n'y a pas réellement de définition générale de ce qu'est un motif d'ARN mais plutôt une description propre à chaque motif d'ARN. Ces motifs sont à la base de notre hypothèse qui est qu'ils confèrent la fonction aux

molécules d'ARN et que les identifier et les localiser au sein des ARN viraux est une information utile à la compréhension du mode de fonctionnement des virus.

Un motif est un fragment d'ARN qui pourrait être comparé à un bloc de construction. Une structure d'ARN contient plusieurs blocs; certains sont récurrents et d'autres sont uniques (Hermann & Patel, 2000; Hendrix et al., 2005). Un exemple est donné à la Figure 11 avec l'ARN de la sous-unité 5S du ribosome de *Haloarcula Marismortui.* On y voit des motifs dans la structure de différentes couleurs montrant ainsi qu'une molécule d'ARN est divisible en plusieurs sous-unités. Le tableau II résume certains motifs d'ARN qui ont été utilisés lors de l'étude des motifs d'ARN structuraux dans le chapitre 3 et qui ont servi à annoter la Figure 11

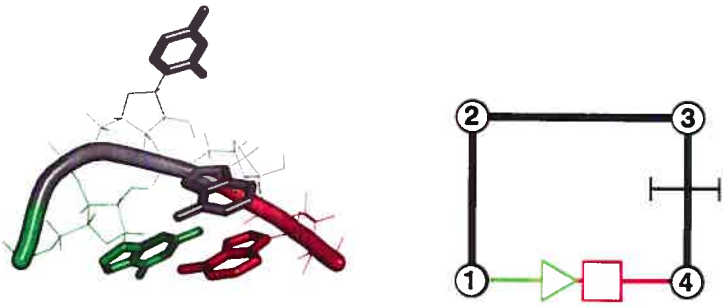
Tableau II Quelques motifs d'ARN

Motif	*Description*
Diloop	Boucle à deux nucléotides fermée par une paire de bases.
Triloop	Boucle à trois nucléotides fermée par une paire de bases canonique.
Tetraloop	Boucle à quatre nucléotides fermée par une paire de bases canonique.
Pentaloop	Boucle à cinq nucléotides fermée par une paire de bases canonique.
GNRA	Motif ayant une séquence GNRA et une paire S/H entre G et A.
YNMG tetraloop	Tetraloop ayant une séquence YNMG et fermée par une paire de bases.

Figure 11 Des motifs dans la sous unité ARN de la Ribonucléase P bactérienne

Image générée avec PyMol 0.99 et *MC-View* (Article mis en annexe). Liste des motifs : En vert, une tetraloop; en cyan, une paire Wobble, en beige, des diloop; en magenta, un bulge; en gris, les paires Watson-Crick. Les régions en bleu foncé ne contiennent aucun des motifs de la table 2.

Il existe beaucoup de motifs d'ARN bien caractérisés et une base de données leur est consacrée (Klosterman et al., 2002; Klosterman et al., 2004). Chaque année, de nouveaux membres de cette famille sont découverts et mis en avant pour leurs caractéristiques structurales ou bien leur fonction (Hendrix et al., 2005). Un exenple de motif très étudié est la tetraloop GNRA ; à la figure 12 est représenté une boucle GCAA et son graphe d'ARN correspondant à son annotation utilisant la nomenclature de Leontis et Westhof où un rond symbolise la face Watson (en cis M ou en trans F), un triangle la face Sucre (en cis ▶ ou en trans ▷), un carré la face Hoogsteen (en cis O ou en trans G) et où une barre perpendiculaire à une ligne symbolise un empilement de bases (Leontis & Westhof, 2001; Leontis et al., 2006). Une ligne épaisse représente une liaison entre deux nucléotides du graphe par le squelette de l'ARN. Chaque nucléotide représente un nœud du graphe et chaque interaction (empilement, appariement ou lien de squelette) une arrête du graphe (Lemieux & Major, 2006).



(A) (B)

Figure 12 Une tetraloop GNRA et son graphe d'ARN

Les nucléotides G et A formant l'appariement S/H sont respectivement en vert et en rouge. (A) Les nucléotides d'une tétraloop GCAA. (B) Son graphe d'ARN correspondant respectant le même code de couleur. Le nucléotide 1 en vert fait un appariement en trans via sa face sucre (▷) avec le nucléotide 4 en rouge qui y implique sa face hoogsteen (G).Les nucléotide 1 à 4 sont reliés par le squelette de l'ARN. Le nucléotide 3 est empilé sur le nucléotide 4.

A.Les cycles

En 2006, Lemieux et Major ont proposé une nouvelle approche à l'analyse des motifs d'ARN : les cycles (Lemieux & Major, 2006). La structure d'une molécule d'ARN peut être représentée sous forme de graphe, où les nœuds sont les nucléotides et les arêtes sont les relations entre les nucléotides où une relation est soit un empilement, soit un appariement, soit une adjacence. Un cycle est un sous-graphe indivisible de ce graphe d'ARN et est une représentation de motif d'ARN facilement manipulable du fait de la simplicité de sa description. La Figure 13, montre le base de cycle d'un fragment de la structure de la sous-unité ARN de la ribonucléase P bactérienne.

Le chapitre 2 présente un exemple d'utilisation des cycles dans la modélisation de structures théoriques d'ARN.

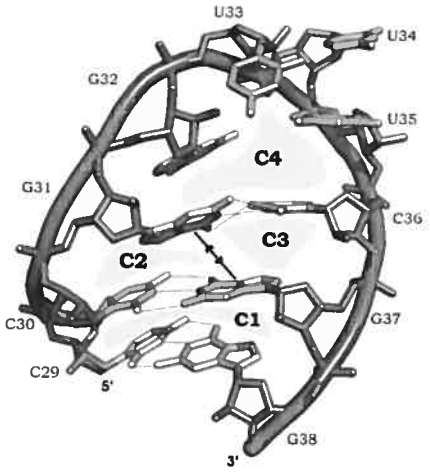


Figure 13 Des cycles dans un fragment de la sous-unité ARN de la Ribonucléase P bactérienne

Le fragment allant des nucléotides 29 à 38 est décomposable en 4 cycles. Le cycle C1 est formé par les bases C29, C30, G37 et G38; le cycle C2 par les bases C30, G31 et G37; le cycle C3 par les bases C36, G37 et G31; et le cycle C4 par les nucléotides G31, G32, U33, U34, U35 et C36.

Sur ce schéma, les ponts hydrogènes sont représentés sous forme de lignes pointillées; et l'empilement de type « outward » par une ligne doublement fléché. Les empilements et appariements entre nucléotides adjacents ne sont pas indiqués.

7.Bio-informatique structurale

Plusieurs outils bio-informatiques permettant l'étude de la structure tridimensionnelle d'ARN ont été développés au Laboratoire de Biologie Informatique et Théorique (LBIT). Ils ont servi de base à la réalisation de nos objectifs de recherche. Voici la liste de ceux utilisés tout au long de la thèse :

MC-Annotate annote un fichier de structure (format PDB) et en ressort toutes les caractéristiques structurales (interactions base-base) (Gendron et al., 2001). La méthode consiste en l'extraction de donnees structurales a partir des coordonnées atomiques contenues dans un fichier de type PDB. L'empilement entre deux base est determine par un critère de distance et les vecteurs normaux au plan des bases. Un appariement est aussi annoté selon un critère de distance, mais aussi grâce a un algorithme de flot maximum (Lemieux & Major, 2002)

MC-Cycle calcule la base minimale de cycle que l'on peut trouver dans un fichier de structure d'ARN (Lemieux & Major, 2006). Un cycle est forme par un graphe cyclique indivisible. L'annotation d'une structure avec *MC-Annotate* permet de la représenter sous forme de graphe ou les nœuds sont les nucléotides et les relations les arêtes de ce graphe. *MC-Cycle* recherche les sous graphes non divisibles du graphe issu de l'annotation d'un fichier PDB.

MC-Fold (Parisien et Major- Manuscrit en préparation) prédit la structure secondaire d'une séquence d'ARN en incluant les appariements non canoniques. En se basant sur des critères thermodynamiques, *MC-Fold,* qui prend en argument une séquence d'ARN, prédit plusieurs structures secondaires potentielles correspondantes.

MC-RMSD donne une distance (en angström) entre des molécules en calculant un écart-type (Root Mean Square Deviation) (Kabsch, 1978). Cet outil permet de calculer une distance moyenne entre deux ou plusieurs structures en utilisant les coordonnées atomiques des nucléotides ou atomes d'intérêt (généralement, l'intégralité des nucléotides). Il permet aussi d'aligner les structures les unes avec les autres afin de pouvoir visualiser un alignement structural. Il est aussi possible de calculer un

« clusterring » pour aider à la classification basée sur la valeur de RMSD de plusieurs structures entre elles.

MC-Search recherche un motif 3D donné en argument dans une base de données contenant des structures d'ARN (Hoffmann et al., 2003; Olivier et al., 2005). Basée sur un algorithme de reconnaissance de sous graphe dans un graphe, *MC-Search* permet la recherche efficace de motifs d'ARN dans un ou plusieurs fichiers PDB (voir glossaire).

MC-Sym permet le modélisation de modèle théorique de molécule d'ARN (Major et al., 1991; Major et al., 1993). Ce système répond à un problème de satisfaction de contraintes que sont les nucléotides et leurs interactions à modéliser. Ces contraintes sont données en argument au programme qui essaye de construire une structure 3D a partir d'une base de donnees de relations extraite de la PDB.

MC-View est une adaptation de *MC-Search* qui permet de rechercher un ensemble de motifs 3D prédéfinis dans un fichier de structure d'ARN (Lavoie LP – manuscrit en préparation en annexe). Un ensemble de motif d'ARN prédéfini (triloop, tetraloop, dinucleotide platform, entre autres) est recherché dans une structure. Le résultat est un script pour le programme de visualisation PyMOL. Celui-ci permet de voir une structure annotée par divers motifs.

Manipulant le même type de données, étant en développement constant et représentant d'excellents outils de travail, ces programmes sont souvent nommés dans les différents chapitres.

8.S'intéresser aux virus

Les virus sont responsables de beaucoup de maladies chez l'humain, certaines courantes et presque anodines comme le rhume provoqué par les rhinovirus ou bien dévastatrices et extrêmement contagieuses comme la fièvre hémorragique foudroyante due à la contraction du virus Ebola. Il en existe une grande diversité et ils touchent tous les organismes des trois grands règnes : les bactéries, les archaebactéries et les eucaryotes. Les scientifiques dénombraient environ 3000 virus en 2005 et pensent actuellement qu'il ne s'agit que de 1% du nombre réel de virus existants (Buchen-

Osmond, 2003). Des hiéroglyphes datant de 1400 avant J.C. décrivent un homme souffrant de poliomyélite. Ceci démontre que les virus pathogènes humains existent depuis plusieurs siècles (Pallansch & Roos, 2001).

Les virus sont des « Microorganismes infectieux rudimentaire [...] qui utilisent, pour la synthèse de leurs propres constituants, les matériaux de la cellule qu'ils parasitent, et qui se reproduisent à partir de leur seul matériel génétique » (adapté du grand dictionnaire terminologique - http://www.granddictionnaire.com). Ils sont composés d'une molécule d'acide nucléique entouré d'une coque de protéine (la capside) (Figure 14) et parfois d'une enveloppe. Ce sont des parasites absolus et bien qu'ils aient le potentiel d'infecter les organismes des trois domaines du vivant, ils ne font partie d'aucun règne.

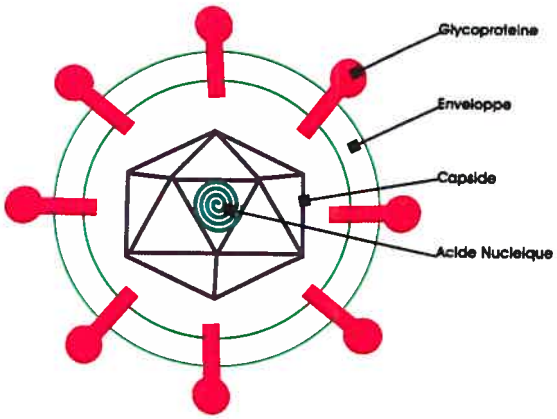


Figure 14 Exemple de structure de virus

Enveloppe : propre à certains virus (virus enveloppés).Capside : protection de l'acide nucléique. Glycoprotéines : protéines transmembranaires. Acide Nucléique : ARN ou ADN (simple brin ou double brin).

La classification des virus peut se faire de plusieurs façons et en impliquant différents critères tels que : la nature du génome, leur géométrie ou forme, leur hôte, etc. Une des classifications des virus la plus utilisée s'appuie sur le type d'acide nucléique qui encode son génome, c'est à dire ADN ou ARN. Cette classification est accessible dans l'ICTVdb (The Universal Virus Database of the International Commitee on

Taxonomy of Viruses) (Buchen-Osmond, 2003). Il existe aussi des virus à transcription inverse dont le génome peut être soit de l'ARN ou de l'ADN (Tableau III).

Tableau III Classification des virus par leur type de génome

Virus à ADN	
Groupe I	ADN à double brin
Groupe II	ADN à simple brin
Virus à ARN	
Groupe III	ARN à double brin
Groupe IV	ARN simple brin à polarité positive (type ARN messager)
Groupe V	ARN simple brin à polarité négative
Virus à ADN ou à ARN à transcription inverse	
Groupe VI	ARN simple brin
Groupe VII	ADN double brin

Il existe ainsi sept groupes de virus différents. Les virus dont le génome est composé d'ARN, groupes III, IV, V et VI, sont responsables de plus de 75% de toutes les maladies dues à des virus et ils incluent la majorité des pathogènes humains. Les virus à ARN seraient plus virulents à cause de leur incapacité d'auto-réparation en cas de mutations. Ils peuvent ainsi évoluer plus rapidement et donc « échapper » aux astuces du système immunitaire de leurs hôtes. Parmi ceux-ci, on retrouve le virus de la grippe, le virus Ebola, le virus du SRAS (Syndrome Respiratoire Aigu sévère), le virus de l'hépatite C, celui de la Dengue, ou bien encore celui de la Rage. L'étude de leur génome et de sa structure constitue donc un enjeu précieux dans l'identification de nouvelles cibles thérapeutiques et dans l'éradication des maladies qu'ils provoquent.

Nous avons choisi d'étudier la structure de l'ARN par le biais des virus à ARN en se basant sur l'hypothèse suivante : le génome d'un virus a un fort taux de mutation (10^{-5} à 10^{-3} mutation par nucléotide et par réplication contre 10^{-9} pour l'ADN chez les eucaryotes) mais son organisation dans l'espace est conservée afin de maintenir les

fonctionnalités permettant son expression. Par exemple, l'IRES (Internal Ribosomal Entry Site) situé en 5' des génomes viraux est fortement conservé au niveau structural et est essentiel à l'expression du génome (il permet l'initiation de la traduction) (Martinez-Salas & Fernandez-Miragall, 2004). Malgré des génomes différents, tous les virus s'exprime à une étape sous forme d'ARN : après la transcription pour les virus à ADN et avant la traduction pour les virus à ARN (Figure 15). Ne souhaitant pas nous limiter aux groupes de virus dont le génome est de type ARN, nous avons inclus dans notre étude tous les virus quelque soit leur génome. Ceci a été fait en se basant sur l'hypothèse qu'il existe des éléments d'ARN impliqués dans des processus d'expression qui sont conservés structuralement dans tous les virus.

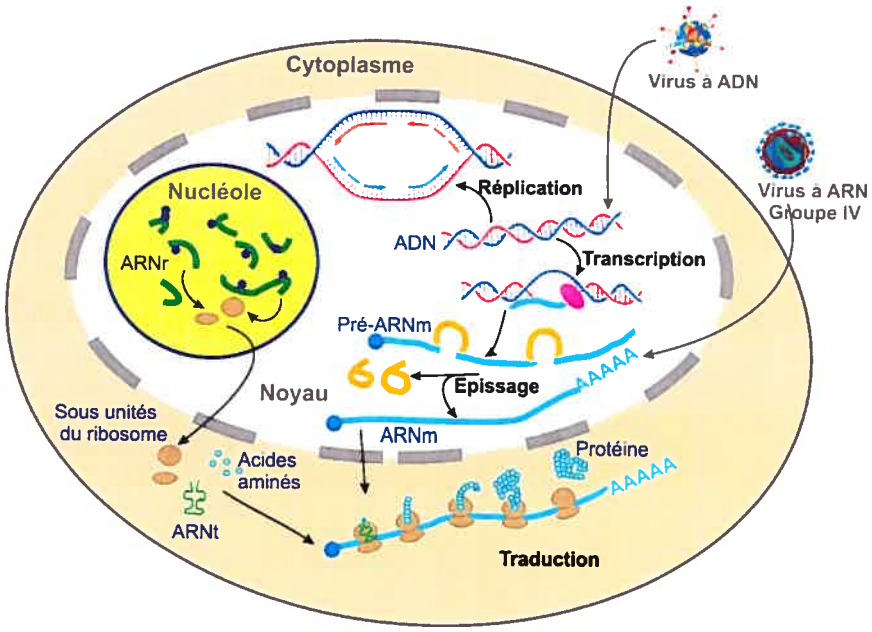


Figure 15 Voies d'expression empruntées par les virus dans une cellule eucaryote

Schéma représentant une cellule eucaryote et les étapes de réplication, de transcription, d'épissage et de traduction. Les deux types de virus pouvant attaquer cette cellule sont aussi représentés, montrant une différence fondamentale entre les deux voies d'expression qu'ils empruntent.

En étudiant les molécules d'ARN des virus au niveau de leur structure tridimensionnelle, nous nous basons sur l'hypothèse que malgré une forte variation en séquence, il y a conservation des motifs structuraux en trois dimensions afin d'assurer la fonction de l'ARN.

9.Hypothèses et objectifs

L'étude de la structure des macromolécules d'ARN est une étape cruciale pour la compréhension de l'expression de l'information génétique. Les motifs ARN, qui sont des blocs de construction de la structure tridimensionnelle des molécules d'ARN confèrent la fonction aux molécules d'ARN.

En analysant la structure tridimensionnelle de l'ARN, il pourrait être possible d'apporter des éléments de réponse sur les virus et leur habilité à détourner la machinerie cellulaire de leur organisme hôte et permettre l'identification de nouvelles cibles thérapeutiques.

10.Présentation des chapitres

Le chapitre 2 (article en préparation pour la revue Bioinformatics) propose une méthode alternative à celle existante pour la modélisation tridimensionnelle de molécule d'ARN par homologie. Cette méthode combine plusieurs outils bioinformatiques permettant d'utiliser les informations d'une structure tridimensionnelle d'ARN servant de structure modèle afin de modéliser des modèles théoriques de structure non déterminées par des méthodes biophysiques. A la différence des autres approches existantes « nucléotide par nucléotide » (qui consistent à remplacer les nucléotides non homologues un par un dans le fichier de coordonnées tridimensionnelles de l'ARN matrice), le processus utilise la modélisation par cycles. Il a été appliqué sur une tige-boucle d'origine virale : la tige-boucle D du site interne d'entrée du ribosome (IRES) dans la famille des *picornaviridae*. La combinaison des caractéristiques des cycles de l'ARN matrice et de l'alignement multiple avec les séquences homologues d'autres virus dont la tige-boucle D était inconnue, a permis de déterminer de nouvelles structures tridimensionnelles de tige-boucle D virale.

Le chapitre 3 (article en préparation pour la revue PLoS Biology) présente la caractérisation et la recherche du motif quartet afin d'appuyer une hypothèse de mimique de la région de la boucle 530 de l'ARNr du ribosome par le motif stem-loop 2 (s2m) du virus du Syndrome Respiratoire Aigu Sévère (SRAS), un élément extrêmement conservé chez les *coronavirus*. Le motif quartet est un motif contenant quatre nucléotides formant quatre paires de bases entre eux, planaire et ne possédant qu'un unique cycle. Le chapitre 3 démontre que ce motif est ubiquitaire dans les molécules d'ARN et qu'il est impliqué dans la stabilisation de l'ARN. Une mimique potentielle du quartet du s2m et du quartet formé par deux nucléotides universellement conservée de l'ARNr, un nucléotide provenant de l'ARNm et son nucléotide complémentaire situé à la deuxième position de l'anticodon l'ARNt dans le centre de décodage du ribosome a aussi été identifiée. Un court article détaillant les travaux effectués autour de cette découverte est actuellement en préparation pour PLoS One.

Le chapitre 4 (article en préparation pour la revue Nature Structural and Molecular Biology) présente l'annotation d'une base de donnée d'éléments structuraux d'ARN par des motifs d'ARN. Le but de ce travail était d'identifier de nouveaux motifs d'ARN potentiel. Deux bases de données ont été constituées, une contenant des structures tridimensionnelles d'ARN viraux et une autre contenant des descriptions de motifs structuraux récurrents et largement étudiés. Un outil *MC-View* a été développé afin d'attribuer sa composition en motifs à chaque structure de notre base de données (article en préparation pour Bioinformatics présenté en annexe). L'analyse des éléments d'ARN ainsi annoté a permis de déceler des caractéristiques intéressantes de certains motifs dont les bourgeons.

Tous les chapitres sont structurés en vue de leur soumission aux revues indiquées entre parenthèses.

CHAPITRE 2 Homology Modeling with RNA molecule: an example on viral hairpins

Homology modeling with RNA molecules: an example on viral hairpins

Emmanuelle Permal[1,2], Philippe Thibault[1,2] and François Major[1,2]

[1]Institute for Research in Immunology and Cancer, [2]Department of Computer Science and Operations Research, Université de Montréal, PO Box 6128, Downtown station, Montréal, QC, H3C 3J7, Canada

ABSTRACT

Motivation: There are few examples of homology modeling with RNA structures, despite its widespread use for proteins. However, the few approaches for RNA homology modeling use the same technique, which is to replace nucleotides within the coordinates of the native structure. As an alternative, we propose a new way for RNA homology modeling with structural fragments combined with a modeling system: *MC-Sym*.

Results: We propose a multi-step example of RNA homology modeling based on fragment modeling and protein homology modeling processes. Each fragment can be either extracted from an existing structure or built with our modeling system *MC-Sym*. We choose to test our approach on a well-known and conserved viral motif: the Stem-loop D that is essential to translation in Coxsackie viruses. Homology models produced with our methods showed a strong similarity (RMSD criterion from 2 Å to 3.5 Å) with the model 1 from PDB file of the stem-loop D from Coxsackievirus B3 (1RFR) used to extract essential information on base-base interactions that were crucial to modelling and also with the PDB file 1IK1 that was used to validate our method (1IK1 is the 3D-structure of the stem-loop D from Human Rhinovirus 14).

Availability: www-lbit.iro.umontreal.ca/mcsym/ or *MC-Sym* version 4.2.0
www.bioinfo.iric.ca/major_f/SLDpdb/

Contact: [redacted]

1 INTRODUCTION

Homology modeling is a combination of methods that allows the construction of an atomic resolution model of a molecule (protein or RNA) from its sequence. It relies on a template 3D structure and a sequence alignment that are used to build a theoretical structural model. Homology models are useful to predict interactions, molecular docking or differences between related RNA structures that have not been solved. Few examples of RNA homology modeling exist in contrast to the numerous ones in proteins structural studies (Kairys et al. 2006; Xiang 2006). This can be explained by the fact that protein sequences are more conserved since amino acids can be encoded by several codons. In this context, it seems obvious that a single nucleotide mutation does not have the same impact on a nucleotide sequence and a polypeptide.

Hairpins are double-stranded RNA structural element formed by a helix part, a loop and can contain bulges (Svoboda and Di Cara 2006). They serve as regulatory element as protein binding site like the TAR RNA in HIV-1 that interacts with TAT protein (Richter et al. 2002), as precursor for crucial molecules like miRNA (Lee et al. 2003), as regulator for localization of mRNA (Palacios and St Johnston 2001) and much more. Studying the 3D structure of hairpins is then a way to understand many molecular mechanisms. Homology modeling with RNA is an interesting alternative to biochemical methods such as Nucleic Magnetic Resonance spectroscopy or X-Ray crystallography to analyze hairpins structures.

Coxsackieviruses are enteroviruses involved in many human diseases. They belong to the *Picornaviridae* viral family that possesses a single-stranded positive RNA genome. In the 5' region of their genomic RNA stands the Internal Ribosome Entry Site (IRES), an essential element for uncapped RNA translation initiation. Here we took as example the coxsackievirus B3 (a B type human enterovirus) whose IRES contains seven domains (I to VII) including a "cloverleaf" domain (domain I) that can be split in four distinguishable parts (A to D); one of which is the stem-loop D (SLD). It is a highly conserved domain that interacts with 3C or 3CD viral proteins and holds a key role in viral translation and replication (Xiang et al. 1995; Kuyumcu-Martinez et al. 2004)3C

protease is essential to viral replication in Human Rhinovirus (Wanga and Chen 2007) that causes common cold ; studying its binding element is thus of therapeutic interest since it will provide information on the possibility of blocking viral expression. In 2004, Ohlenschläger and colleagues solved the structure of the SLD sub domain of Coxsackievirus B3 by Nucleic Magnetic Resonance spectroscopy (NMR) and suggested that its structural features (the D loop and the central pyrimidine bulge) were the keys to specific interaction with the 3C protein (Ohlenschlager et al. 2004). This is the motivation for choosing the stem-loop D as an example for cyclic RNA homology modeling since the structure allows the binding; not the nucleotide sequence. We also assume that the structure is conserved between viruses to ensure the function of the stem-loop D.

Here we introduce a way to do RNA homology modeling on viral hairpins with theoretical fragments as well as with PDB extracted structural fragments. We modeled the SLD of five different *Picornaviridae* viruses: Human Enterovirus A, Human enterovirus C, Human Enterovirus D, Human Rhinovirus 14, and Poliovirus 1 (respectively referred as HEV-A, HEV-C, HEV-D, HRV-14 and Polio in the rest of the document). We also completed the modeling for Human Enterovirus B (HEV-B) as a control since it is a coxsackievirus of type B. For each virus we choose to select the best model from results obtained by both methods (theoretical fragments and extracted fragments) by Root Mean Square Deviation (RMSD) distance criterion with the model 1 of PDB file 1RFR. We obtained twelve modeled structures with structural interaction based on model 1 from PDB 1RFR with a maximum RMSD distance to the template structure of 3.5 Å on all atoms but H.

2 METHODS

Each section of Methods follows the workflow displayed on figure 1.

2.1 Construction of the Datasets

Nucleotides sequences used for alignments and modeling were downloaded from the National Center for Biotechnology Information (NCBI) *Entrez genome* database (Accession codes: *NC_001612.1, NC_001472.1, NC_001428.1, NC_001430.1,*

NC_002058.3, K02121.1 for HEVA, HEVB, HEVC, HEVD, Poliovirus 1 and HRV-14 respectively). (cf. figure 1 – Sequences box)

The NMR structure of the SLD from Coxsackievirus B3 with accession code 1RFR (Ohlenschlager et al. 2004) was downloaded from the Protein Data Bank (PDB) ; this file contains 20 structures showing a maximal RMSD distance of 1.34 Å (on all atoms). Based on this observation, we arbitrarily chose one as a structural model: the model 1 of the 1RFR file that will be referenced as pdb1rfr-01.). (cf. figure 1 – Template structure box)

PDB file 1IK1 that contains one model of the NMR structure of SLD from Human Rhinovirus 14 (Huang et al. 2001) was also obtained from the Protein Data Bank. This structure was used to confirm the homology models that possessed triloops instead of tetraloops in their SLD loop part of the structure (e. g. HEV-C and HRV-14 models).

2.2 Annotation of the template model and cycle basis determination

The annotation is the critical part of the homology modeling process since it is at this step that relationships between nucleotides are collected (cf. figure 1 – Template annotation box). Using *MC-Annotate (Gendron et al. 2001)*, a program that take a PDB file as input and outputs a structural RNA graph displayed as a text file containing base-pairing annotation (using Leontis and Westhof nomenclature (Leontis and Westhof 2001)), pairing orientation, base-stacking, ribose pucker mode and base glycosidic bond torsion, we extracted all structural data from pdb1rfr-01 (figure 2a and 2b).

The cycle basis of pdb1rfr-01 was built using *MC-Cycle* (Lemieux and Major 2006)without considering base-backbone interactions (cf. figure 1 – Minimum cycle basis box) and figure 2c). We did not consider backbone-hydrogen bonding since the *MC-Sym* modeling tool did not support them. The *MC-Cycle* software computes the minimal cycle basis of a RNA molecule from a structural file in PDB file format where a cycle is a cyclic minimal graph.

2.3 Sequence alignment and secondary structure prediction

Sequence alignment was performed using clustalw (cf. figure 1 – Alignment box and figure 4a) (Thompson et al. 1994) with the six sequences to model and the sequence from the template structure extracted from PDB file 1RFR (figure 2a). Because the sequence of pdb1rfr-01 differed at positions 1 and 30 from other sequences used in the study (G1 becomes A1 and C30 becomes G30), this A: G pairing was excluded from the homology models. Two viruses, the HEV-C and the HRV-14, presented a deletion in their sequence of nucleotides; therefore, to confirm or infirm that the impact of the deletion was at the loop level (tetraloop changed into a triloop) we used *MC-Fold* (Parisien and Major – Unpublished). The aim of this step is to verify if within a deletion in a sequence a tetraloop is kept or not (cf. figure 1 – Alignment correction box). It gives a prediction of the most stable secondary structure based on specific RNA cyclic motifs and thermodynamics parameters for the six sequences (figure 4b). *MC-Fold* results showed the loss of pairing 14-17 for viruses HEV-A and Poliovirus but the diloop formed by these nucleotides (nucleotides 14 to 17) in the original annotation of 1RFR was kept intact for the modelling of their SLD. It also confirmed that for HEV-C and HRV-14, the loop was a triloop. We adapted the modeling with the alignment and the cycle basis conservation according to *MC-Fold* results which showed that a deletion in the sequence lead to the formation of a triloop.

2.4 Cyclic building blocks

We used two different methods to build our cyclic building blocks database (cf. figure 1 – Fragment production box). In the first method, structural fragments were extracted from the PDB database, whereas in the second one, we used theoretical cyclic blocks produced with the modeling system called *MC-Sym* (Major et al. 1991; Major et al. 1993). Note that in both cases, the annotation of relations from pdb1rfr-01 was used to define each cycle for all viruses. For example, cycle 6 from HEV-A was described as two contiguous canonical stacked pairing with the same annotation as the helical base pairs G9-C20 and U10-A19 from pdb1rfr-01. All blocks are PDB format files that contain 3D coordinates for each cycle.

In the first approach, all cycles from the SLD NMR structure are computed using MC-Cycle (Lemieux and Major 2006), and the structural cycle database corresponding

to this molecule is then built using the *MC-Search* program (Hoffmann et al. 2003; Olivier et al. 2005). *MC-Search* is a tool that allows us to describe each cycle from SLD structure with attributes such as base stacking and pairing and then to look for them in PDB structures. Each cyclic building block was output by *MC-Search* as a fragment PDB structure corresponding to the RNA graph description given as input.

All structural fragments, matching the descriptions given to *MC-Search* program, were then extracted from all RNA molecules available in the Protein Data Bank (PDB) considering all resolutions and methods (accessed on January 26th 2007).

In the second approach, we built a cycle database with *MC-Cycle* as previously described but we used *MC-Sym* modeling tool to produce theoretical structural building blocks instead of extracting them from the existing RNA structures of the protein databank. *MC-Sym* uses a filtered set of homogeneous transformation matrices (HTM) built from the set of structures with a resolution of at least 3Å (Major et al. 1991; Major et al. 1993). The input is a RNA graph where nodes are nucleotides and edges are interactions between nucleotides. Each edge of this graph is then search in the *MC-SYM* database contains edges encoded as transformations matrices.

Transformation matrices are obtained by combining nucleotides atomic coordinates in 4 dimensions with translations and rotations information needed to build a base-base interaction in 3D space (Major 2007). Since those matrices are based on nucleotides, they do not impose any constraints from the backbone (e.g. ribose and Phosphate group). Each cyclic structural fragment has been constructed (see examples in supplementary data) by giving a structural graph that encodes relations between nucleotides as input to the *MC-Sym* modeling tool. As output, we get theoretical cyclic fragments in PDB format file matching the annotation of the corresponding cycle in model 1 from the 1RFR file.

2.5 Merging the building blocks

As a final modeling step, we used *MC-Sym* to merge all blocks to form each structures for both methods used to produce the cyclic fragment databases (either *MC-Search* fragment or *MC-Sym* theoretical cycles). Merging is done on a common relation shared by the fragments, starting from the end loop (cf. figure 1 – Merging and selection

box). For example, cycle 1 and 2 are merge together by the relation between nucleotide 13 and 15. The process is repeated iteratively, with each step adding to the growing solution structure until all the constraints are satisfied and the entire stem-loop D structure is formed. Partial results are discarded as soon as incompatibilities are encountered.

2.6 Selection of the best models

One model per method and per type of virus was kept and submitted to minimization, where the selection criteria was the model's distance to the template SLD (pdb1rfr-01) in RMSD with *MC-RMSD*(Kabsch 1978) (calculated on all atoms except hydrogen and backbone atoms). See table 3 for RMSD distance in Å (cf. figure 1 – Merging and selection box).

2.6 Model minimization and backbone reconstruction with Amber-8.0

Using sander, a tool from the molecular dynamics package AMBER-8.0 that minimizes electrostatic and covalent links in a force field space (Case et al. 2005), we minimized all best models to replace the backbone in a two steps process: first a minimization on all atoms, then another minimization with all bases restrained (cf. figure 1 – Minimization box). This allowed us to refine our models in terms of energy and backbone torsion. We then checked the RMSD distances with pdb-1rfr-01.pdb with *MC-RMSD* (See table 3). We also compared the models obtained for HEV-C and HRV-14 with 1IK1 PDB file.

3 RESULTS

3.1 Sequence conservation in cycles

We present in table 1 the number of conserved cycle, the number of occurrences found for each cycle in PDB on nearly 800 structural RNA files and the number of models generated. In cases where the cycle was not found in PDB database we inserted a *MC-Sym* modeled fragment in the solution structure (indicated as **"modeled"** in the table).

The number of conserved cycles gave us an insight into how the sequence changes while still preserving each cycle's identity. For HEV-A, cycles were less

conserved in sequence (5/15) than with the cycles from other viruses (see figure 3) but substitutions were able to keep cycles integrity (G9A, C20U, A11U, U18A, U12C, C13U, and A28C where the first nucleotide is from pdb1rfr-01 reference sequence, the number is the residue number and the second nucleotide is from HEV-A sequence). Stable (at least two hydrogen bonds) base-pairs G9-C20, A11-U18, U12-G17, C13-G16 and G2-A28 were substituted by other stable pairing A9-U20, U11-A18, C12-G17, U13-G16 and G2-C28.

In HEV-C and HRV-14, cycles 1 to 3 were lost, according to *MC-Fold* predictions, but cycles that were included in the helical part of the SLD were conserved. For cycle description of their triloops, we used their sequences information and the type of interaction of the closing base-pair (e.g. pairing in 1RFR that closes the tetraloop). We found no occurrence for HEV-C triloop and one occurrence for HRV-14 triloop. This result shows that the triloop does not support all sequence possibilities with a wobble closing base-pair and that the triloop from HRV-14 is unique since the occurrence came from the 1IK1 PDB file that contains the NMR structure of HRV-14 SLD (Huang et al. 2001). This was confirmed by a recent study (Lisi and Major – in press) that enumerated all existing triloops in RNA molecules and showed that only one triloop is closed with a UG base-pair.

3.2 Modeling with structural fragments cycle

We used two types of fragments to build our models: Fragments extracted from the PDB database and fragments created with the modeling system *MC-Sym*. Table 1, summarizes results obtained with the first type of modeling blocks.

The number of occurrences of each cycle found in the RNA subset of PDB indicated that some cycles were very rare. For example, cycle2, cycle3, cycle9 and cycle10 from pdb1rfr-01 are unique since no other occurrences of those cycles were found elsewhere. Interestingly, some cycles, those marked as **"modeled"**, did not exist with differing sequences. This is the case for cycle3 and cycle9 in HEV-A, HEV-D and Poliovirus and demonstrates that a cycle formed by the sequence 5'CY3' and 5'YU3' paired by two Watson / Watson relations (cycle9) is unique. Since this feature was already mentioned (Ohlenschlager et al. 2004), this finding is a confirmation. For HEV-

C, we did not find the cycle11 that includes a pyrimidine-pyrimidine base-pair (U5-U24), showing that this type of pairing in a helix might not be favourable to form a stable structure (assuming that the most stable interactions are preferred in structural elements). One other surprising fact is that all cycles are not equivalent; for example a simple change in cycle8 (U22 in sequence from HEV-B to C22 in sequence from HEV-A) can drastically lower the number of occurrences (92 to 18 fragments) extracted from the structural database. But when, in the same cycle8, U7 and U22 paired by one hydrogen bond are substituted by a C7-C22 base-pair, the number of occurrences is then slightly enhanced (92 to 108 fragments).

This method gives few models but since only at least one model per sequence was expected, it is a quite satisfying result: for each virus all cycles merged together produced a full SLD structure (Table 1). This low number is a consequence of the fact that for each sequence at least one cycle is not abundant, for example there only one occurrence of the cycle 1 for the poliovirus cycle in the Protein Databank.

Selecting the best homology model per virus by RMSD distance with the template structure, based on all atoms but hydrogen, we saw that the best models had acceptable RMSD distance (less than 4 Å).

3.3 Modeling with theoretical cycles

We wrote for each cycle a script describing all relations needed by the modeling system *MC-Sym*. In contrast to models built with MC-Search fragment, those produced by *MC-Sym* fragments are numerous (table 2). It is quite interesting to see that the modeling system is able to produce cycles that were not in the database with relations extracted from the same database. But as with cycles collected with the first method, some cycles are less abundant than others (Data not shown); for example cycle1 from HEV-D that is formed by three consecutive C was only found once with *MC-Search* and built in only nine potential configurations with *MC-Sym*. All models show RMSD distances as good as those obtained with the previous method.

3.4 Comparison with stem-loop D structures

Table 3 displays all RMSD distances between the homology models and the pdb1rfr-01, computed with *MC-RMSD* (on all atoms but H), with both methods. The

RMSD distance for all SLD structures with tetraloop (HEV-A, HEV-B, HEV-D and Polio) is below than 3Å. These values suggest that we can trust the proposed pipeline to do homology modeling with hairpins without good sequence alignment. The best models from the first method (*MC-Search* cycles) are shown in figure 5 aligned with the template model pdb1rfr-01 using *MC-RMSD* program.

To compare results for triloops SLD structures, we used PDB file 1IK1 that contains the cap of the SLD from HRV-14 (nucleotide 9 to 22 of the numbering used for HEV-C and HRV-14 in our homology models). RMSD distances displayed on table 4 showed that our method succeeded even with the missing cycles caused by deletions in alignment. In the first method's model of HRV-14, the cycle1 came from the PDB file 1IK1; the RMSD distance indicates that the loop is a major element in the distance score since the model built with the *MC-Sym* fragment shows a larger RMSD value. The two best models from the *MC-Sym* method are shown aligned with 1IK1 (Figure 6).

4 DISCUSSIONS

In this study we modeled 12 RNA structures of stem-loop D for HEV-A to D, Poliovirus and HRV-14. Using a new approach for homology modeling with RNA molecules, we provide models for 4 different viruses: HEV-A, HEV-C, HEV-D and Poliovirus.

PDB files of models and PyMol sessions containing structure alignments are available at: www.bioinfo.iric.ca/major_f/SLDpdb/ where HEV-A is the coxsackievirus A16 responsible of the infectious "Hand, Foot, Mouth" disease; HEV-B is the coxsackievirus B1 (same serotype as coxsackievirus B3) which induces cardiomyopathy; HEV-C is the coxsackievirus A21 that causes common cold like HRV-14; HEV-D is the coxsackievirus A24 implicated in haemorrhagic conjunctivitis; and Poliovirus 1 causes poliomyelitis. We offer then a possibility to study SLD from other viruses and thus an interaction of considerable interest: the 3C protein – SLD interaction.

While we were writing this article, a new NMR model of the SLD from HRV-14 has been published (Headey et al. 2007). We were then able to compare our results obtained for MC-Sym modeled HRV-14 SLD with this structure using RNA graphs (See

supplementary material, figure 8). The two SLD shares the same annotation for pairings except for base-backbone pairings that are not supported by *MC-Sym* modeling system since those pairings are not included in the database of transformation matrices. The main differences are in stacking interactions but the RMSD distance on all atoms but H between the two structures (3.28 Å) shows that, despite those different relations, they are very close. From these observations, we can confirm that the methodology we applied for homology modeling with RNA is suitable to make models when the sequence alignment contains substitutions and deletions.

The starting hypothesis from Ohlenshlager et al. that the structure of the SLD primes on the sequence is supported here by the low RMSD values between all the models. The topological site formed by nucleotides 5 and 21 in the major groove of the SLD that is part of the 3C protein recognition site is also present in all our models that were modeled from independent cycles and different sequences (see Supplementary material, figure 7). Replacing the closing base-pair UG from the template molecule to CG in HEV-A and HEV-D does not affect the overall structure as predicted by Ohlenshlager et al. when they tested the binding with 3C protein in their study. Homology modeling with RNA can therefore help in data validation.

In this study, we used two methods to produce the RNA fragments: *MC-Search* and *MC-Sym*; despite the fact that for tetraloop SLD the RMSD values were lower with *MC-Sym* fragments, modeling with the *MC-Search* fragments has some advantages. The first one is that it is quicker because it takes less time to extract a cycle with *MC-Search* than to build a cyclic fragment with *MC-Sym*. The second is the information brought by the extraction process (cf. figure 1 – Fragment production box) on the number of occurrences of the cycles. The third one is the number of cycles extracted from the PDB that is drastically lower than with *MC-Sym* method (Data not shown). The last point is also an inconvenient since not all the wanted cycles are in the solved structures from PDB; in this case *MC-Sym* has a strong advantage on *MC-Search* because all cycles can be modeled.

Homology modeling with fragments could use fragments of any size. For example, a structure could be built by elements such as helical parts and non-helical

parts. This would avoid the use of numerous cycles fragments as in this study; however, in our example of the stem-loop D from enteroviruses, the models show a better RMSD distance from the native structure with cyclic fragments than with bigger fragments (Unpublished data – Modeling made with two fragments: an helical part and a loop part with resulting RMSD > to 3 Å). In this study we chose one model per sequence to illustrate the method, however, we could have consider all the models since the closest one in RMSD criterion may not be the most interesting. In further work, it would be better to analyze all models.

The pipeline used to do the homology modeling is quite simple and can be done with other bioinformatics tools than those used in this study. It is applicable to all RNA molecules that can be divided in cycles and could also be applied to RNA molecules with mixed fragments (cyclic and non-cyclic parts) but the merging would need to take this into account. It is also able to support deletion in the aligned sequences by a simple step of alignment correction with *MC-Fold.*

Further work would be to apply this approach to other examples, like bigger RNA molecules such as pri-pre-miRNAs or complex molecules composed of several hairpins such as the P4-P6 domain of Group 1 intron ribozyme.

ACKNOWLEDGMENTS

We would like to thank Martin Larose for helping with *MC-Search,* Marc Parisien for full access to *MC-Fold*, Romain Rivière for pdb2pdf tool, Véronique Lisi, Louis-Philippe Lavoie for useful discussions and reviewing this manuscript. We also like to thank Stephen Headey and Steven Pascal for sharing their PDB file of HRV-14 SLD with us. This work is supported by grants from the Canadian Institutes of Health Research (CIHR). FM is a member of the Institute for Research in Immunology and Cancer and of the Centre Robert Cedergren.

Conflict of interest: none declared

Figures, Tables and Legends

Table 1: Fragment cycle conservation and abundance in PDB.

This table contains the number of cycles found in our RNA structural database for each cycle described with a *MC-Search* script. Each cycle was described by a RNA graph where edges were interactions annotation from pdb1rfr-01 and nodes nucleotides from the aligned sequence. The upper line shows the conservation of cycles in sequence. The bottom line of the table gives the number of models that we got with the fragment produced with the first method (e.g. extraction of fragment with *MC-Search*).

Table 2: Theoretical models.

As an indication of the huge quantity of model that can be build the second method that uses *MC-Sym* to produce cycle, we show in this table the number of models obtained in our study.

Table 3: Root Mean Square Deviation table with parent template 1RFR in Å.

RMSD distances (on all atoms but H) were computed using the *MC-RMSD* program on all selected best models and pdb1rfr-01 before and after minimization with AMBER-8.0. RMSD values after minimization are shadowed in grey. The two first lines are for the first method of cycle extraction (*MC-Search*) and the two bottom lines are for the second method of cycle production (*MC-Sym*). Columns in light grey highlight the two stem-loop D structure that possess triloops instead of tetraloops: HEV-C and HRV-14 SLD models.

Table 4: Root Mean Square Deviation table with the SLD from HRV-14 1IK1 in Å.

RMSD distances (on all atoms but H) were computed using the *MC-RMSD* program on the two triloop SLD best models after minimization with AMBER-8.0 and PDB file 1IK1. The first line shows the first method distance (*MC-Search*) and the bottom line shows the second method distance (*MC-Sym*).

Figure 1: The RNA Homology modeling Flow.

Here we propose a process for homology modeling with RNA hairpins. All steps are in grey boxes, useful tool are in the right side and some descriptions on events are in the left side.

Figure 2: The Stem-loop D RNA from Coxsackievirus B3.

a) A view of the overall structure of 1RFR SLD made with PYMOL 1.0; backbone is represented as cartoon and nucleotides as sticks.

b) The SLD and its structural annotation: RNA graph using Leontis and Westhof nomenclature (Leontis and Westhof 2001) for pairings and one defined by Major and Thibault for stackings (Major 2007).

c) The SLD cycles and their number; they are the fragment that we take as sub template for each cycle that we extracted or modeled through this work.

Figure 3: Cycle conservation.

This figure shows all cycles needed to build a SLD by homology modeling. Cycles in dark gray are different in sequence from the template structure 1RFR. All substitutions in residue are circled on the RNA-graph. The arrow shows the place of a missing nucleotide that leads to the apparition of the triloop.

Figure 4: Alignment and secondary structure prediction.

a) Using clustalw we did a multiple alignment of the sequence to be modeled. Shadowed column show identities, dash show deletion in sequence. The sequence of the template PDB 1RFR is in bold at the bottom of the alignment. All sequences are referenced with their GI numbers. Numbers on the top are sequence numbering in HEV-B genomic sequence and numbers at the bottom are residues numbers from 1RFR PDB file.

b) Secondary structure prediction of all sequences present in the alignment, except the one from the template, made with *MC-Fold.* Those were used to correct the above alignment on cycle conservation for the two gapped sequences.

Figure 5: Superimposition of models built with database extracted fragments an their template.

Superimposition was made with the *MC-RMSD* tool on all atoms and the figure with PyMol 1.0 program. Backbone are shown as cartoon and bases as sticks. The parent structure 1RFR is colored in dark grey; HEV-A, HEV-B, HEV-D and Polio (produced with *MC-Search* method) are in red, blue, pink, and green respectively.

Figure 6: Superimposition of HEV-C and HRV14 models with PDB 1IK1.

As in figure 5, superimposition was made with the *MC-RMSD* tool on all atoms and the figure with PyMol 1.0 program. Backbone are shown as cartoon and bases as sticks. The parent structure 1IK1 is colored in dark grey; HEV-C and HRV-14 (produced with *MC-Sym* method) are in light orange, and dark cyan respectively. HEV-C and HRV-14 modeled structures have been shortened to nucleotide 9 to 22 to make the structural alignment with PDB file 1IK1.

Supplementary material

Figure 7: Sphere representation of important nucleotides for 3C protein binding.

Nucleotides 5, 15 and 21 important for 3C protein binding on SLD according to (Ohlenschlager et al. 2004) are represented as sphere on superimpositions. Nucleotides 16 and 23 that are crucial for activity are colored in orange.A) SLD with tetraloop B) SLD with triloop.

Figure 8: Comparison of the RNA graph of NMR SLD and *MC-Sym* SLD from HRV-14.

RNA graph of NMR SLD is at left and *MC-Sym* SLD at right. Red arrows indicate the differences between the two structures. Dashed blue lines show backbone-base interactions. A) Three-dimension structures. Left: the NMR model. Right: the MC-Sym model.

Table 1: Fragment cycle conservation and abundance in PDB.

	Virus					
	HEV-A	**HEV-B**	**HEV-C**	**HEV-D**	**Poliovirus**	**HRV-14**
Minimum cycle basis sequence conservation (based on 1RFR)	5/15	14/15	7/15	7/15	7/15	8/15
Cycle 1 occ.	40	32	**Lost cycle**	12 (2NOQ)	1	**Lost cycle**
Cycle 2 occ.	31	20 (1RFR)	**Lost cycle**	20 (1RFR)	31	**Lost cycle**
Cycle 3 occ.	**Modeled**	23 (1RFR)	**Lost cycle**	**Modeled**	**Modeled**	**Lost cycle**
Cycle 1b occ.	NA	NA	**Modeled**	NA	NA	1 (1IK1)
Cycle 4 occ.	4	252	6	252	252	6
Cycle 5 occ.	360	257	257	257	257	257
Cycle 6 occ.	554	2093	2093	2093	2093	2093
Cycle 7 occ.	1665	2999	2999	2999	2999	2999
Cycle 8 occ.	18	92	92	108	108	92
Cycle 9 occ.	**Modeled**	27 (1RFR)	27 (1RFR)	**Modeled**	**Modeled**	27 (1RFR)
Cycle 10 occ.	20 (1RFR)	20 (1RFR)	20 (1RFR)	20 (1RFR)	20 (1RFR)	20 (1RFR)
Cycle 11 occ.	92	92	**Modeled**	92	92	92
Cycle 12 occ.	2093	2093	5	2093	2093	2093
Cycle 13 occ.	1653	1653	818	818	818	818
Cycle 14 occ.	1843	1843	818	818	818	818
Cycle 15 occ.	8264	8264	2093	2093	2093	2093
Total number of SLD models	**6**	**2**	**19**	**4**	**3**	**4**

Table 2: *MC-Sym* fragments models

	Virus					
	HEV-A	**HEV-B**	**HEV-C**	**HEV-D**	**Poliovirus**	**HRV-14**
Total number of SLD models	605	438	72	266	694	554

Table 3: RMSD (all atoms but H) distance table from PDB 1RFR in Å.

Fragments Method	**HEV-A**	**HEV-B**	**HEV-C**	**HEV-D**	**Poliovirus**	**HRV-14**
MC-Search	2.36	2.74	2.73	2.32	2.23	3.10
MC-Search (After minimization)	2.31	2.70	2.71	2.26	2.16	3.06
MC-SYM	2.12	2.37	3.67	1.97	2.18	3.58
MC-SYM (After minimization)	2.05	2.31	3.65	1.84	2.17	3.52

Table 4: RMSD (all atoms but H) distance table from PDB 1IK1 in Å.

Fragments Method	**HEV-C**	**HRV-14**
MC-Search (After minimization)	2. 05	1.01
MC-SYM (After minimization)	2.36	3.09

Figure 1: Homology modeling for hairpin RNA pipeline

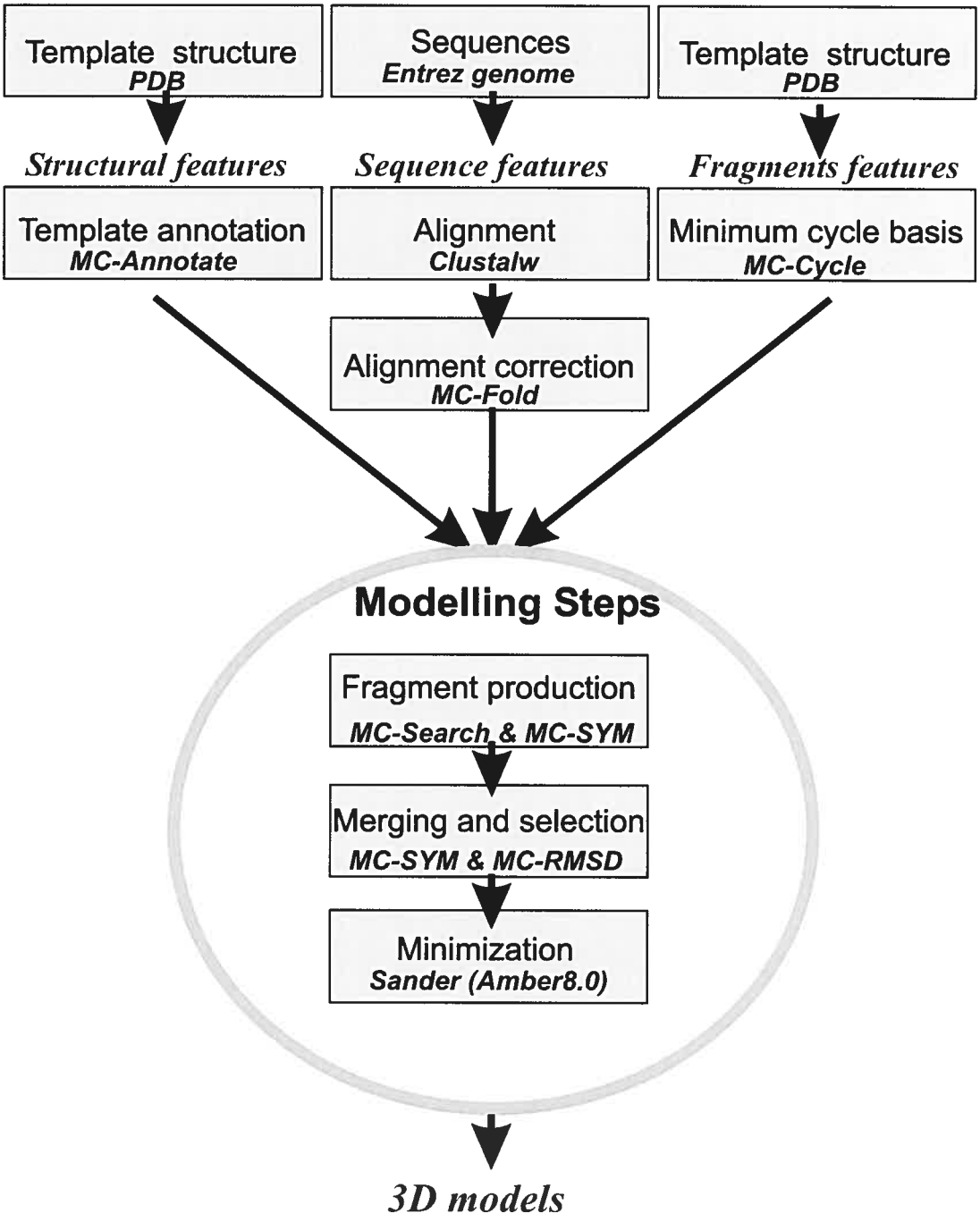
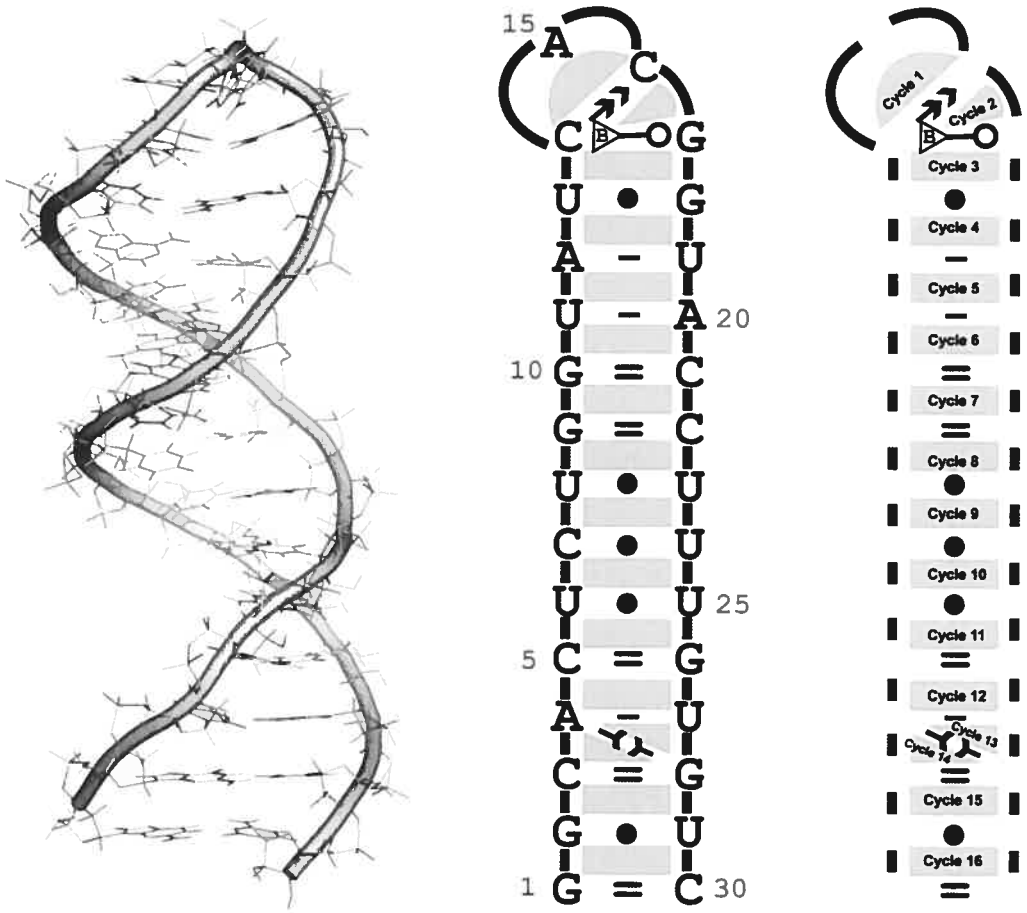


Figure 2: The Stem-loop D RNA from Coxsackievirus B3.



(a) (b) (c)

Figure 3: Cycles transposed from template to future models

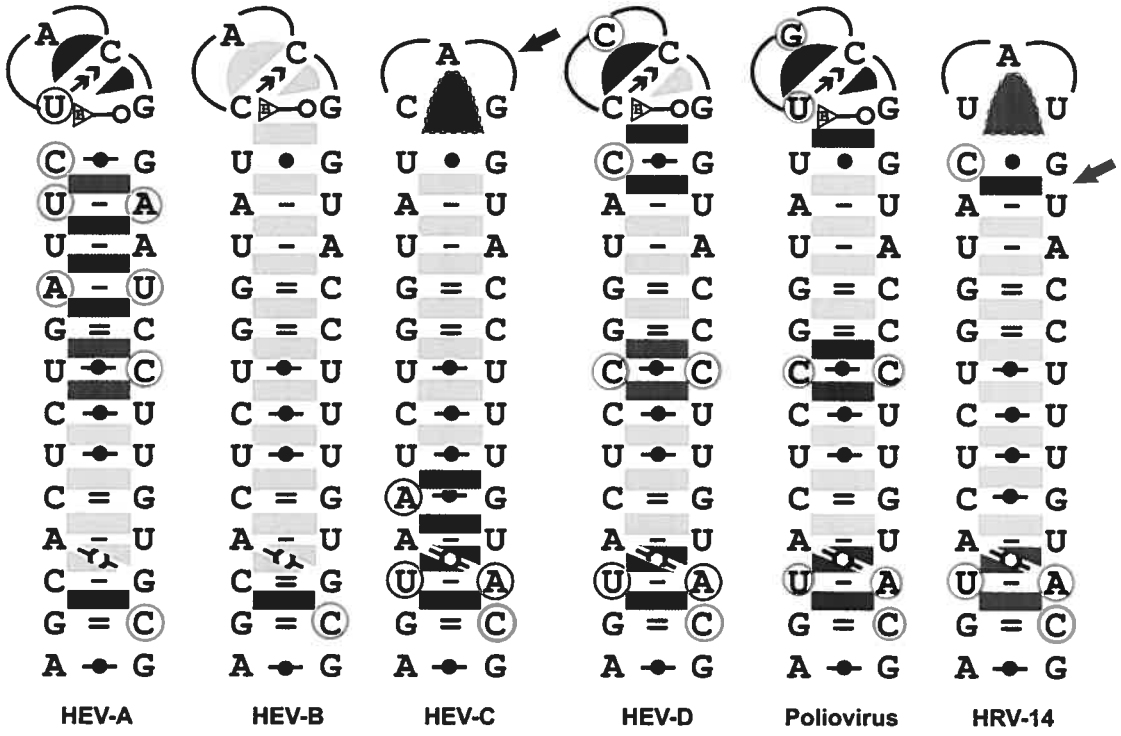


Figure 4: Alignment and secondary structure prediction

(a)

```
                                        48      54   57   61       67   71   74   75
gi|9627719|Human Enterovirus A          AGCAC UCU GAUU CUACGG AAUC CUU GUGCG
gi|9626677|Human Enterovirus B          AGCAC UCU GGUA UCACGG UACC UUU GUGCG
gi|9626433|Human Enterovirus C          AGUAA UCU GGUA UCA-GG UACC UUU GUACG
gi|9626436|Human Enterovirus D          AGUAC UCC GGUA CCCCGG UACC CUU GUACG
gi|12408699|Human Poliovirus            AGUAC UCC GGUA UUGCGG UACC CUU GUACG
gi|330029|Human Rhinovirus type 14      AGUAC UCU GGUA CUAUG- UACC UUU GUACG
1RFR|Stemloop D Coxsackievirus B3       GGCAC UCU GGUA UCACGG UACC UUU GUGUC
                                        1    6    9    13       19   23   26   30
```

(b)

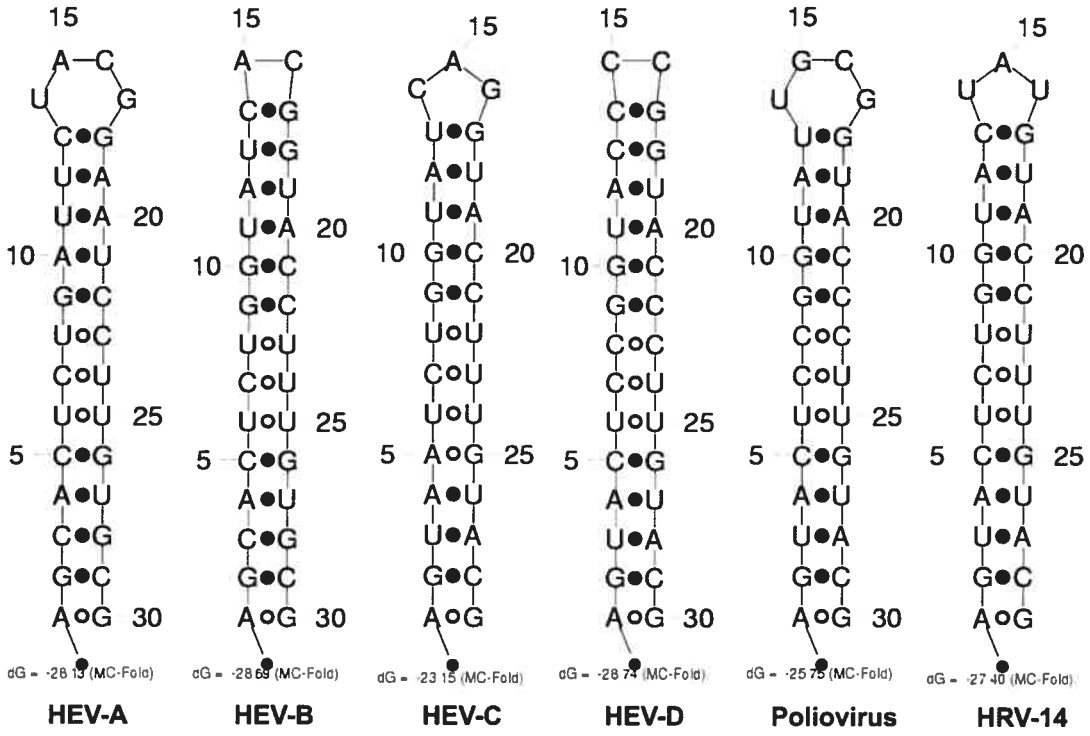
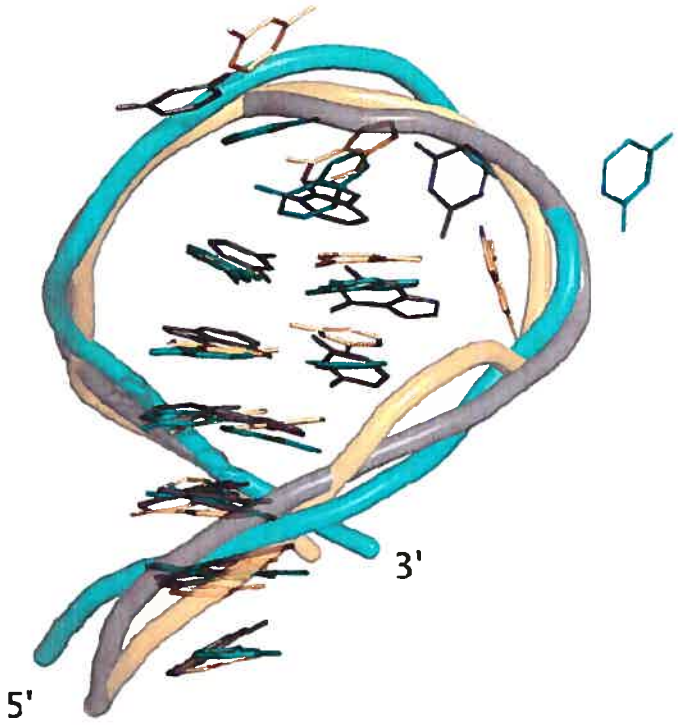


Figure 5: Superimposition of SLD models that possess tetraloops with the template structure.

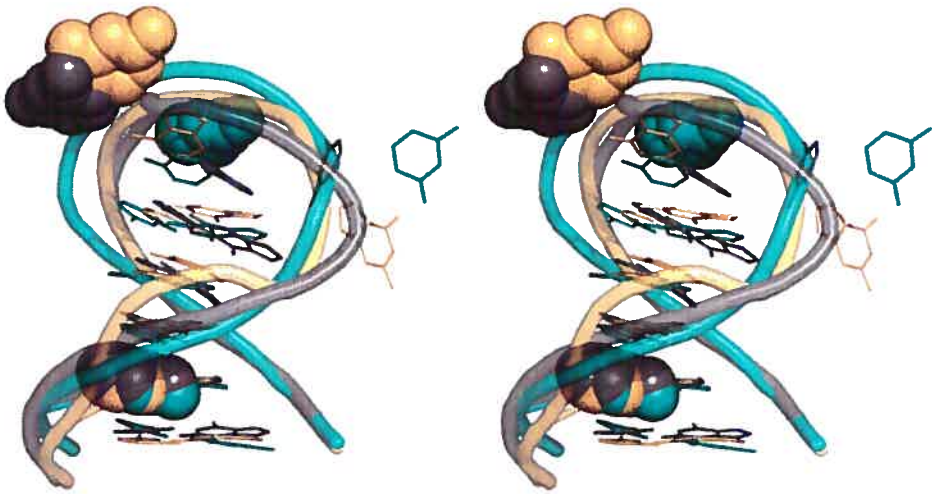
Figure 6: Superimposition of the SLD models that possess triloops with 1IK1.



SUPPLEMENTARY MATERIAL

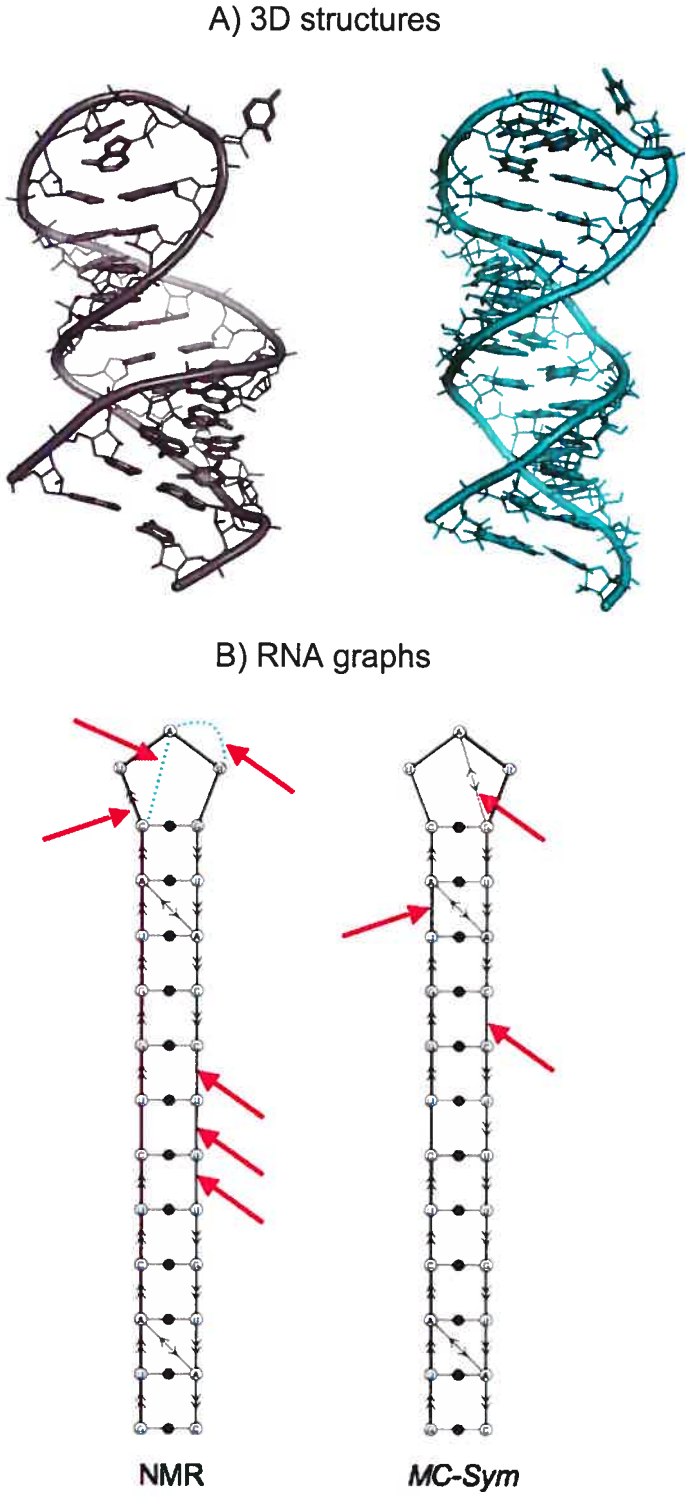
Figure 7: Sphere representation of important nucleotides for 3C protein binding.

SLD with tetraloop



SLD with triloop

Figure 8: Comparison of the RNA graph of NMR SLD and *MC-Sym* SLD from HRV-14.



REFERENCES

Case DA, Cheatham TE, 3rd, Darden T, Gohlke H, Luo R et al. (2005) The Amber biomolecular simulation programs. J Comput Chem 26(16): 1668-1688.

Gendron P, Lemieux S, Major F (2001) Quantitative analysis of nucleic acid three-dimensional structures. J Mol Biol 308(5): 919-936.

Headey SJ, Huang H, Claridge JK, Soares GA, Dutta K et al. (2007) NMR structure of stem-loop D from human rhinovirus-14. Rna 13(3): 351-360.

Hoffmann B, Mitchell GT, Gendron P, Major F, Andersen AA et al. (2003) NMR structure of the active conformation of the Varkud satellite ribozyme cleavage site. Proc Natl Acad Sci U S A 100(12): 7003-7008.

Huang H, Alexandrov A, Chen X, Barnes TW, 3rd, Zhang H et al. (2001) Structure of an RNA hairpin from HRV-14. Biochemistry 40(27): 8055-8064.

Kabsch H (1978) A discussion of the solution for the best rotation to relate two sets of vectors. ACTA Cryst Sec A: Cryst Phys Diff Theo Gen Cryst 34A: 827-828.

Kairys V, Gilson MK, Fernandes MX (2006) Using protein homology models for structure-based studies: approaches to model refinement. ScientificWorldJournal 6: 1542-1554.

Kuyumcu-Martinez NM, Van Eden ME, Younan P, Lloyd RE (2004) Cleavage of poly(A)-binding protein by poliovirus 3C protease inhibits host cell translation: a novel mechanism for host translation shutoff. Mol Cell Biol 24(4): 1779-1790.

Lee Y, Ahn C, Han J, Choi H, Kim J et al. (2003) The nuclear RNase III Drosha initiates microRNA processing. Nature 425(6956): 415-419.

Lemieux S, Major F (2006) Automated extraction and classification of RNA tertiary structure cyclic motifs. Nucleic Acids Res 34(8): 2340-2346.

Leontis NB, Westhof E (2001) Geometric nomenclature and classification of RNA base pairs. Rna 7(4): 499-512.

Major F (2007) RNA tertiary structure prediction. In Lengauer,T. (ed.), Bioinformatics: From Genomes to Therapies Wiley-VCH, Weinheim, Germany, Vol. I; Lengauer T, editor. 491–539 p.

Major F, Gautheret D, Cedergren R (1993) Reproducing the three-dimensional structure of a tRNA molecule from structural constraints. Proc Natl Acad Sci U S A 90(20): 9408-9412.

Major F, Turcotte M, Gautheret D, Lapalme G, Fillion E et al. (1991) The combination of symbolic and numerical computation for three-dimensional modeling of RNA. Science 253(5025): 1255-1260.

Ohlenschlager O, Wohnert J, Bucci E, Seitz S, Hafner S et al. (2004) The structure of the stemloop D subdomain of coxsackievirus B3 cloverleaf RNA and its interaction with the proteinase 3C. Structure 12(2): 237-248.

Olivier C, Poirier G, Gendron P, Boisgontier A, Major F et al. (2005) Identification of a conserved RNA motif essential for She2p recognition and mRNA localization to the yeast bud. Mol Cell Biol 25(11): 4752-4766.

Palacios IM, St Johnston D (2001) Getting the message across: the intracellular localization of mRNAs in higher eukaryotes. Annu Rev Cell Dev Biol 17: 569-614.

Richter S, Cao H, Rana TM (2002) Specific HIV-1 TAR RNA loop sequence and functional groups are required for human cyclin T1-Tat-TAR ternary complex formation. Biochemistry 41(20): 6391-6397.

Svoboda P, Di Cara A (2006) Hairpin RNA: a secondary structure of primary importance. Cell Mol Life Sci 63(7-8): 901-908.

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22): 4673-4680.

Wanga QM, Chen SH (2007) Human rhinovirus 3C protease as a potential target for the development of antiviral agents. Curr Protein Pept Sci 8(1): 19-27.

Xiang W, Harris KS, Alexander L, Wimmer E (1995) Interaction between the 5'-terminal cloverleaf and 3AB/3CDpro of poliovirus is essential for RNA replication. J Virol 69(6): 3658-3667.

Xiang Z (2006) Advances in homology protein structure modeling. Curr Protein Pept Sci 7(3): 217-227.

CHAPITRE 3 The Quartet Motif

The RNA Quartet Motif

Emmanuelle Permal[1, 2] and François Major[1, 2*]

1 Institute for Research in Immunology and Cancer, Université de Montréal, Montréal, Québec, Canada, 2 Department of Computer Science and Operations Research, Université de Montréal, Montréal, Quebec, Canada

*To whom correspondence should be addressed: E-mail: [redacted]

We have searched a new motif, the quartet motif, into all the available X-ray and NMR RNA structures of the Protein Data Bank; it is planar and composed by four nucleotides from multiple strands of RNA forming four base pairs. The motif was first identified in the stem loop II of SARS coronavirus near an atypical loop that was thought to mimic the 530 loop of ribosome. We did the study, using *MC-Search* and *MC-Annotate* programs, to test a potential mimicking of a GC-quartet located in the stem loop II of Severe Acute Respiratory Syndrome coronavirus and the decoding center of ribosome, a structure close to the 530 loop during translation of mRNA, composed by two nucleotides from the 16s rRNA (G530 and A1492), one from the mRNA codon (N35) and one from the tRNA anticodon (N2). We found 66 occurrences of this quartet in different RNA types: 16s rRNA, 23s rRNA, synthetic RNAs, viruses pseudoknots, group I intron, riboswitches and GTP aptamers. The quartets showed no consensus sequence, since the multiple strands repartition of the motif prevents us to track it in genomic data. But they are characterized by canonical and non-canonical pairing types and many hydroxyl-base interactions. This analysis also revealed that the GC-quartet found in stem loop 2 from SARS CoV also exists in the 16S rRNA of at least two species: *Thermus thermophilus* and *Escherichia coli* and that other quartets showed interesting aspects. Strong similarities also exist between the decoding center quartet and the GC-quartet found in SARS CoV, supporting the SARS CoV mimicking the ribosome for the recruitment of

ribosomal proteins. In conclusion, the RNA quartet is a highly structured and conserved ubiquitous motif that contributes to the stabilization of multiple strands interactions in RNA.

Abbreviations: DNA, Deoxyribonucleic Acid; NMR, Nucleic Magnetic Resonance Spectroscopy; PDB, Protein Data Bank file ID; RNA, Ribonucleic Acid; s2m, stem-loop II motif; SARS CoV, Severe Acute Respiratory Syndrome coronavirus; X-Ray, X-Ray crystallography

Introduction

It is well known that RNA molecules adopt a three-dimensional structure that is constitute of RNA motifs which hold key to RNA folding and function (Moore, 1999; Hendrix et al., 2005). A RNA motif can be described by its sequence (consensus in nucleotides composition), the interactions (pairing, stacking, adjacency), its function (binding site for a protein for example), or as a cycle of interactions (Lemieux & Major, 2006). Several structural motifs are described in literature or classified in some databases like SCOR (Klosterman et al., 2002).

Some of them like the tetraloops (Heus & Pardi, 1991; Jucker & Pardi, 1995; Lemieux & Major, 2006) or sarcin-ricin (Chen et al., 2006; St-Onge et al., 2007) have been widely studied. New motifs are discovered each year (Lee et al., 2006)(Steinberg & Boutorine, 2007).They are often found with new released X-Ray crystallography or NMR RNA tertiary structure. The GC-quartet motif of the stem loop II (s2m) from Severe Acute Respiratory Syndrome coronavirus (SARS CoV) was described by Robertson et al. in 2005 (Figure 1A); it differs from previously known quartets by its mixed nucleotide composition such as purine tetrads (Robertson et al., 2005) or pyrimidine tetrad (Pan et al., 2006). It was the first naturally occurring RNA nucleotide quadruplex ever observed.

The stem-loop 2 from SARS CoV containing this motif is located in the 3' untranslated region (3'UTR) of the viral genome, and exists in two other viral families: astroviruses and picornaviruses, indicating a possible role in viral pathogenicity (Jonassen et al., 1998; Bartlam et al., 2005). The GC-quartet motif from s2m is highly conserved in sequence through these families (see Robertson et al. 2005 for an alignment of the s2m of several viral genomes) except for avian nephritis virus where covariation in sequence is observed; the C20-G28 base-pair is replaced by a G20-C28 base-pair. A similarity between the loop of s2m and the 530 loop of the 16S ribosomal RNA (rRNA) subunit was noticed. It was hypothesized that the apical loop (GNRA-like pentaloop) mimics the rRNA 530 loop, from which could result a competition between the virus and the ribosome to recruit ribosomal proteins (Robertson et al., 2005). In order to characterize the structural foundation of the proposed mimicry, we computed a shortest cycle basis of the crystal structure of the SARS CoV s2m motif (Lemieux

& Major, 2006). A cycle is an indivisible sub graph of a RNA graph that represents the 3D structure with nodes as nucleotides and edges as interactions between nucleotides. As shown in (Figure 1B), the s2m structural motifs reported by Robertson et al., and in particular the GNRA-like pentaloop motif (Legault et al., 1998) and the GC-quartet, are included in its shortest cycle basis. Residues 22 to 26 of the s2m crystal structure form the GNRA-like pentaloop. Close to the pentaloop in the structure, Robertson and coworkers described the GC-quartet (residues 19, 20, 28 and 31) as a putative new motif because it had no resemblance with other known quartets (Robertson et al., 2005). We decided then to characterize the RNA quartet motif, also known as quadruplex, tetrad, tetraplex or G4 structures in DNA, by searching it in all available RNA tertiary structures in the Protein Data Bank (PDB) (Berman et al., 2000).

Mimicking is a mean to fool well established processes. Diverse examples exits : mimivirus that mimic a bacteria (Claverie 2005); tmRNA that mimic tRNA acceptor branches (felden 1998, Bessho 2007); HIV-1 A-rich RNA loop that mimics the tRNA anti-codon structure (Puglisi 1998); a Foot and Mouth IRES region that has evolved to create a mimic to tRNA (Serrano 2007);.

Quartets have been mentioned before as redundant motifs in DNA : G-quadrulex (Phan et al., 2006), A-quadruplex (Kondo et al., 2004) or C-quadruplex through i-motif (Snoussi et al., 2001) but never within a single double-helical structure like in SARS CoV s2m motif. In physiological ionic conditions, tetrad can form involving K+ or Na+ ion (Burge et al., 2006). In telomere, the free 3' single stranded ends can organize themselves as multiple stacked G-quartets. This characteristic reminds us of RNA that is able to form stabilizing long-range interactions and that a U-quadruplex exists in synthetic RNA (Pan et al., 2006). Recently, Burge et al. rigorously described DNA quadruplex topology and suggested that other topologies and structures may exist.

Here we report the analysis of heterogeneous quadruplexes in RNA molecules that we will call quartet to not interfere with previous definitions of nucleic quadruplexes (Burge et al., 2006). We define the RNA quartet as a structural motif formed when four nucleotides (adjacent or not) are paired together by four pairings and that possesses a unique cycle in its RNA graph. It has been analyzed on results obtained from on all RNA structural files

available during our study (PDB last accessed on November 17th 2006) and showed some interesting features on topology, sequence conservation and structural motif mobility. We have found nearly 300 occurrences of quartets in all PDB files with many different nucleotide compositions. This motif presented no consensus sequence, no apparent function but a role in the stabilization of the RNA molecule, was evolutionary conserved in some cases, can be shifted in the molecule when bound to a transcription factor or an antibiotic and can be sorted in topological classes.

Results

About Quartets Occurrences

After cleaning of the database redundancy by comparing all occurrences with each other, the number of single occurrences felt from 352 occurrences to 66 (see Supporting Information at Table S1 and Table S2 respectively).

Nucleotide Distribution in Quartets

Unlike known quadruplexes in DNA, RNA quartets are always composed at least by two different nucleotides (one exception: a AAAA quartet in *E. Coli* 23S rRNA). All quartets possess a minimum of two purines except one, the CCCG-quartet found in the P4-P6 region of group I intron (PDB 1GID) which has only one purine; and can exist without any pyrimidine at all, like the GAGA-quartet found in *H. Marismortui* 23s subunit (PDB 1S72).

Quartets composed by nucleotides coming from multiple strands made a sequence analysis awkward. However, we used circular string linearization (see Materials and Methods) to infer a unique sequence to each quartet. 24 sequences were found for the quartets extracted from PDB Database accessed on November 17th 2006 (AAAA, AAAC, AAAG, AAAU, AACG, AAGU, ACAU, ACCG, ACGG, AGAG, AGGG, ACGC, ACGU, AGAU, AGCG, AGCU, AGGU, AUGU, CCCG, CCGG, CGCG, CGUG, GGGG, UUUU) showing that not all sequences (out of $4^4 = 64$ possibilities) can be supported by the quartet.

Base-pairs in Quartets

The base-pair characteristics are described using the nomenclatures described by Leontis and Westhof in 2001 and Lemieux and Major in 2002 and detailed in Materials and Methods section {Lemieux, 2002 #2; Leontis, 2001 #30}.

A canonical base-pair, meaning Watson-Crick or Wobble pairing, is not required for the stability of the quartet, for example the GAGA-quartet in rRNA from the Large Subunit of the ribosome of *Haloarcula Marismortui* (PDB 1S72) has four pairings with the backbone: two 02'/Hoogsteen base-pair and two O2'/Bifurcated Hoogsteen base-pair.

The most common canonical and non-canonical interactions are between nucleotide G and C or G and A. All permutations of pairing are observed except UU and (UC/CU), reflecting that Uridine may only occurs once in this motif if we exclude the synthetic UUUU-quartet from our observations. All occurrences of quartet possess at least one hydroxyl-base interaction e.g. O2'-Hoogsteen or O2'-Sugar or O2'-Watson pairings. We did not observe all base-pairs showing that there might be preferred sides of nucleotides involved in the quartet motif; but the majority of pairings were of types ●● (Watson/Watson cis), ○□ (Watson/Hoogsteen trans) or O2'▶ (O2'/sugar) and no ◁▷, ◁□, ◁○, □□, or ■■ (see Material and Method for explanation on symbols). We also noticed that O2P-base pairing appears always with a trans relation in the quartet and that bifurcated-base pairing was often shared by O2' or O2P oxygen atoms. These 2 atoms showed a great potency in creating H-bonds with any nucleotides faces because all but one, O2P● pairing, O2'-base and O2P-base pairing have been observed. An interesting feature of this motif is the great involvement of nucleotide-backbone pairing because it is often forgotten in RNA motif analysis. However, we cannot make any statistics on these observations since the number of unique occurrences is very low, but it is possible to give some information on the nucleotides pairings mainly involved. We defined, then, two types of base-pairs: Base-Base and Base-Backbone and observed their contribution. Base-Backbone relations represented around 28% percent of all pairings involved in quartets.

Quartet Topological Classes

We can classify quartets in four topological categories: "canonicals" quartets (class I), "trans" quartets (class II), "bulged" quartets (class III) and "backbone" quartets (class IV). Some quartets belong to two classes (Figure 2). The first class, canonical quartet, is a highly

structured quartet with numerous and strong interactions (it is the s2m GC-quartets class with at least 8 hydrogen bonds; see Figure 2A) containing at least two canonical pairings (Watson-Crick or Wobble). Trans quartets possess a trans relation that introduces a mild twist in the quartet (Figure 2B; example of the trans quartet that we found in the decoding center quartet). Bulged quartets contain one canonical base-pair and a base expulsed outside from the quartet but linked by nucleotide/backbone pairing (Figure 2C; example of the bulged UAGA moving quartet). Backbone quartets have less H-bonds but a high participation of hydroxyl linkage (Figure 2D; an example in the 23S rRNA of *H. Marismortui* ribosomal subunit). A quartet can belong to two classes; one occurrence that we found was trans and bulged.

Conserved Quartets

We found two evolutionary conserved quartets in several species: a GC-quartet (Figure 2A) and a GGCA-quartet in ribosomal rRNAs. The GC-quartet is conserved in 16S rRNA between *Thermus thermophilus* and *Escherichia coli* 16s rRNA and contain two Watson-Crick base-pairs and two to three non-canonical base-pair including at least one with the backbone (involving the O2' from the sugar of nucleotide G1255 in *E. coli* or C1282 in *T. thermophilus*). In *T. thermophilus,* the two Watson-Crick pairing G1255 - C1282 and C1259 - G1276 are linked by a Sugar/Hoogsteen base-pair (G1255-C1259), one O2'/Sugar base-pair (C1282-G1276) and one O2'/Bifurcated Sugar (C1282-G1276) base-pair. In *E. coli*, non-canonical pairings are a Sugar/Hoogsteen base-pair (G1276-C1282) and one O2'/Hoosteen base-pair (G1255-1259). The GC-quartet is conserved in nucleotide composition, sequence and in tertiary interaction between the two procaryotes *T. thermophilus* and *E.coli.* The distance by 1.1 Å RMSD (all atoms but no H with HTM {Gendron, 2001 #3}) of these two quartets illustrates the conservation for this motif. However, there is no evidence of binding of this quartet from helix 41 (H41) of the small ribosomal subunit with protein or RNA. This quartet is connecting two non-canonical (e.g. not closed by a canonical base-pair) loops together: A1256-U1257-G1258 triloop and C1277-U1278-A1279-A1280-U1281 pentaloop. It may thus have an important role in structure stabilization of the molecule. We also observed mobility for this quartet in presence of the kasugamycin antibiotic in 16S rRNA from *T. thermophilus* {Schluenzen, 2006 #101}. The only two occurrences of GGCA-quartets are found in the two prokaryotes *Deinococcus radiodurans* (G697-G788-C804-A801) and *Escherichia coli* (G684-G775-C791-A788). They share the same nucleotide composition, the

same spacing between nucleotides positions (91 nucleotides between G and G-16 nucleotides between G and C -3 between C and A -104 between A and G) and interactions.

Moving Quartets

We called a moving quartet a motif that involves different residues can form the same network of interaction depending on the conformation changes. We found in the 16S rRNA the UAGA-quartet (Figure 2C) with exactly the same interactions in two different structural files of the 30s rRNA of *Thermus thermophilus* at two different positions. It seems that this quartet is kept in the structure of the ribosome when bound to a transcription factor but at a different location of the RNA. In the U810-A836-G846-A849 quartet, A836 belongs to a region of the ribosome (833-839) that interacts with transcription factor IF3C (PDB 1I94). We believe that this event causes the "jump" of the quartet from its "regular" position that is U827-A872-G869-A859 (PDB 1IBL). The GC-quartet evolutionary conserved is also moving when the 16S rRNA is bound to the kasugamycin antibiotic. In *Thermus thermophilus* ribosome, it is shifted from positions G1255-C1282-G1276-C1259 (PDB 1J5E) to G1237-C1264-G1258-C1241 (PDB 2HHH).

Double-stacked Quartet

In *E. coli* (PDB 1VOZ) we found two quartets stacked together: G27-U296-G301-C556 and G28-C295-G302-C555, building a combined motif that we named double stacked quartet (Figure 3 and 4). It is located in the 16s domain I of *E. coli*'s ribosome and seems to stabilize the tetraloop G297-A298-G299-A300 from helix H12 where the nucleotide G299 is thought to be important in ribosome because mildly deleterious if mutated {Yassin, 2005 #118}. This motif is highly stable since it is constituted by strong interactions: each base of the first quartet stacks with its backbone-linked base (for example G27 stacks with G28), each quartet possess a GC Watson-crick base-pair and either another GC watson-crick base-pair or GU wobble base-pair, both canonical interactions are linked by O2' – Bifurcated bonds to close the quartet. This double-stacked quartet is similar to DNA quadruplex since it obeys to the rules described by Burge and coworkers in 2006 except for the nucleotide composition. Using *MC-RMSD* we superimposed the double-stacked quartet with three DNA-quadruplexes (PDB 143D, 2GKU and 1XAV) and obtained an average RMSD distance on all atoms but H of 9Å for our combined motif and the quadruplexes. However, it is still possible to

hypothesize that this double-staked motif could be an ionophore like the DNA-quadruplexes (Figure 7).

A Quartet Formed by the Decoding Center

Two bases-quartets were found in the decoding center of the ribosome of *T.thermophilus* that is composed by nucleotides G530 and A1492 from 16S rRNA and U35 (PDB 1XMO) or A35 (PDB 1IBL) from mRNA and A2 (PDB 1XMO) or U2 (PDB 1IBL) from tRNA. The presence of these three types of RNA molecules in the same motif proves that the quartet is a multiple strands motif that plays an important role in stabilization of base-base interactions. This quartet that is close to the 530 loop was thought to be the one mimicked by the quartet from SARS CoV s2m. But the RMSD of GC-quartet located in helix 41 with the viral GC-quartet is pointing that there might be two candidates for the structural mimicry in the 16S rRNA (RMSD distances on all atoms but hydrogen with the s2m GC-Quartet were 2.3 Å and 1.53 Å, for the decoding center AAUG-quartet and for the helix 41 GC-quartet respectively)

Quartets' Context in RNA Molecules

Quartets formed by the decoding center (in PDB files 1XMO and 1IBL) are near a A-minor motif composed by nucleotides G37-U1-A1493 {Ogle, 2002 #95} which has been defined as a motif that "stabilize contacts between RNA helices, interactions between loops and helices, and the conformations of junctions and tight turns." by Nissen and colleagues {Nissen, 2001 #93}. One occurrence from *E. Coli* (C2500-G2523-U2625-A2545) also is near two type I A-minor interaction of Adenine with U2506-G2583 (with A of tRNA bound to A site) and/or A2450-C2501 (with A of tRNA bound to P site) {Nissen, 2001 #93}. All quartets found in 16S rRNA from *Thermus thermophilus* are indicated on Figure 4. All GC-quartets are located in the 3'-major domain of the 16S ribosomal RNA in helix 41 where ribosomal protein S9 and S14 interact. Some quartets are located at junction of multiple strands of RNA. These observations led us to conclude that the quartets are ubiquitous stabilizing elements in RNA.

Conservation and Covariation in Ribosomal and Viral RNA

Based on multiple alignments of bacterial 16s and 23s rRNA sequences and literature, we were able to see covariation and / or conservation of nucleotides involved in the different quartets. GC-quartets (G1255-C1259-G1276-C1282) are highly conserved (no variation in

sequence except in bacteria *Prevotella intermedia* where C1259-G1276 is replaced by G1259-U1276) among bacterial species (on 184 sequences analyzed without any redundancy) (Figure 5). GC-quartet from SARS CoV s2m is also highly conserved among astrovirus and coronavirus {Robertson, 2005 #100}. An other GC-quartet (G28-C295-G302-C555 that belong to the double stacked quartet) from 16s rRNA in bacteria showed covariation for G28 and C555 to A28 and U555. The second quartet of the double stacked quartet (G27-U296-G301-C556) is also highly conserved in 16s rRNA sequences. The decoding center quartet possesses a base-pair that co varies with the interaction of the second positioned base of the codon from mRNA and anti-codon from tRNA and a universally conserved base pair G530 and A1492 {Murphy, 2004 #92}. Some other quartets are highly conserved in bacterial 16S rRNA.

On Potential Mimicry of Quartets

The GC-quartet from SARS CoV s2m and the GC-quartet found in all 16S rRNA share a RMSD distance lower than 2Å (Figure 6). They also are strongly conserved in both cases in many species. Then, this motif might be important either for structure stabilization or for molecular functionalities. We first thought that the mimicry was at the level of the decoding center, but now it might be in the GC-quartet from 16S rRNA helix 41. However we cannot make any conclusion on a functional mimicry, since these hypothesis need to be experimentally tested, but the short RMSD distances between these quartets show that there is strong structural motif mimicry potential.

Quadruplexes in DNA and Quartets in RNA

Using *MC-RMSD* we compute a RMSD distance between 3 DNA-quartets and the RNA-double staked quartet and superimposed them (Figure 7). The RMSD distance was of 8.5 Å, which did not point to a strong similarity between RNA-quartet and DNA-quartet. However, the superimposition revealed that the double-stacked quartet could be an ionophore such as a DNA-quadruplex. This hypothesis is consistent with the fact that RNA molecules have shown that they can support diverse roles in cell such as being enzymes (ribozymes). Once again this hypothesis needs to be experimentally tested.

Discussion

Ribosomal Quartets Conservation Between Species

We observed on bacterial 16S rRNA alignments that many quartets possess highly conserved nucleotides. This confirmed the hypothesis that this multiple strands stabilizing structural motif might be essential for these highly structured RNA molecules and that further investigations and experimental confirmation need to be done.

A Structurally-conserved Stabilizing Motif

The quartet motif exists in many structures: 16s rRNA, 23s rRNA, synthetic RNAs, viruses pseudoknots, group I intron, riboswitches and GTP aptamer and is often localised at helix junctions; its frequency and its conservation shows that it is important in RNA structure. The quartet is, then, a new motif that has interesting and new features as the extensive use of base-backbone interactions. It is involved in stabilization of tertiary structure of RNA and is often found near stabilizing motif such as GNRA loop, A-minor motif or other quartet. We have made the proof that the GC-quartet from SARS CoV s2m is not unique and that it shares a strong similarity with the GCGC-quartet conserved in 16s from *E.coli, T. thermophilus*. The quartet motif is, then, a ubiquitous stabilizing motif involving multiple strands that shows no consensus sequence despite its strong conservation through species and that need to be further experimentally investigate to determine its involvement in RNA structure and folding.

Materials and Methods

PDB Structural Files

We studied all structural files available from the Protein Data Bank (PDB) accessed on November 17th 2006 {Berman, 2000 #28}. See supporting information for the list of all PDB file codes that contained quartets at http://www.bioinfo.iric.ca/major_f/Quartet/QTT_occurences.xls.

S2m SARS CoV PDB File and Motif Composition

As seen on Figure 1A, the s2m SARS CoV (PDB file 1XJR) can be decomposed in several motifs. This motif annotation describes a GNRA-like pentaloop in yellow, the planar quartet motif in red, and what remains of the structure in lightcyan (unknown motif) or in grey (watson-crick base-pairing forming helical part of the s2m). The file 1XJR is accessible from

the Protein Data Bank. Using the *MC-Annotate* {Lemieux, 2002 #2; Gendron, 2001 #3} computer program, we determined the tertiary structure interactions of the s2m X-ray crystal structure from SARS CoV (PDB 1XJR). We used the annotation of the quartet, i.e. the description of all its interactions in a general context, as a template to describe a four-nucleotide quartet (see descriptor *MC-Search* input in Figure 9 and explanations in RNA tertiary structure pattern matching). Four pairs of nucleotides paired together without making any cross pairing compose the motif.

Minimum Cycle Basis

We extracted from the PDB file 1XJR the apical part of the s2m motif (nucleotide 19 to 31) and we computed the cycle basis of this fragment with the program *MC-Cycle* (Figure 1B) {Lemieux, 2006 #1}. *MC-Cycle* outputs all cycles (e.g. all non-divisible sub-RNA graphs from the overall RNA graph formed by all nucleotide-nucleotide interactions that are present in the RNA motif s2m) from the 1XJR file. The cycles from the GNRA-like pentaloop are in yellow, the cycle from the quartet is red, the cycle in grey represents a canonical cycle involving Watson-crick base-pairs and the lightcyan cycle is formed by two bulged out nucleotide and a non-canonical GC pair.

Nomenclature

We label the base pairs observed in a structure by using a nomenclature that was first proposed by Leontis and Westhof {Leontis, 2001 #30}, and then augmented by Lemieux and Major {Lemieux, 2002 #2}. We use the following symbols to describe the edges of the chemical groups involved in Hydrogen bonds: ▶ sugar edge, ■ Hoogsteen edge, ● Watson-Crick edge, O2' when the 2'OH is involved, O2P when the O2 atom of the phosphate group is involved, B when the Hydrogen bond bifurcates between two groups, and the purine C8, when it is involved. There is a total of 34 different base-pairing types if we distinguish the relative orientation of the backbone: *cis* for the same side of the median that split in halves the base pair plane, indicated by the black symbols (cf. ■), and *trans* for two backbones on the opposite sides of the median, indicated by white symbols (cf. □) (Figure 8). In this nomenclature, the canonical Watson-Crick base pair is noted ●. Bh stands for a Bifurcated base pair that involves a chemical group on the Hoogsteen edge and Bs is for a Bifurcated base pair that involves the sugar edge (Figure 8).

RNA Tertiary Structure Pattern Matching

The comparison between all quartets motifs was based on Root-Mean-Square-Deviation {Kabsch, 1978 #84} computed with *MC-RMSD*; this allows us to reduce the set to analyse because of the database redundancy (e.g. at least 31 structures of the 23S rRNA of *H. marismortui* ribosome are available in the PDB). All the motifs were extracted from their respective PDB files {Berman, 2002 #71}. Instances of the four-nucleotide platform were searched using the *MC-Search* computer program developed in our laboratory {Hoffmann, 2003 #50; Olivier, 2005 #49}. *MC-Search* allows the research of several motifs in PDB formatted 3D structural file solved by NMR or crystallography. Helped by a nomenclature that describes three major parameters (the sequence respecting the IUPAC code (http://www.iupac.org/dhtml_home.html), the residues and the relations between them), an exhaustive script of a 3D structural motif can be defined. It works by applying the Ullman's algorithm of isomorphism {Ullman, 1976 #111} on the motif script translated into a structural graph and on the database of structural files, also translated into graphs, where the motif is searched. Translation into structural element graph of the database is performed by *MC-Annotate* {Lemieux, 2002 #2; Gendron, 2001 #3}. The input pattern was defined from the s2m tertiary structure {Robertson, 2005 #100}, and was obtained from the results of *MC-Annotate*. The pattern describes the nucleotides' base pairing and stacking interactions. The sequence is busted by introducing the nucleotides in four "sequence" statements, and the nucleotide types were generalized by using the letter 'N'. The pattern was searched in all the structures of the PDB. The identified structures were compared by Root-Mean-Square-Deviation.

Results Filtration

We filtrated the results according to the following criterion: 1) A quartet must form a unique indivisible cycle, 2) A quartet must be almost planar (to be compared efficiently with the model quartet: GC-quartet from SARS CoV) and 3) Pertinence of the quartets (all quartets from structural files showing some interpretation problems were removed.)

Circular String Linearization

To assign a unique sequence to each quartet, we computed a script that choose a place to cut in the circular string formed by the base from the quartet. This produced linear strings, four in our case that we read in both directions because the cycle that we cut could be read

clockwise or anti-clockwise. We then generated eight strings per quartet's base. We allocated the lexically smallest string to the quartet resulting of a unique string per quartet.

16s rRNA Quartet Conservation and Covariation analysis

Structural multiple alignments of 16s bacterial rRNA were provided by Westhof and colleagues. These structural alignments that possess near 800 sequences were reduced to 184 sequences by avoiding species redundancy. We used BioEdit software to visualize and analyze quartet conservation and covariation in these sequences {Hall, 1999 #69}.

Supporting Information

Coordinates and annotations supported by the RNA Ontology consortium {Leontis, 2006 #90} of all quartets of our study are available at http://www.bioinfo.iric.ca/major_f/Quartet/.

Accession Numbers

The RSCB Protein Data Bank accession number for all RNA structures discussed in this paper are given through the text and are available in Table S1 and Table S2 of Supporting Information.

Acknowledgements

We thank Martin Larose and Philippe Thibault in our laboratory, as well as Daniel Lamarre and Claude Perreault at the IRIC for useful discussions, Eric Westhof for bacterial 16s rRNA sequences alignments and Robin Gutell for the 16sRNA picture. This work was supported by a grant from the Canadian Institutes of Health Research (CIHR) (MT-14604) to FM. FM is a CIHR investigator and a member of the Centre Robert-Cedergren of the Université de Montréal.

Competing interest. The authors have declared that no competing interests exist.

Author contributions. The experiments were conceived, designed, and interpreted in a cooperative effort among all of the authors. EP determined all the RNA quartet occurrences. FM designed the project.

References

Bartlam M, Yang H, Rao Z (2005) Structural insights into SARS coronavirus proteins. Curr Opin Struct Biol 15(6): 664-672.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN et al. (2000) The Protein Data Bank. Nucleic Acids Res 28(1): 235-242.

Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE et al. (2002) The Protein Data Bank. Acta Crystallogr D Biol Crystallogr 58(Pt 6 No 1): 899-907.

Burge S, Parkinson GN, Hazel P, Todd AK, Neidle S (2006) Quadruplex DNA: sequence, topology and structure. Nucleic Acids Res 34(19): 5402-5415.

Chen C, Jiang L, Michalczyk R, Russu IM (2006) Structural energetics and base-pair opening dynamics in sarcin-ricin domain RNA. Biochemistry 45(45): 13606-13613.

Gendron P, Lemieux S, Major F (2001) Quantitative analysis of nucleic acid three-dimensional structures. J Mol Biol 308(5): 919-936.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41: 95-98.

Hendrix DK, Brenner SE, Holbrook SR (2005) RNA structural motifs: building blocks of a modular biomolecule. Q Rev Biophys 38(3): 221-243.

Heus HA, Pardi A (1991) Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. Science 253(5016): 191-194.

Hoffmann B, Mitchell GT, Gendron P, Major F, Andersen AA et al. (2003) NMR structure of the active conformation of the Varkud satellite ribozyme cleavage site. Proc Natl Acad Sci U S A 100(12): 7003-7008.

Jonassen CM, Jonassen TO, Grinde B (1998) A common RNA motif in the 3' end of the genomes of astroviruses, avian infectious bronchitis virus and an equine rhinovirus. J Gen Virol 79 (Pt 4): 715-718.

Jucker FM, Pardi A (1995) Solution structure of the CUUG hairpin loop: a novel RNA tetraloop motif. Biochemistry 34(44): 14416-14427.

Kabsch H (1978) A discussion of the solution for the best rotation to relate two sets of vectors. ACTA Cryst Sec A: Cryst Phys Diff Theo Gen Cryst 34A: 827-828.

Klosterman PS, Tamura M, Holbrook SR, Brenner SE (2002) SCOR: a Structural Classification of RNA database. Nucleic Acids Res 30(1): 392-394.

Kondo J, Adachi W, Umeda S, Sunami T, Takenaka A (2004) Crystal structures of a DNA octaplex with I-motif of G-quartets and its splitting into two quadruplexes suggest a folding mechanism of eight tandem repeats. Nucleic Acids Res 32(8): 2541-2549.

Lee JC, Gutell RR, Russell R (2006) The UAA/GAN internal loop motif: a new RNA structural element that forms a cross-strand AAA stack and long-range tertiary interactions. J Mol Biol 360(5): 978-988.

Legault P, Li J, Mogridge J, Kay LE, 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(2): 289-299.

Lemieux S, Major F (2002) RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. Nucleic Acids Res 30(19): 4250-4263.

Lemieux S, Major F (2006) Automated extraction and classification of RNA tertiary structure cyclic motifs. Nucleic Acids Res 34(8): 2340-2346.

Leontis NB, Westhof E (2001) Geometric nomenclature and classification of RNA base pairs. Rna 7(4): 499-512.

Leontis NB, Altman RB, Berman HM, Brenner SE, Brown JW et al. (2006) The RNA Ontology Consortium: an open invitation to the RNA community. Rna 12(4): 533-541.

Major F (2007) RNA tertiary structure prediction. In Lengauer,T. (ed.), Bioinformatics: From Genomes to Therapies Wiley-VCH, Weinheim, Germany, Vol. I; Lengauer T, editor. 491–539 p.

Murphy FVt, Ramakrishnan V, Malkiewicz A, Agris PF (2004) The role of modifications in codon discrimination by tRNA(Lys)UUU. Nat Struct Mol Biol 11(12): 1186-1191.

Nissen P, Ippolito JA, Ban N, Moore PB, Steitz TA (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. Proc Natl Acad Sci U S A 98(9): 4899-4903.

Ogle JM, Murphy FV, Tarry MJ, Ramakrishnan V (2002) Selection of tRNA by the ribosome requires a transition from an open to a closed form. Cell 111(5): 721-732.

Olivier C, Poirier G, Gendron P, Boisgontier A, Major F et al. (2005) Identification of a conserved RNA motif essential for She2p recognition and mRNA localization to the yeast bud. Mol Cell Biol 25(11): 4752-4766.

Pan B, Shi K, Sundaralingam M (2006) Base-tetrad swapping results in dimerization of RNA quadruplexes: implications for formation of the i-motif RNA octaplex. Proc Natl Acad Sci U S A 103(9): 3130-3134.

Phan AT, Kuryavyi V, Patel DJ (2006) DNA architecture: from G to Z. Curr Opin Struct Biol 16(3): 288-298.

Robertson MP, Igel H, Baertsch R, Haussler D, Ares M, Jr. et al. (2005) The structure of a rigorously conserved RNA element within the SARS virus genome. PLoS Biol 3(1): e5.

Schluenzen F, Takemoto C, Wilson DN, Kaminishi T, Harms JM et al. (2006) The antibiotic kasugamycin mimics mRNA nucleotides to destabilize tRNA binding and inhibit canonical translation initiation. Nat Struct Mol Biol 13(10): 871-878.

Snoussi K, Nonin-Lecomte S, Leroy JL (2001) The RNA i-motif. J Mol Biol 309(1): 139-153.

St-Onge K, Thibault P, Hamel S, Major F (2007) Modeling RNA tertiary structure motifs by graph-grammars. Nucleic Acids Res. Vol. 35, No. 5, pp.1726-1736, 2007.

Steinberg SV, Boutorine YI (2007) G-ribo: A new structural motif in ribosomal RNA. Rna 13(4): 549-554.

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

Yassin A, Fredrick K, Mankin AS (2005) Deleterious mutations in small subunit ribosomal RNA identify functional sites and potential targets for antibiotics. Proc Natl Acad Sci U S A 102(46): 16620-16625.

Figure 1. Crystal Structure of SARS s2m and Minimum Cycle Basis of the Loop

(A) The SARS CoV s2m motif: The helical parts are shown in grey; the GC-quartet in red; and, the GNRA-like pentaloop in yellow. All other nucleotides are shown in light cyan. We used the PyMOL (PyMOL.sourceforge.net) computer program to create this image. (B) Shortest cycle basis for the region of nucleotides 19 to 31: Solid shapes represent the interaction cycles. The colors match those in (A). Solid lines represent the backbone; doted lines for a single H-bond; double arrow for stacking; double lines for canonical Watson-Crick base pairs; and, triangle-square for the Sugar/Hoogsteen trans base pair, using the Leontis and Westhof nomenclature representation for base pairing {Leontis, 2001 #98} and the Major and Thibault nomenclature representation for base-stacking {Major, 2007 #103}.

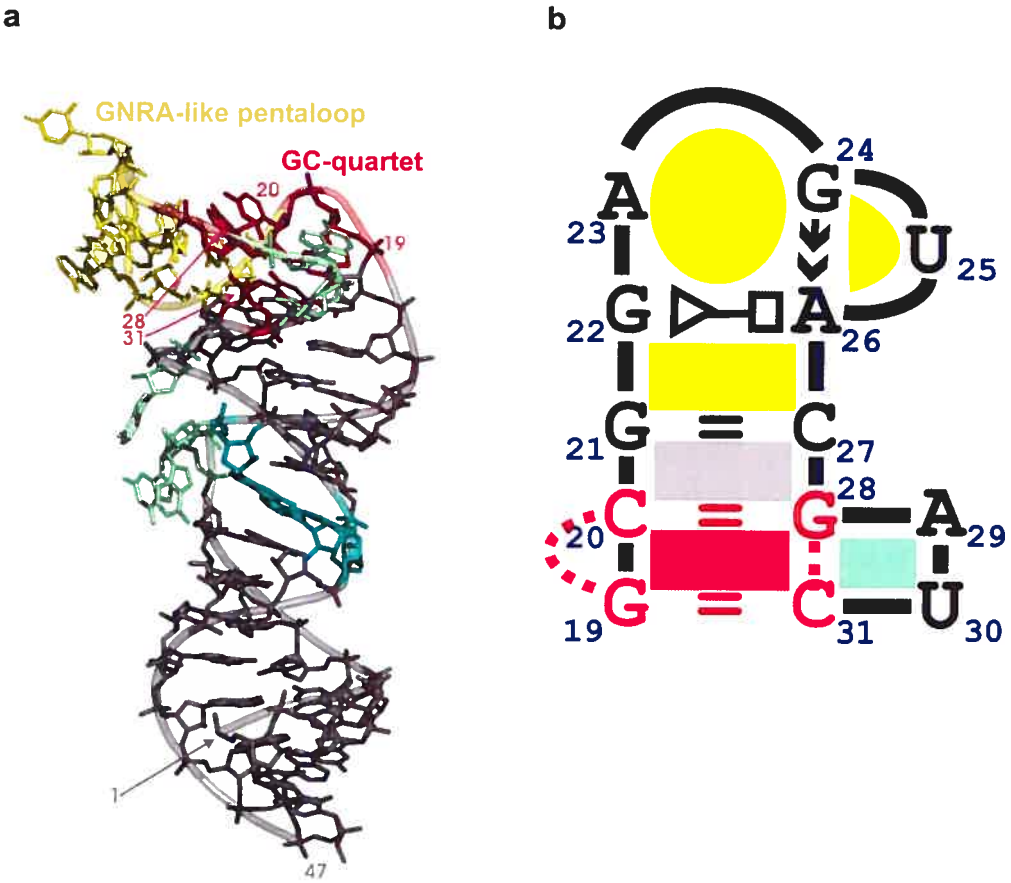


Figure 2. The Four Topological Classes of Quartets

(A) Class I = Canonical class; example of the conserved GC-quartet (B) Class II = Trans class; example of the decoding center quartet (C) Class III = Bulged class, example of the moving quartet (D) Class IV = Backbone class. Classes were made from the annotation given by *MC-Annotate* {Gendron, 2001 #69} and RMSD computation on alla quartets occurences.

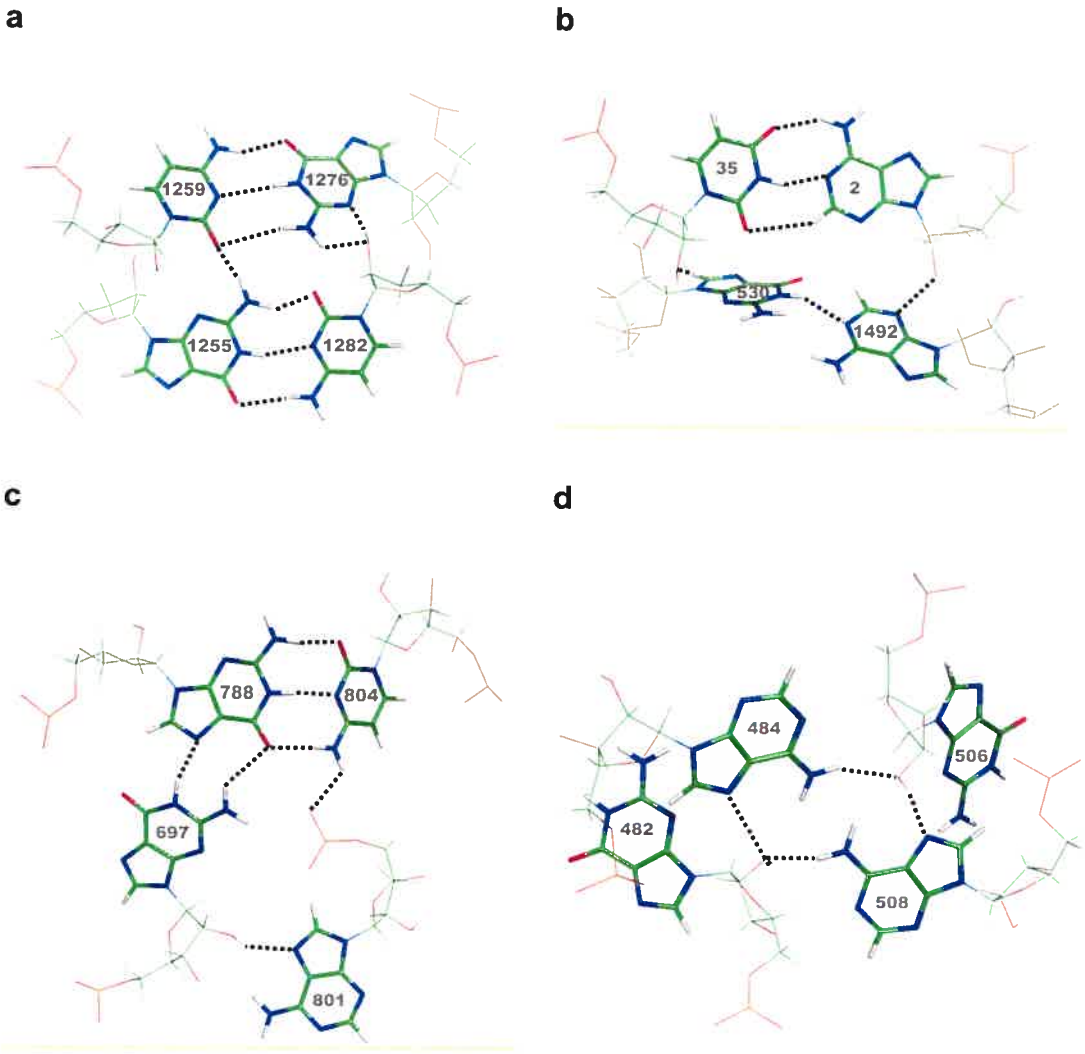


Figure 3. Double-stacked Quartet

The Double-stacked base quartet is represented as a structural graph composed by the two quartets stacked together, GCGC and GCUG quartets, and linked by the backbone. Dotted lines represent hydrogen bonds. This motif is located near H12 tetraloop in 16S rRNA domain I of *E. Coli* ribosome.

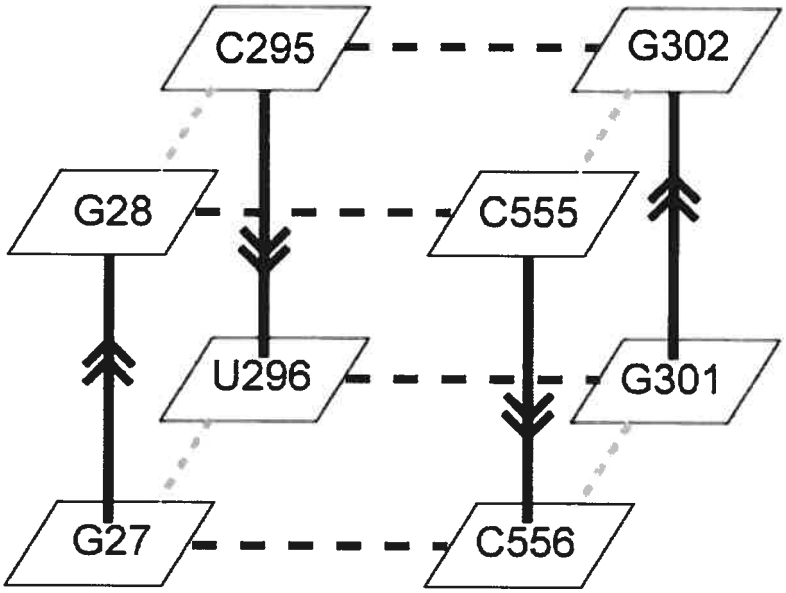


Figure 4. Quartets in the Secondary Structure of the 16s rRNA from *Thermus thermophilus*

Structural localization of all quartets found in *Thermus thermophilus* 16s rRNA. They are showed as colour circles linked by dashed lines. Two quartets, that form the double-stacked quartet motif, are also represented; their nucleotides are symbolised with squares. Numberings and nucleotides respect PDB file annotation.

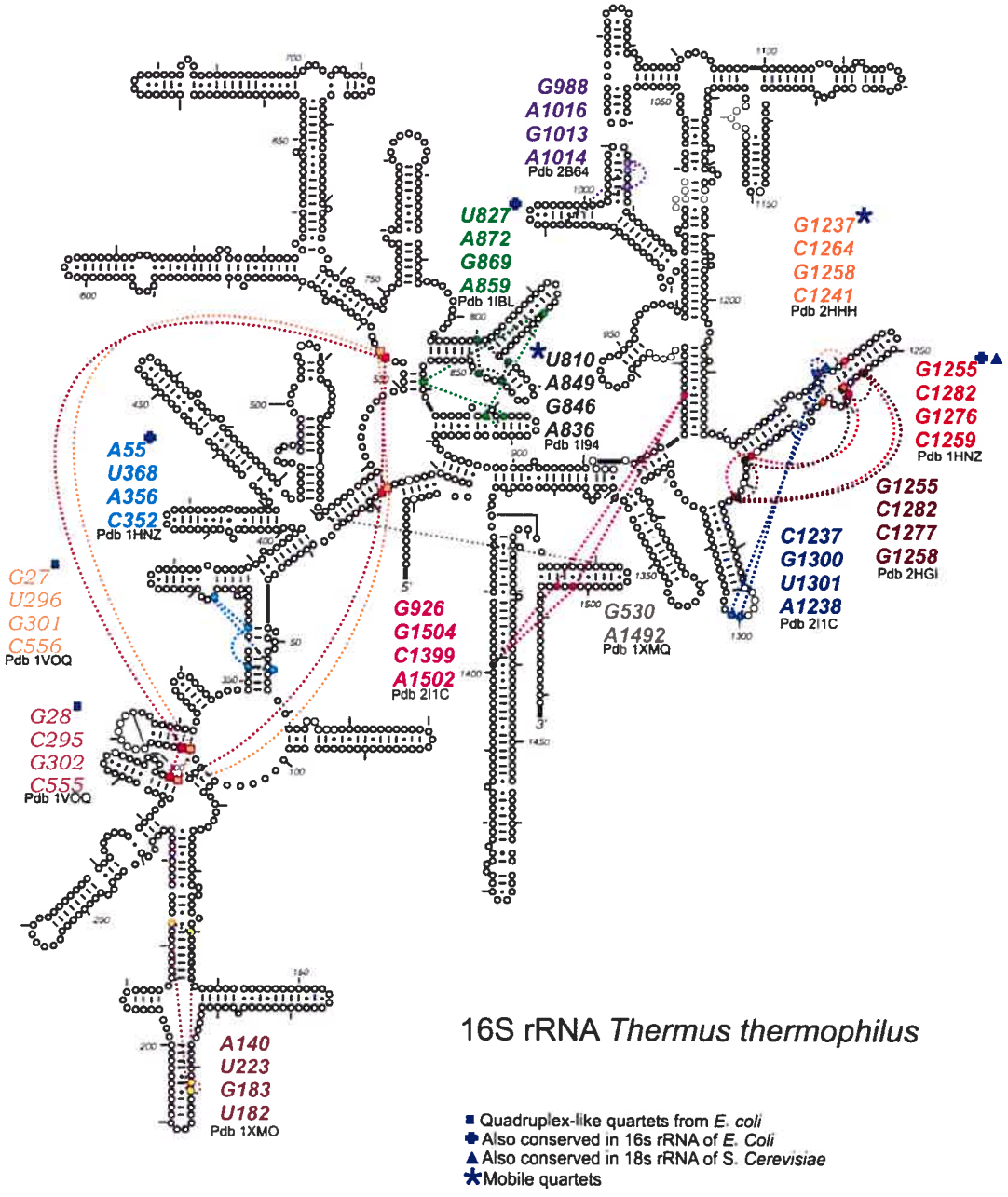


Figure 5. Bacterial 16s rRNA Alignment

Eric Westhof and colleagues have aligned 184 sequences with respect of secondary structure (Alignment is a courtesy of Westhof's laboratory). This figure shows the conservation of the G1255-C1259-G1276-C1282 quartet in bacterial 16S rRNA sequences. Conserved nucleotides are shadowed in light red. Screenshot of an alignment session made with BioEdit.

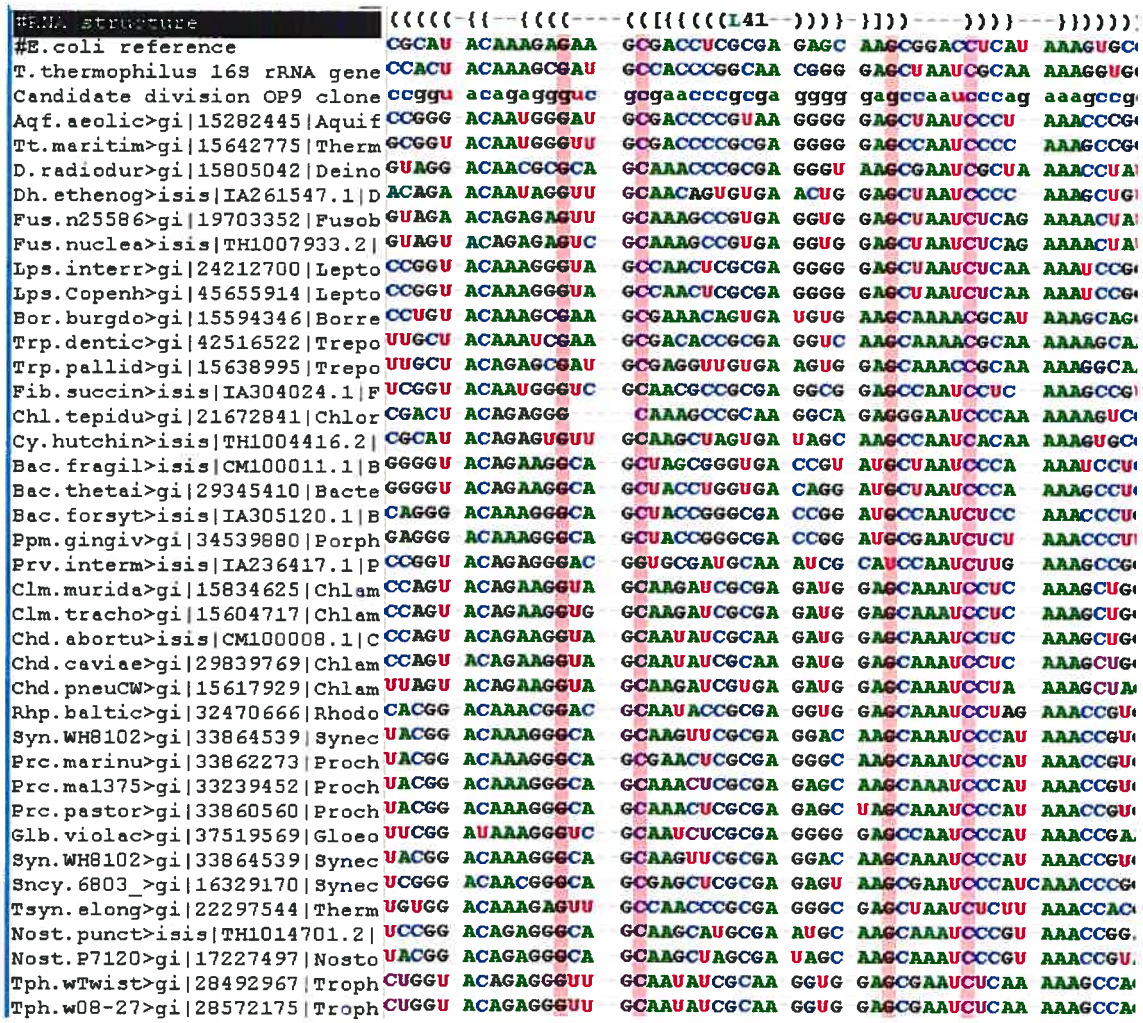


Figure 6 Mimicking Superimposition of the SARS s2m GC-quartet and the *T. thermophilus* GC-quartet.

Superimposition of GC-quartet from crystal structure of *Thermus Thermophilus* ribosomal small subunit (in ruby red- 1HNZ) with GC-quartet from SARS Coronavirus stem-loop II (in pale green – 1XJR).

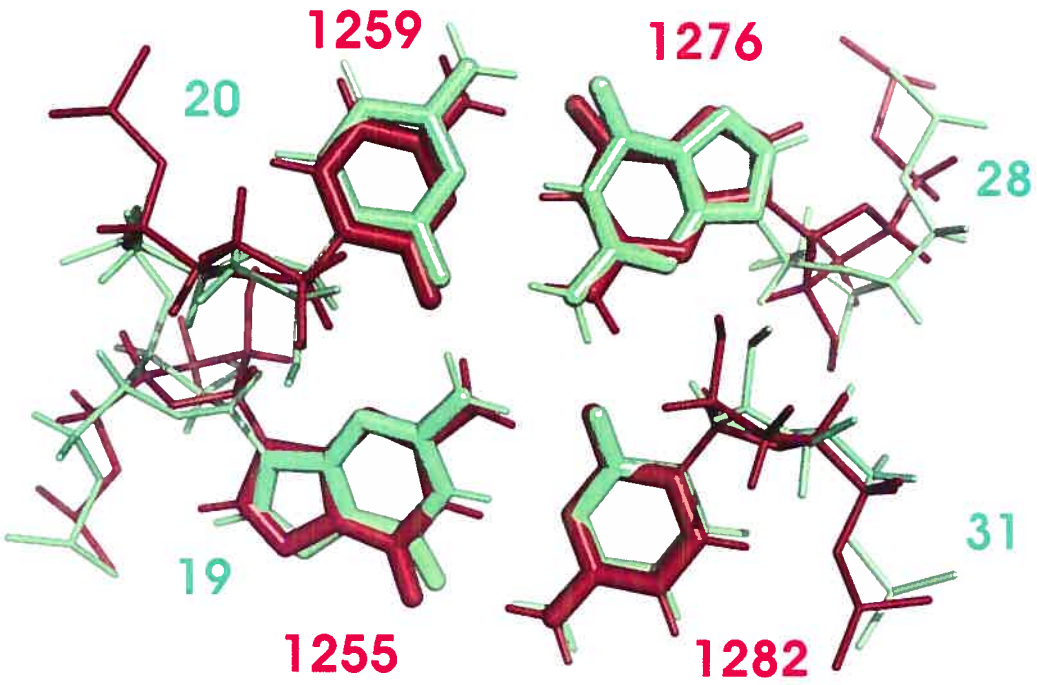


Figure 7. Superimposition of Three DNA Double-stacked Quartets and of the RNA Double-stacked Quartet

Superimposition of the RNA double-stacked quartet motif and three DNA double-quadruplex extracted from PDB files 143D, 1XAV and 2GKU. RNA quartet is in pale green, 1XAV quadruplex in red, 2KGU in blue, and 143D in yellow. A purple sphere has been added to show a hypothetical ion location.

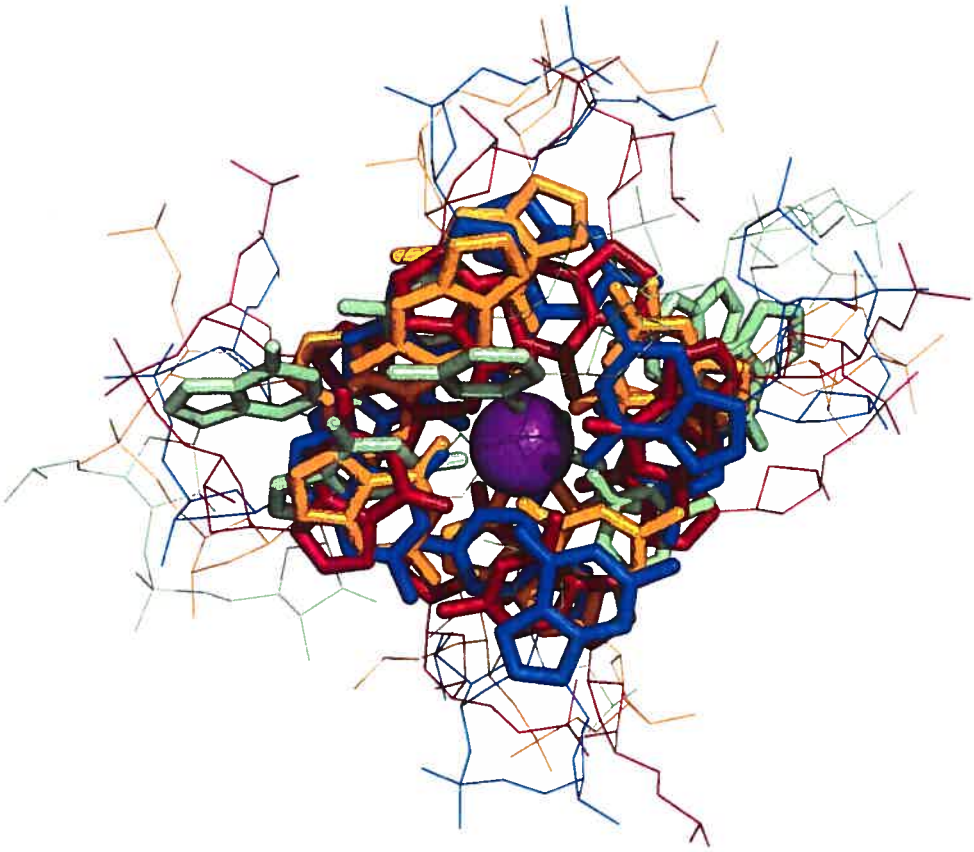
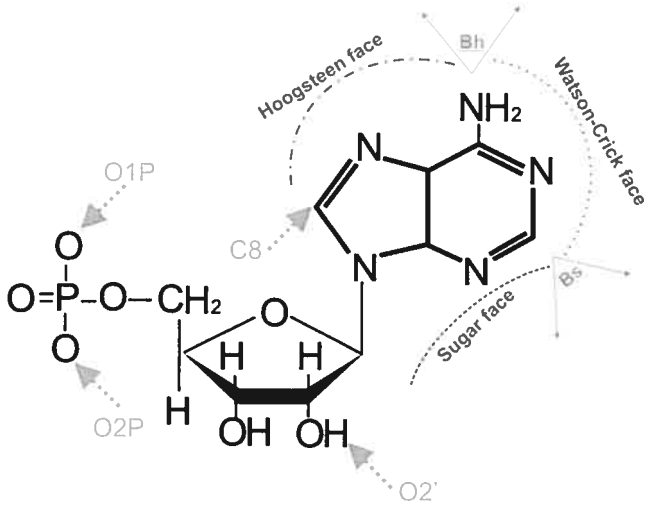


Figure 8. Base-Pairing Nomenclature

a. Faces and Backbone pairing nomenclature as implemented in MC-Annotate program (Gendron,et al. 2001) A nucleotide possesses three faces (W for Watson-Crick face, H for Hoogsteen face and S for Sugar face) and backbone emplacements (O2', O1P, O2P and C8) that can share Hydrogen atoms to create H-bonds. Bh and Bs represent "twilight zone" between Hoogsteen and Watson faces and Sugar and Watson faces respectively where H-bonds can form. B stands for Bifurcated.

b. Pairing orientation nomenclature. When the two ribose of each nucleotide involved in the pairing are in the same side of the median, the pairing is cis and when they are opposite the pairing is labelled trans.

a



b

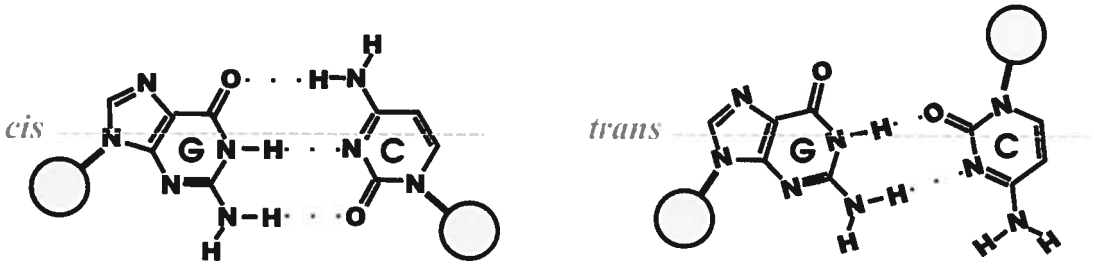
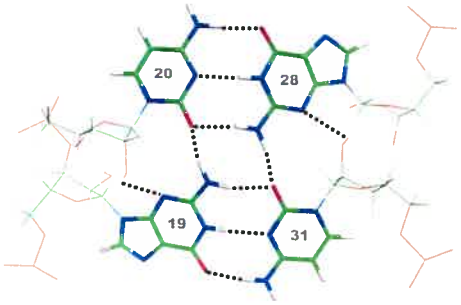


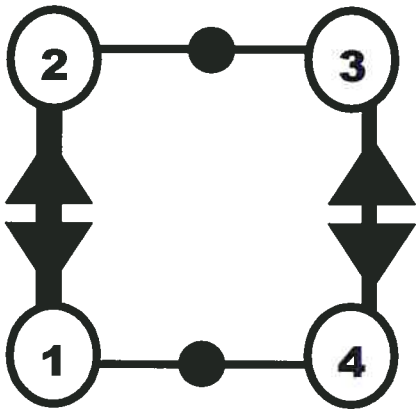
Figure 9. *MC-search* Descriptor

a. 3D representation using PyMOL of the GC-quartet instance from PDB 1XJR. Dotted lines represent h-bonds. b. Structural graph corresponding to the GC-quartet from SARS CoV using the Leontis-Westhof nomenclature. A larger line represents a backbone connection. c. *MC-Search* script used to search quartets. Where capital letters name each nucleotide described by the sequence parameter and where pairing means any type of base pair including those with backbone and not (stack) means that there is no stacking, between the two bases named before by the relation parameter. The RNA graph in b. has been generalised in the descriptor c. by searching any nucleotide (labelled N) and any type of pairing (labelled pairing).

a



b



c

```
sequence (A1 N)
sequence (B2 N)
sequence (C3 N)
sequence (D4 N)


relation
(
A1 D4 {pairing}
B2 C3 {pairing}
A1 B2 {pairing && not (stack)}
C3 D4 {pairing && not (stack)}
)
```

CHAPITRE 4 On Structural Motifs in Viral RNA

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 (colored regions) and the localization of potentially new RNA motifs (uncolored regions) in the RNA molecule **(Fig. 1).** Using a combination of two methods: annotation and motif research, we analyzed the results in all the uncolored 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 lead made on a larger dataset of RNA structure 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.

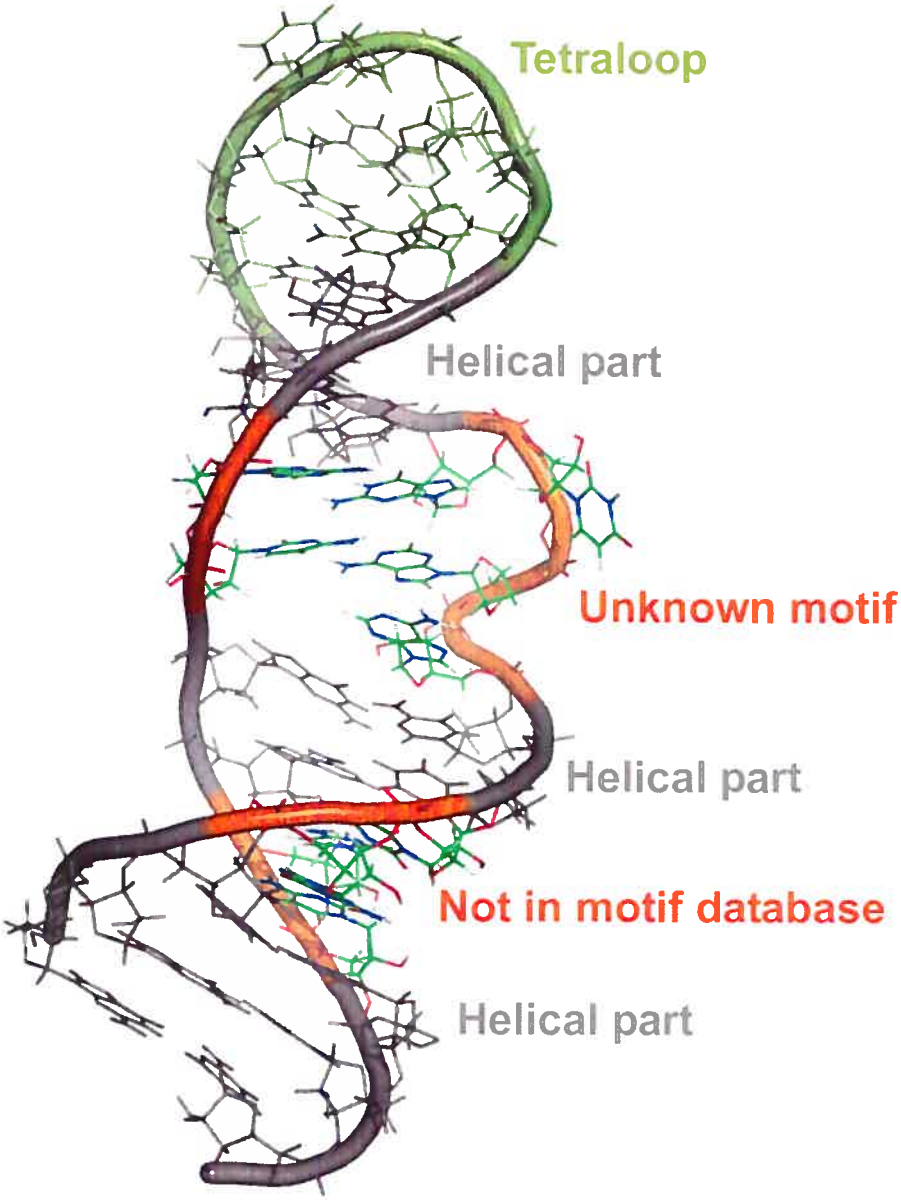


Figure 2 The Bovine Immunodeficiency Virus Trans-Activation Response RNA element (BIV TAR) in complex with TAT protein annotated with *MC-View*. The close-up shows the bulged part of BIV TAR that contains two motifs searched with *MC-Search*; the double-bulge motif and the triple G-Arg motif. In Green,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

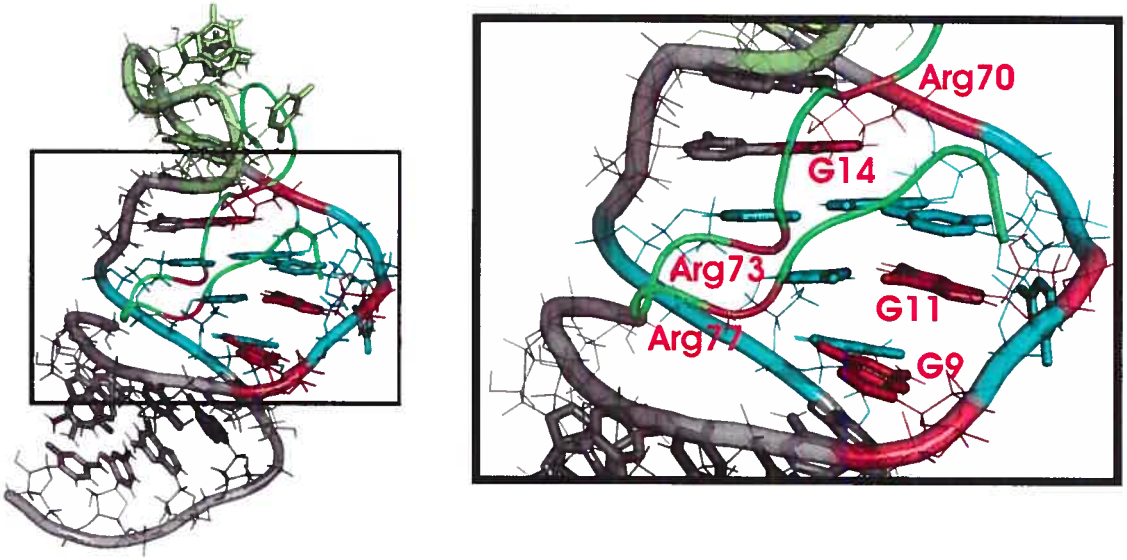


Figure 3 Hepatitis B virus pseudo-triloop (1MFY) In Blue the triloop; in cyan and blue the bulged out cytosine, in Grey helical Watson-crick basepair; in cyan, a wobble base pair

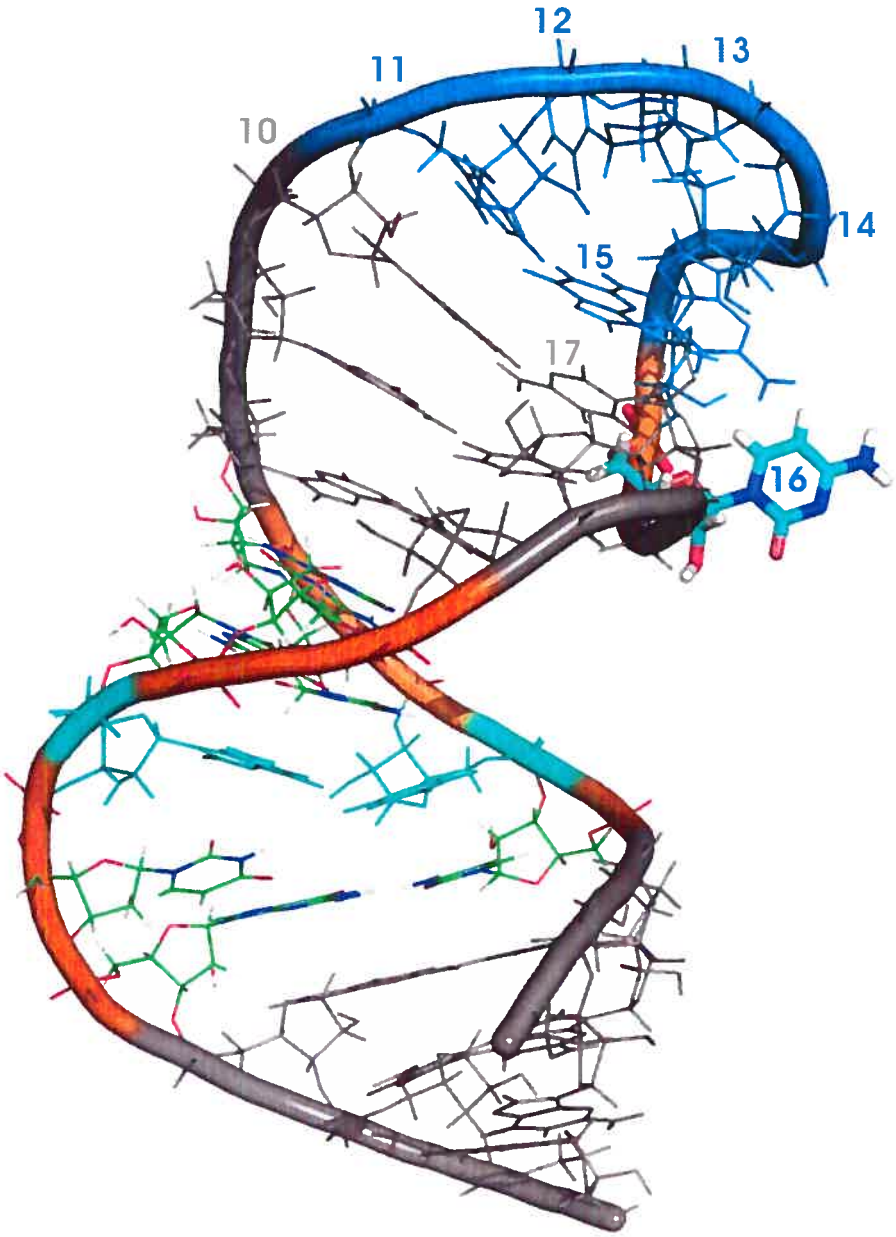


Figure 4 The K-loop motif in 2IXY. Two models of the C4 promoter influenza A are superimposed. In Green, a tetraloop; in grey, helical Watson-crick basepair; in cyan, a wobble base pair; in pink, the k-loop motif.

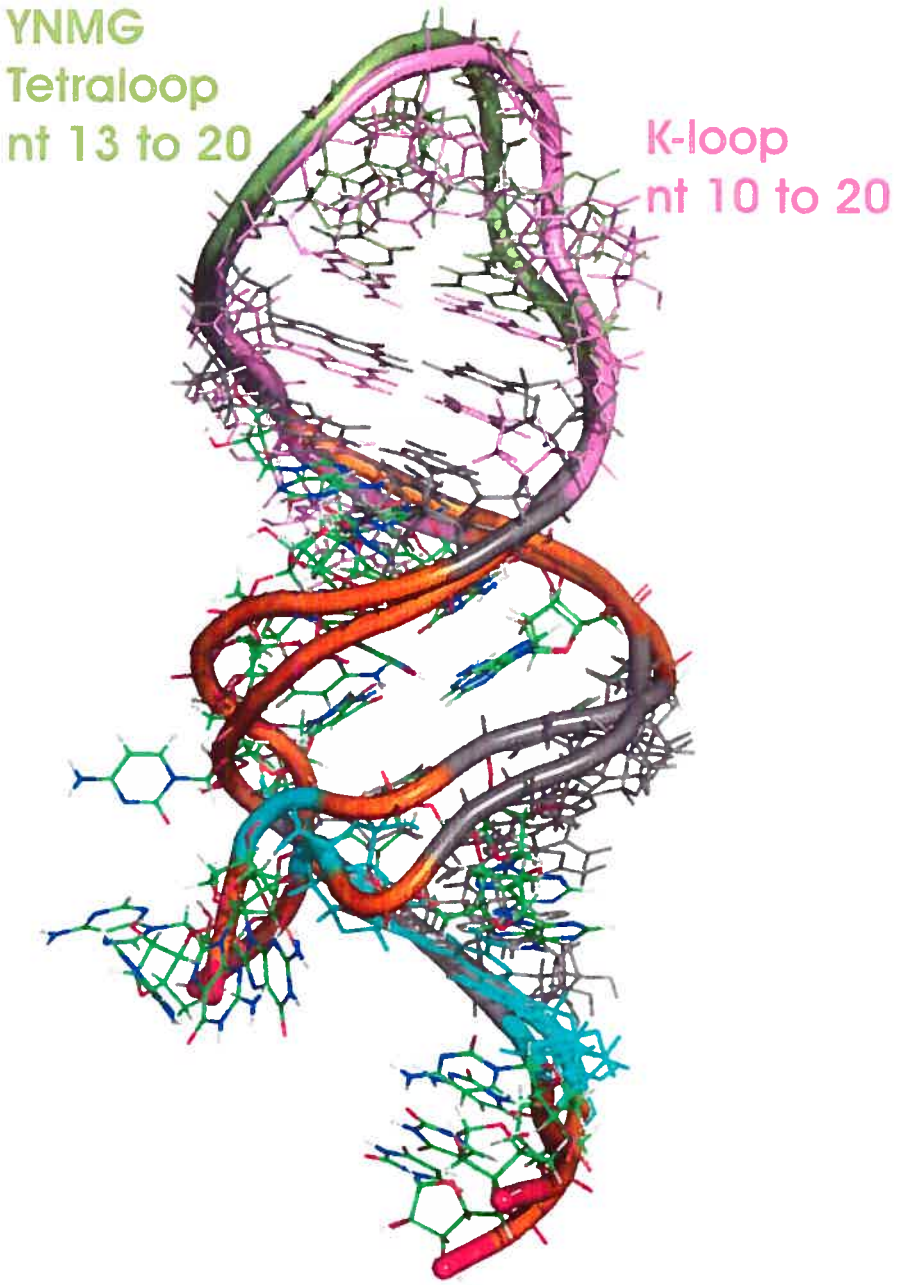


Table 1 Viral bulges frequencies. Lower case letters stand for nucleotides involved in Watson-crick or wobble base pairs. Specific means that there is no other occurrence of the motif in the PDB. Non-Specific motifs were found in other RNA molecules. H.M. is for Haloarcula marismortui; E.C. for Escherichia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

CHAPITRE 5 CONCLUSIONS

L'objectif de ma thèse était l'étude de la structure tridimensionnelle de l'ARN et plus particulièrement chez les virus. J'en ai exploré différents aspects à travers la modélisation par homologie de la tige-boucle D de quelques *Picornaviridae* qui exploite l'hypothèse de conservation de la structure tridimensionnelle de l'ARN et de ses cycles afin d'assurer la fonction (chapitre 2), et avec la recherche de motif qui montre que certains sont ubiquitaires (chapitre 3) et d'autres uniques (chapitre 4).

Les travaux présentés, qui touchent à plusieurs aspects de l'organisation des motifs au sein de la structure de l'ARN, ont permis de développer deux procédés originaux à l'étude de la structure tridimensionnelle de l'ARN. Le chapitre 2 présente la première approche de modélisation par homologie de séquences utilisant la reconstruction par cycles. Celle-ci montre qu'il est possible d'appliquer le même procédé fréquemment utilisé pour la modélisation par homologie des protéines chez les ARN. Le chapitre 4 et l'article mis en annexe relatent des travaux innovants sur l'annotation des molécules d'ARN virale avec des motifs et le développement d'une méthode liée à cette analyse. Ces deux approches pourront être appliquées à d'autres exemples et constituent d'intéressantes avancées dans le domaine de la bio-informatique structurale. La caractérisation du motif quartet apporte quant à elle plusieurs informations sur un motif mal connu des molécules d'ARN.

1.Problèmes rencontrés

A.Les données

Les données utilisées tout au long de ma thèse venaient de bases de données accessibles publiquement : PDB et *Entrez nucleotides*. Il est intéressant de noter que la Protein DataBank qui contient tous les fichiers de structures tridimensionnelles de protéines et d'acide nucléique n'a pas été élaborée dans le but d'être traitée de façon automatisée comme il a été fait dans tous mes travaux. En effet, pendant l'avancé de cette thèse plusieurs problèmes dus au traitement de l'information sont apparus. Il n'est pas anodin que des nucléotides soient insérés ou mutés par les chercheurs pour faciliter la résolution d'une structure; par exemple, une décaloop (boucle à dix nucléotides) provenant de l'ARN du spliceosome appelé U1A-RBD a été fréquemment insérée dans

plusieurs structures de ribozymes dont le HDV ribozyme (Ferre-D'Amare et al., 1998). Pour refléter les réelles données expérimentales, les fichiers au format PDB contiennent la structure exacte résolue mais ne font aucune référence au fait qu'une partie de la molécule a été changée, seule une lecture approfondie de l'article publié avec la structure peut fournir cette information. L'analyse à grande échelle des motifs d'ARN avec l'outil *MC-Search* ne permet malheureusement pas de déceler de telles choses. Une solution à ce problème serait d'améliorer la définition des structures en ajoutant la description de telles modifications dans la section REMARK des fichiers PDB prévue afin d'insérer des commentaires sur la structure tridimensionnelle contenue dans le fichier. Le développement d'un simple analyseur syntaxique de fichier PDB pourrait alors fournir les renseignements concernant les insertions de nucléotides qui ont été nécessaires à la résolution de la structure.

Un autre problème auquel les travaux ont été confrontés est la numérotation des fichiers. En effet, il n'est pas rare que cette numérotation soit différente dans le texte d'un article commentant une structure et le fichier PDB de cette même structure amenant ainsi un lot de désagréments quand à l'identification de nucléotides ou d'interactions d'intérêt.

Quand on parle de base de données, on peut se poser plusieurs questions concernant la complétude, la redondance et la fiabilité des données. La base de données PDB contient à l'heure actuelle environ 3000 fichiers contenant de l'ARN ce qui permet d'avoir à disposition une grande quantité d'information sur la structure de l'ARN. L'une des molécules les plus informative de cette base de données est l'ARNr qui contient beaucoup de motifs d'ARN et de types d'interactions non identifiées avant la résolution de sa structure en 2000. De plus, elle contient une grande diversité de molécule d'ARN. On peut donc considérer que ses structures d'ARN sont représentatives du monde ARN dans la limite de ce qui peut être résolue avec les techniques biophysiques actuelles et donc observable. Il existe 31 structures de ribosomes dans la base de données PDB, amenant une certaine redondance dans les données qui doit être considérée dans toute analyse. Afin de résoudre ce problème, une base de données contenant une structure de référence pour chaque structure d'ARN pourrait être construite. Cependant, la redondance n'est pas mauvaise car elle permet aussi l'observation des molécules dans

plusieurs états, ce qui est un atout majeur à l'étude de la relation entre structure et fonction.

A.Les limites de l'annotation

A la base de tous les travaux de cette thèse se trouve l'annotation des ARN. Cette annotation qui est décrite en détail dans l'introduction est extraite des fichiers PDB grâce au programme *MC-Annotate* (Ducharme et al., 2001). L'annotation est aussi utilisée par les programmes *MC-View, MC-Search, et MC-Sym.* Elle repose sur un traitement automatique de l'information tridimensionnelle fournie par les fichiers PDB, il est donc toujours indispensable de conserver un esprit critique sur elle et se méfier des cas limites. Par exemple, deux nucléotides peuvent être empilés et ne pas être annotés comme tel. Une inspection visuelle à l'aide de programme de visualisation comme rasmol (Sayle & Milner-White, 1995) ou PyMOL (Delano, 2002) s'avère souvent utile afin de gérer ce genre de cas.

Lors de la définition des motifs pour les fournir en argument aux différents programmes, une part de subjectivité est ajoutée. Elle correspond à l'interprétation que l'on a de la structure d'intérêt soit après son observation grâce à un logiciel de visualisation en 3D soit après avoir lu les résultats de l'annotation obtenue avec *MC-Annotate.* Afin de diminuer cette subjectivité, il est possible de décrire un même motif de plusieurs façons comme par exemple le c-motif de deux manières dans *MC-View.*

2.Perspectives

A.L'accès à la méthode de modélisation par homologie

L'approche de modélisation par homologie utilisant des cycles d'ARN pourrait être rendu accessible à la communauté en développant un serveur comme il en existe pour les protéines (Schwede et al., 2003). Tous les outils bioinformatiques utilisés pendant la thèse sont développés et mis à jour régulièrement au laboratoire et ne sont pas tous disponibles pour la communauté scientifique. Ce serveur permettrait à un utilisateur de soumettre une requête sous la forme d'une structure modèle en format PDB et d'une

séquence en format FASTA à modéliser. Une automatisation de chaque étape du processus représente un grand défi bioinformatique qu'il serait intéressant de relever.

B.L'implication du motif quartet dans la structure de l'ARN

Afin de déterminer si la présence du motif quartet est essentielle à la formation de la structure tridimensionnelle de l'ARN, il serait intéressant de faire des expériences de mutations dirigées contre les nucléotides participants aux différents quartet identifiés lors de notre étude. Les résultats indiqueraient par exemple si l'importance du motif réside dans la séquence du quartet ou bien simplement dans sa formation et donc sa propriété de stabilisation de la structure tridimensionnelle de l'ARN. On s'attendrait plutôt à la deuxième hypothèse n'ayant pas réussi à déterminer une séquence consensus de quartet lors de nos travaux. Une première piste serait d'abord de porter une attention particulière aux nucléotides du quartet présent dans le ribosome et qui est très proche en RMSD de celui de la tige-boucle 2 du virus du SRAS afin de vérifier l'hypothèse de mimique structurale.

C.La distribution des motifs

Appliquer l'approche présentée dans le chapitre 3 d'annotation des structures par un ensemble de motifs récurrent à d'autres types d'ARN pourrait permettre d'identifier de nouveaux motifs d'ARN. Utilisée sur des résultats de NMR, elle faciliterait l'observation des motifs variables dans un ensemble de structures.

Le bourgeon (ou bulge en anglais) se révèle dans le chapitre 3 comme étant un motif complexe portant souvent une signature unique lui permettant d'interagir avec les acides aminés des protéines. La dernière version du programme *MC-Search* permet de rechercher efficacement ce type d'interaction. Une étude à grande échelle des bourgeons et de leurs caractéristiques d'appariement avec les protéines avec *MC-Search* découvrirait certainement de nouvelles informations sur la structure des ARN et celle de leurs ligands.

D.Une description par les motifs

La manipulation des motifs montre qu'une molécule est divisible sous forme de cycle ou de motifs décrits selon leurs interactions. Il est donc possible d'envisager de « coder » les molécules en terme de bourgeon, boucle, hélice pour les comparer plus efficacement entre elles. Ce type de linguistique serait proche de celle développée pour les gènes par David Searls (Dong & Searls, 1994) mais appliquée à la structure tridimensionnelle de l'ARN.

GLOSSAIRE

Adjacence :

C'est la relation qui existe entre deux nucléotides consécutifs dans une séquence d'acide nucléique.

Cristallographie et rayons x :

La molécule d'ARN est mise dans des conditions particulières de température, de pression et ioniques permettant de faire croître un cristal ; ce dernier est ensuite bombardé par des rayons X. Les amplitudes de diffraction déterminent la structure tridimensionnelle.

Élément :

Sous-ensemble d'une molécule d'ARN contenant plusieurs motifs structuraux.

Empilement ou « stacking » :

L'empilement ou « stacking » est une énergie qui permet de stabiliser plusieurs nucléotides entre eux.

Interaction :

Relation entre deux nucléotides de type empilement, appariement ou adjacence.

MC-Annotate :

L'annotation d'un fichier de structure (type PDB) consiste à donner des informations sur la conformation des nucléotides et leurs interactions spatiales.

MC-View:

Voir article mis en annexe.

MC-RMSD :

Outil permettant de calculer l'écart type (root mean square deviation en anglais) entre 2 structures d'ARN, c'est-à-dire, une distance moyenne reflétant une ressemblance quantifiable entre deux structures. La méthode de Kabsch a été d'abord développée pour l'évaluation de l'écart type entre des protéines, elle a été adaptée pour l'ARN avec *MC-RMSD*.

$$RMSD=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$$

MC-Search:

Outil permettant la recherche des occurrences d'un motif dans des fichiers contenant des structures 3D déterminées par RMN ou par cristallographie. Grâce à une nomenclature définissant 2 éléments majeurs : la séquence et les relations ; il est possible de définir de façons exhaustive une structure que l'on souhaite rechercher dans un ensemble de fichiers. La nomenclature emploie de façon simple les termes utilisés usuellement lors de l'étude de la structure d'un ARN; par exemple : Stack pour le « stacking » ou bien pairing pour tous les types d'appariements.

Entrée : Description symbolique du motif (empilement, appariement, etc…) et base de données de structure.

Sortie : des fichiers pdb (coordonnées atomiques 3D) des fragments correspondant à la description.

Fonctionnement :

La base de données est annotée avec *MC-Annotate* (voir plus haut) produisant ainsi les éléments d'un graphe structurel, puis le descripteur est transformé en un graphe cible qui

sera recherché selon l'algorithme d'isomorphisme d'Ullmann. Les fragments retenus par *MC-Search* sont ceux répondant aux exigences du graphe cible. *MC-Search* nous permet donc de rechercher des motifs d'ARN et de les caractériser sous forme de classe selon leur composition ou bien leur environnement.

MC-Sym (MaCromolecular modeling by Symbolic programming):

Outil permettant de faire de modélisation 3D de molécules d'ARN.

Il utilise des cordonnées et des relations entre résidus extraites de structures (cristallographies, RMN et modèles théoriques).

Il répond à un problème de satisfaction de contrainte (CSP), c'est-à-dire que l'on contraint la structure afin d'obtenir un modèle valide. Pour résoudre le problème de satisfaction de contrainte, un algorithme combinatoire de retour-arrière (« backtrack ») explore un espace de conformations de nucléotides et de relations (appariement et empilement).

Chaque nucléotide est placé en utilisant l'ordre contraint par le descripteur (ordre de « backtrack ») et une relation avec un nucléotide précédent.

Modèle :

Représentation abstraite permettant de synthétiser un ensemble d'informations sur un système étudié. En biologie structurale, un modèle est une représentation de la structure satisfaisant les informations connues sur la molécule étudiée.

Modélisation :

Principe qui consiste à construire un ou plusieurs modèles d'une molécule correspondant à un ensemble d'informations.

Moment angulaire d'un atome :

Caractérise le comportement de l'atome sous l'effet de la rotation dans l'espace.

Moment magnétique nucléaire :

Caractérise la capacité d'un noyau à produire des interactions magnétiques avec son environnement.

Relation :

Voir Interaction.

Résonance magnétique nucléaire (RMN) :

La molécule d'ARN est mise en solution puis soumise à un champ magnétique. Des protons entrant en vibration, on déduit un spectre de résonance dont on peut dériver des contraintes de distance et d'angle de torsion.

Spin :

moment angulaire d'une particule.

Système de modélisation :

Programme ou ensemble de programmes permettant de passer d'une liste d'informations structurales à un ou plusieurs modèles 3D correspondants à un ensemble d'informations.

BIBLIOGRAPHIE

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

Basavappa R, Sigler PB. 1991. The 3 A crystal structure of yeast initiator tRNA: functional implications in initiator/elongator discrimination. *Embo J 10*:3105-3111.

Berman H, Henrick K, Nakamura H, Markley JL. 2007. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res 35*:D301-303.

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

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

Blum B, Bakalara N, Simpson L. 1990. A model for RNA editing in kinetoplastid mitochondria: "guide" RNA molecules transcribed from maxicircle DNA provide the edited information. *Cell 60*:189-198.

Buchen-Osmond C. 2003. The universal virus database ICTVdB. *Computing in Science & Engineering [see also IEEE Computational Science and Engineering] 5*:16-25.

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

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

Dong S, Searls DB. 1994. Gene structure prediction by linguistic methods. *23*:540-551.

Ducharme F, LeVesque L, Gendron L, Legault A. 2001. Development process and qualitative evaluation of a program to promote the mental health of family caregivers. *Clin Nurs Res 10*:182-201.

Felden B, Himeno H, Muto A, Atkins JF, Gesteland RF. 1996. Structural organization of Escherichia coli tmRNA. *Biochimie 78*:979-983.

Ferre-D'Amare AR, Zhou K, Doudna JA. 1998. Crystal structure of a hepatitis delta virus ribozyme. *Nature 395*:567-574.

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

Forster AC, Symons RH. 1987. Self-cleavage of virusoid RNA is performed by the proposed 55-nucleotide active site. *Cell 50*:9-16.

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

Gottesman S. 2004. The Small RNA Regulators of Escherichia coli: Roles and Mechanisms. *Annual Review of Microbiology 58*:303-328.

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. 2003. Rfam: an RNA family database. *Nucleic Acids Res 31*:439-441.

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. 2005. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res 33*:D121-124.

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.

Guignon V, Chauve C, Hamel S. 2005. An edit distance between RNA stem-loops.335-347.

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

Hermann T, Patel DJ. 2000. RNA bulges as architectural and recognition motifs. *Structure 8*:R47-54.

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.

Kabsch H. 1978. A discussion of the solution for the best rotation to relate two sets of vectors. *ACTA Cryst Sec A: Cryst Phys Diff Theo Gen Cryst 34A*:827-828.

Kazantsev AV, Krivenko AA, Harrington DJ, Holbrook SR, Adams PD, Pace NR. 2005. Crystal structure of a bacterial ribonuclease P RNA. *Proc Natl Acad Sci U S A 102*:13392-13397.

Kim SH, Rich A. 1968. Single crystals of transfer RNA: an x-ray diffraction study. *Science 162*:1381-1384.

Kim SH, Suddath FL, Quigley GJ, McPherson A, Sussman JL, Wang AH, Seeman NC, Rich A. 1974. Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science 185*:435-440.

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

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

Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. 1982. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell 31*:147-157.

Laferriere A, Gautheret D, Cedergren R. 1994. An RNA pattern matching program with enhanced performance and portability. *Comput Appl Biosci 10*:211-212.

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

Lee RC, Feinbaum RL, Ambros V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell 75*:843-854.

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 NB, Altman RB, Berman HM, Brenner SE, Brown JW, Engelke DR, Harvey SC, Holbrook SR, Jossinet F, Lewis SE, Major F, Mathews DH, Richardson JS, Williamson JR, Westhof E. 2006. The RNA Ontology Consortium: an open invitation to the RNA community. *Rna 12*:533-541.

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

Limbach PA, Crain PF, McCloskey JA. 1994. Summary: the modified nucleosides of RNA. *Nucleic Acids Res 22*:2183-2196.

Macke T, Case DA. 1998. Modeling unusual nucleic acid structures. *Molecular Modeling of Nucleic Acids*:379–393.

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

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

Major F, Gautheret D, Cedergren R. 1993. Reproducing the three-dimensional structure of a tRNA molecule from structural constraints. *Proc Natl Acad Sci U S A 90*:9408-9412.

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.

Markham NR, Zuker M. 2005. DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res 33*:W577-581.

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

Massire C, Westhof E. 1998. MANIP: an interactive tool for modelling RNA. *J Mol Graph Model 16*:197-205, 255-197.

Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol 288*:911-940.

Moss GP. 1996. Basic terminology of stereochemistry. *68*:2193–2222.

Mount DW. 2004. *Bioinformatics: Sequence and Genome Analysis*: CSHL Press.

Mueller F, Sommer I, Baranov P, Matadeen R, Stoldt M, Wohnert J, Gorlach M, van Heel M, Brimacombe R. 2000. The 3D arrangement of the 23 S and 5 S rRNA in the Escherichia coli 50 S ribosomal subunit based on a cryo-electron microscopic reconstruction at 7.5 A resolution. *J Mol Biol 298*:35-59.

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.

Pallansch MA, Roos RP. 2001. Enteroviruses: polioviruses, coxsackieviruses, echoviruses, and newer enteroviruses. *Fields Virology 4*:723-775.

Robert KZT, Harvey SC. 1993. Yammp: development of a molecular mechanics program using the modular programming method. *Journal of Computational Chemistry 14*:455-470.

Sayle RA, Milner-White EJ. 1995. RASMOL: biomolecular graphics for all. *Trends Biochem Sci 20*:374.

Schuler M, Connell SR, Lescoute A, Giesebrecht J, Dabrowski M, Schroeer B, Mielke T, Penczek PA, Westhof E, Spahn CM. 2006. Structure of the ribosome-bound cricket paralysis virus IRES RNA. *Nat Struct Mol Biol 13*:1092-1096.

Schwede T, Kopp J, Guex N, Peitsch MC. 2003. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res 31*:3381-3385.

Thomas B, Akoulitchev AV. 2006. Mass spectrometry of RNA. *Trends Biochem Sci 31*:173-181.

Watson JD, Crick FH. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature 171*:737-738.

Williams KP, Bartel DP. 1996. Phylogenetic analysis of tmRNA secondary structure. *Rna 2*:1306-1310.

Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JH, Noller HF. 2001. Crystal structure of the ribosome at 5.5 A resolution. *Science 292*:883-896.

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

Article annexe. *MC-View*: An online tool for RNA motif visualization

Dans ce travail, j'ai participé à l'élaboration du projet, à une partie du développement informatique et à la description d'une grande partie des motifs de la base de données de motifs utilisés pour annoter les structures d'ARN.

Structural bioinformatics

MC-View : An online tool for RNA motif visualization

Louis-Philippe Lavoie[1], Emmanuelle Permal[1], Véronique Lisi[1], Caroline Louis-Jeune[1] and François Major[1,*]

[1]Theoretical and Computer Science Biology Lab, Institute for Research in Immunology and Cancer, Université de Montréal, Montréal, QC, H3C 3J7, Canada

ABSTRACT

Summary: *MC-View* is an online tool that allows for searching and visualization of structural motifs in submitted RNA structures. It uses predefined structural motifs such as GNRA tetraloop or sarcin-ricin motif and searches for them 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://bioinfo.iric.ca/mcview.

Contact: [redacted]

*To whom correspondence should be addressed.

Introduction

RNA structural motifs are the building blocks of RNA molecules (Hendrix et al., 2005). A large number has been found and 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, visualization is still one of the simplest and yet most effective recourse 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, the input molecule is searched for all instances of the selected motifs. The result is a fully annotated secondary structure diagram and a PyMOL script overlaying the initial three-dimensional structure with the studied motifs. Putative protein-RNA hydrogen bonds (h-bonds) are also identified.

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 results.

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 cosmetic customizations of the output. A complete description of the syntax used to describe the search motifs is available at http://www-lbit.iro.umontreal.ca/wiki/index.php/MC-Search

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. 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.

Example

Figure 1 shows the 5S rRNA fragment of *Haloarcula Marismortui* and surrounding proteins (PDB 1S72) displayed with all available predefined motifs existing in this molecule.

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.

Future developments

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

Acknowledgments

We thank Romain rivière for his contribution with pdb2pdf tool. 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 education 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 secondary 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. 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. *Nucleic acids research 31*:3450-3460.

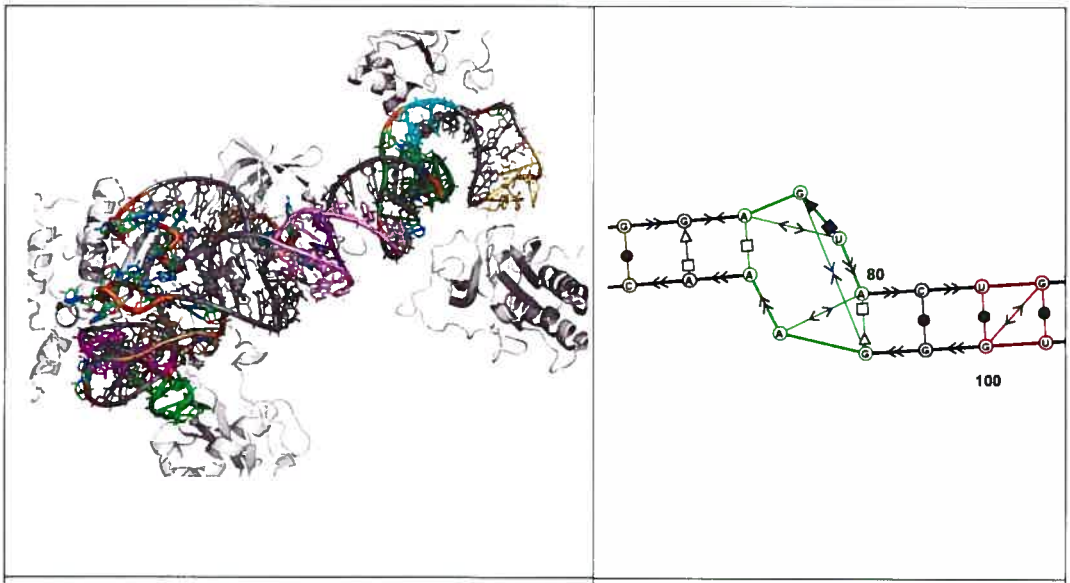


Fig. 1. The results of MC-View for the 5S rRNA of Haloracula Marismortui. The molecule is annotated by the motifs found by MC-View. (A) Full structure as displayed by PyMOL. (B) Displayed as a PDF file of the secondary structure : close-up on the sarcin-ricin motif (Green in both images (A) and (B)).