THESE

Présentée à l'U.F.R. des Sciences de la Vie et de la Terre de l'Université Louis Pasteur de Strasbourg

pour obtenir le grade de

Docteur de l'Université Louis Pasteur de Strasbourg

par

Aurélie LESCOUTE-PHILIPPS

Extraction de contraintes structurales à partir de comparaisons de séquences et de structures tridimensionnelles d'ARN

Soutenue le 24 Février 2006 devant le jury composé de :

Mme Christine GASPIN	Rapporteur externe
M. Brice FELDEN	Rapporteur externe
Mme Anne-Catherine DOCK-BREGEON	Rapporteur interne
M. Alain KROL	Examinateur
M. Eric WESTHOF	Directeur de thèse

THESE

Présentée à l'U.F.R. des Sciences de la Vie et de la Terre de l'Université Louis Pasteur de Strasbourg

pour obtenir le grade de

Docteur de l'Université Louis Pasteur de Strasbourg

par

Aurélie LESCOUTE-PHILIPPS

Extraction de contraintes structurales à partir de comparaisons de séquences et de structures tridimensionnelles d'ARN

Soutenue le 24 Février 2006 devant le jury composé de :

Mme Christine GASPIN	Rapporteur externe
M. Brice FELDEN	Rapporteur externe
Mme Anne-Catherine DOCK-BREGEON	Rapporteur interne
M. Alain KROL	Examinateur
M. Eric WESTHOF	Directeur de thèse

Remerciements

Tout d'abord je souhaite exprimer ma profonde gratitude à mon directeur de thèse le professeur Eric Westhof. Durant ces 5 années passées dans son laboratoire et malgré les tâches nombreuses qui lui incombent, il a toujours été à l'écoute. Il a su également par son enthousiasme et sa patience me faire découvrir et apprécier les problématiques que cache la structure des ARN.

Je remercie le professeur Bernard Ehresmann de m'avoir acceuillie déjà en stage de maîtrise puis en DEA et en doctorat dans l'unité UPR9002 qu'il a dirigée jusqu'en janvier 2005.

J'exprime ma respectueuse reconnaissance à Mesdames les Docteurs Christine Gaspin et Anne-Catherine Dock-Bregeon, Monsieur le Docteur Alain Krol et Monsieur le professeur Brice Felden pour l'honneur qu'ils me font de juger cette thèse.

Je remercie les collaborateurs avec qui j'ai eu la chance de travailler et plus particulièrement Neocles Leontis.

Un grand merci aux membres de l'unité 9002 qui par leur expérience m'ont aidé au cours de ce travail.

Je remercie tous les membres du laboratoire Westhof ; j'ai appris beaucoup à leur côté. Je remercie très chaleureusement Rym et Pascal pour leur soutien.

Merci aux collègues-du-labo-d'à-côté pour les discussions parfois (souvent ?) pas scientifiques ...

Merci à mes chers amis-du-midi Aurélie, Boris, Michel et Simon. Nous nous sommes soutenus en ces temps compliqués et nos moments m'accompagneront longtemps.

Enfin, je remercie Michael, mon complice.

SOMMAIRE

1. INTRODUCTION GENERALE

L'ARN joue des rôles biologiques divers et nombreux : support de l'information génétique (ARNm, ARN génomique viral...), catalyseur de réaction chimique (ribozyme à tête de marteau, RNase P, introns, ribosome ...) ou encore régulateur de l'expression génétique ('riboswitch' ou interrupteur moléculaire). La fonction biologique des ARN est déterminée de manière générale par leur séquence, et, pour beaucoup d'ARN non codants, par leur structure tridimensionnelle. Les différences d'architecture entre ARN et protéine sont nombreuses. Il y a quatre types de monomères (les nucléotides), six angles de torsion du squelette au lieu de deux dans les protéines et la structure de l'ARN n'est pas organisée autour d'un cœur hydrophobe comme la majorité des protéines. Le repliement est dicté essentiellement par deux principes tout d'abord mis en évidence dans les structures de double hélices d'ADN et d'ARN : liaisons hydrogène entre bases et empilement des bases. Il apparaît donc essentiel de comprendre comment la structure primaire de l'ARN, une chaîne primaire polyanionique de nucléotides, détermine sa structure fonctionnelle. Dans ce travail de thèse nous n'avons pas abordé ce thème sous l'angle de la cinétique du repliement, mais en considérant l'aspect modulaire et hiérarchique de formation des différents motifs secondaires et tertiaires composant un ARN structuré fonctionnel.

1.1. Repliement hiérarchique et modulaire de l'ARN

De nos jours, suite à de nombreuses expériences, il est admis que le repliement de l'ARN est modulaire et hiérarchisé (Westhof & Michel, 1994a). Les ARN structurés sont composés d'éléments structuraux récurrents et ubiquitaires. La modulation de l'assemblage de ces différentes « briques », mène à une multitude de structure d'ARN. Ces molécules d'ARN se replient séquentiellement de manière plus au moins concomitante à leur synthèse. Les hélices composées d'appariements Watson-Crick se forment d'abord, suivies de la structuration des régions en simple brin. Finalement, les hélices s'organisent dans l'espace en domaines pour constituer la structure tridimensionnelle qui sera maintenue par des interactions tertiaires, appelées encore interactions à longue distance ou

interactions ARN-ARN. Les ions monovalents et divalents jouent un rôle important dans la formation de la structure tridimensionnelle.

On distingue donc, pour une molécule d'ARN structuré, la structure primaire qui est la séquence nucléotidique orientée de l'extrémité 5' vers l'extrémité 3', la structure secondaire et la structure tertiaire. Ces deux dernières seront décrites ci-dessous.

La structure secondaire d'un ARN est formée à 60-70% par l'appariement des bases en paires Watson-Crick (ou canoniques), qui, par empilement, forment une double hélice de type A. Les nucléotides non impliqués dans des doubles hélices appartiennent à des régions en simple brin qui entrent et sortent des hélices ou qui joignent les hélices dans des jonctions à plusieurs hélices ou encore forment des boucles terminales, des bulles internes symétriques ou asymétriques. Les régions en simple brin et les hélices sont appelées éléments de structure secondaire ou encore modules élémentaires bidimensionnels. La représentation schématique de la structure secondaire d'un ARN est constituée d'un assemblage plan d'hélices formées de paires Watson-Crick et liées par des lignes non sécantes (Figure 1).

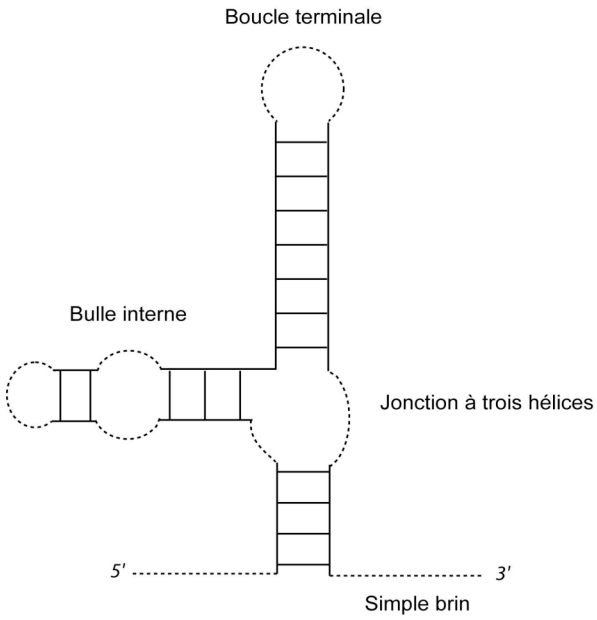


Figure 1 : Représentation schématique de la structure secondaire d'un ARN. Seules les paires Watson-Crick sont représentées (lignes pleines). Les nucléotides non impliqués dans ces paires (en pointillés) appartiennent aux boucles terminales, aux bulles internes (symétriques ou asymétriques), aux brins qui lient les différentes hélices d'une jonction ou aux simples brins.

La structure tridimensionnelle implique l'organisation dans l'espace des modules élémentaires bidimensionnels c'est-à-dire l'empilement des hélices et l'assemblage des éléments de structure secondaire, puis le positionnement des motifs architecturaux les uns par rapport aux autres, et enfin, la formation d'interactions tertiaires à plus ou moins longues distances entre nucléotides (Westhof & Michel, 1994b). Cette structuration implique la formation de paires de bases non Watson-Crick jouant un rôle fondamental dans l'architecture de l'ARN.

1.2. Nomenclature et classification des appariements

Ces cinq dernières années ont vu les banques de données de structures ARN augmenter considérablement grâce à la résolution au niveau atomique de structures de très grands ARN comme les ARN ribosomaux (Berman et al., 1992; Berman et al., 2000; Batey et al., 2004; Holbrook, 2005). La complexité et la diversité des interactions qui composent ces structures sont telles qu'il est souvent difficile de les visualiser et par conséquent de les comprendre. Toutefois, des études ont révélé la présence de motifs récurrents (Moore, 1999; Leontis & Westhof, 2003; Holbrook, 2005; Noller, 2005). Le plus fondamental, qui détermine la première étape de repliement de la chaîne polynucléotidique, est l'hélice Watson-Crick composée d'appariements Watson-Crick ou canoniques. Les deux bases impliquées dans une paire Watson-Crick, interagissent par leur côté Watson-Crick en formant des liaisons hydrogène et ont leur liaison glycosidique orientée en *cis* par rapport à l'axe des liaisons hydrogène.

D'autres paires de bases, dites non Watson-Crick ou non canoniques, ont été mises en évidence dès les années 80 dans la structure des ARNt où elles permettent à des segments éloignés d'interagir (boucle T et boucle D) (Robertus et al., 1974; Rich & Kim, 1978; Moras et al., 1980). Ces paires non canoniques interviennent également dans la structuration des motifs tridimensionnels c'est-à-dire entre deux segments en simple brin ou un segment en simple brin et une hélice ou des boucles internes ou terminales. Elles structurent les motifs récurrents tels que le tournant U (Quigley & Rich, 1976), la boucle E (Varani et al., 1989; Wimberly et al., 1993), la boucle sarcine/ricine (Szewczak et al., 1993), le motif tournant-K (Klein et al., 2001) ou le motif C (Leontis & Westhof,

2003). Les boucles terminales aux extrémités d'hélices contiennent la plupart du temps une ou plusieurs paires de bases non Watson-Crick à l'interface entre l'hélice et la boucle ; par exemple, la tetra-boucle GNRA (Woese et al., 1990; Jucker & Pardi, 1995), le motif boucle-T (Krasilnikov & Mondragon, 2003) (Nagaswamy & Fox, 2002) ou la boucle anticodon (Auffinger et al., 1999). Elles sont présentes également aux jonctions à plusieurs hélices où elles modulent l'empilement des hélices. Enfin, elles interviennent de facon très importante dans les interactions à longue distance en contact boucle-boucle comme dans le domaine S de SRP ou en contact bulle interne hélice comme le motif en A mineur (Nissen et al., 2001). Ainsi les paires non Watson-Crick (non canoniques) jouent un rôle fondamental dans le repliement de l'ARN (Westhof & Fritsch, 2000).

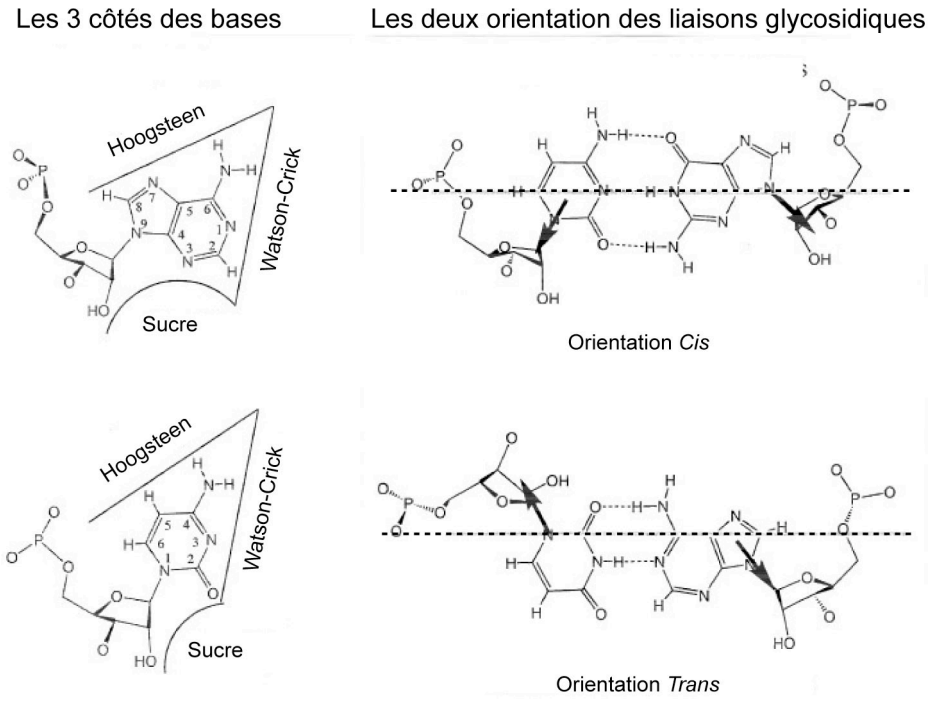


Figure 2 : Identification des côtés des bases et orientation *cis* ou *trans* des liaisons glycosidiques au sein d'un appariement. D'après Leontis et Westhof (2001)

Afin d'exploiter au mieux la quantité phénoménale d'information tridimensionnelle disponible, Leontis et Westhof ont proposé une nomenclature et une classification des paires de bases présentes dans la base de données de structures tridimensionnelle (Leontis & Westhof, 2001). Elle prend en compte les côtés des deux bases impliquées dans l'interaction et la position de la liaison

glycosidique (*cis* (noir) ou *trans* (blanc)) par rapport à la ligne médiane des liaisons hydrogène formées (Figure 3). Chacune des 4 bases, A, U, G et C comporte trois côtés : Watson-Crick (o), Hoogsteen (□) et Sucre (∆) chacune représentée par un symbole (Figures 3 et 4). Une base interagit avec une autre par l'un de ses trois côtés Watson-Crick, Hoogsteen ou Sucre (qui inclut le 2'OH), et les liaisons glycosidiques d'une paire peuvent être orientées en *trans* ou en *cis* pour chaque interaction. Par conséquent, l'ensemble des combinaisons d'appariements mène à 12 familles géométriques possibles (2 bases x3 côtés x2 orientations=12). Chacune des familles géométriques est représentée par une paire de symboles indiquant les côtés impliqués et l'orientation des liaisons glycosidiques (Figure 3). Un appariement impliquant les côtés Sucre de deux nucléotides est représenté par un triangle. La base du triangle fait face au nucléotide qui "donne" son 2'OH et l'extrémité du triangle pointe vers le nucléotide qui le reçoit.

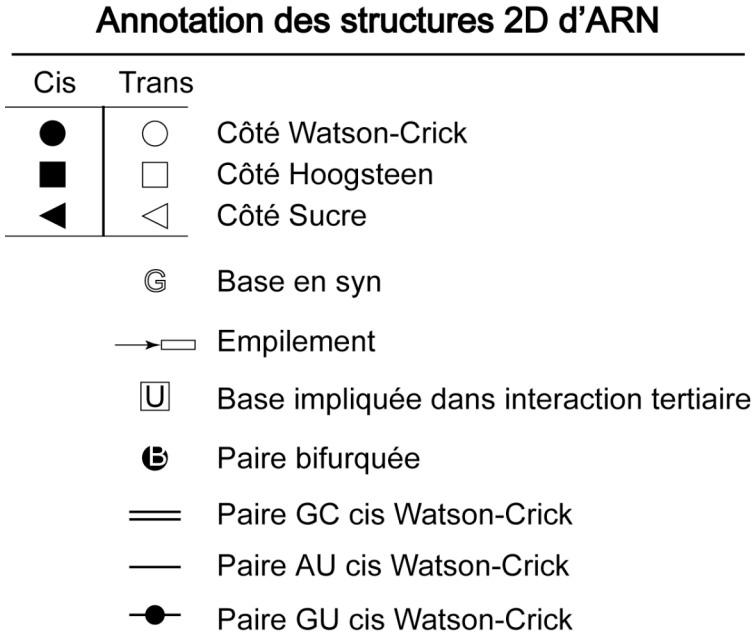


Figure 3 : Nomenclature de Leontis et Westhof. Symboles proposés pour représenter les interactions tertiaires et les caractéristiques structurales d'une structure d'ARN sur un schéma en deux dimensions (Leontis & Westhof, 2001).

Cette nomenclature se révéle d'une grande utilité. Elle permet, tout d'abord, de résumer toute l'information de structure tridimensionnelle d'une molécule d'ARN sur un schéma bidimensionnel. Quelques représentations bidimensionnelles de structures cristallographiques utilisant cette nomenclature

seront présentées dans le chapitre 2.1.3. D'autre part, cette nomenclature permet l'identififcation de motifs structuraux récurrents dans une nouvelle structure cristallographique (Figure 4). Enfin, elle est nécessaire à la prédiction de motifs par analyse comparative de séquences dans des alignements structuraux de séquences homologues d'ARN et permet également leur affinement comme nous le verrons dans le chapitre 2.2.

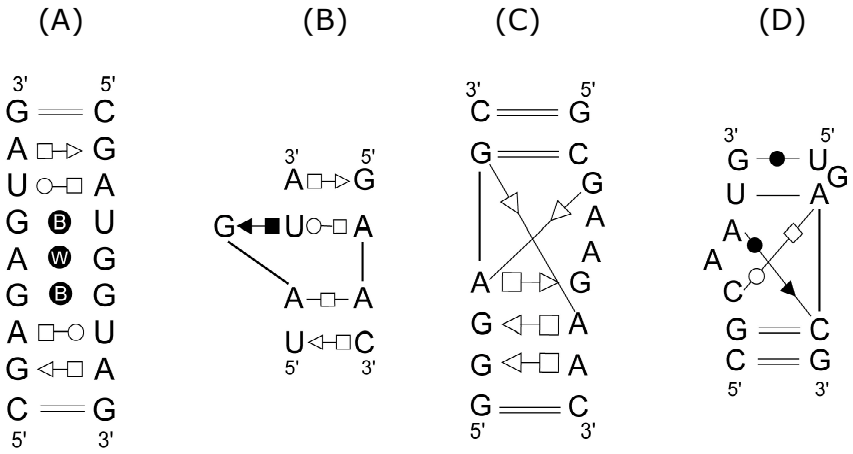


Figure 4 : Annotations à l'aide la nomenclature géométrique de quelques motifs récurrents d'ARN. Les cercles indiquent les côtés Watson-Crick, les carrés les côtés Hoogsteen et les triangles les côtés Sucre. Les symboles pleins/vides signifient que les bases qui interagissent, sont orientées en *cis*/*trans*. **(A) Double boucle E de l'ARNr 5S**. Le « B » dans un cercle indique une paire bifurquée et le « W » une interaction médiée par une molécule d'eau. Ce motif est également présent dans les ARNr 16S et 23S. **(B) Motif S**. Ce motif est retrouvé dans la boucle sarcine-ricine de l'ARNr bactérien 23S. Une G en plateforme interagit avec une U formant ainsi une interaction triple. **(C) Tournant K**. **(D) Motif C**.

1.3. Les paires de bases isostériques et les matrices d'isostérie

La détermination de la structure secondaire est une étape essentielle dans l'étude des relations structure-fonction d'un ARN. Elle peut se faire de deux manières. Si une seule séquence est connue, la structure secondaire peut être déterminée en combinant les résultats proposés par un logiciel de repliement par minimisation d'energie (Zuker, 2003) et les résultats de sondes enzymatiques et

chimiques (Ehresmann et al., 1987). Si plusieurs séquences homologues sont connues, les paires Watson-Crick, composant la structure secondaire de l'ARN, peuvent être mises en évidence par analyse comparative de séquences. L'analyse comparative de séquence exploite le fait que les structures tridimensionnelles de molécules d'ARN homologues, au cours de l'évolution, changent moins vite que leurs séquences. Le maintien des appariements Watson-Crick et des appariements non Watson-Crick de la structure est assuré par des changements compensatoires (ou covariations). Ainsi, une paire de base peut être mutée en une autre paire isostérique c'est-à-dire géométriquement similaire. Deux paires de bases isostériques appartiennent à la même famille géométrique et présentent des distances C1'-C1' proches. Sur la figure 5 sont représentés trois appariements de la famille *trans* Hoogsteen/Sucre. Une double hélice d'ARN est régulière car les appariements qui la composent (A-U, U-A, G=C, C=G) sont isostériques. La paire wobble GoU, bien que non isostérique avec les paires canoniques citées précédemment, est compatible car elle perturbe peu la régularité de l'hélice. L'analyse comparative de séquence est basée sur l'isostérie des appariements. A une même position dans différentes séquences, la géométrie d'un appariement doit être conservée et les mutations observées doivent obéir à la règle d'isostérie. Ainsi, dans un alignement dit structural, c'est l'information d'appariement qui est prise en compte et non la nature des bases. C'est en partie grâce au modèle de Michel et Westhof (1990) qu'il a été possible d'étendre l'analyse comparative de séquences aux régions dites périphériques et ainsi de démontrer certaines interactions essentielles au repliement tertiaire des ribozymes de groupe I (Jaeger et al., 1991; Michel et al., 2000).

Comme nous l'avons vu dans la figure 5, au sein d'une famille géométrique on distingue une ou plusieurs sous-familles d'isostérie chacune caractérisée par une distance C1'-C1' spécifique. Les paires appartenant à une sous-famille d'isostérie (numérotée I1, I2, etc) sont dites isostériques ; elles peuvent se substituer les unes aux autres sans que la structure tridimensionnelle de la molécule ne soit perturbée. A chacune des 12 familles géométriques, correspond une matrice d'isostérie (Figure 6). Une matrice d'isostérie est un tableau à double entrée ; chaque entrée comporte les 4 bases A, C, G, U et à chaque croisement le numéro de la famille d'isostérie à laquelle la paire correspondante appartient. Certaines combinaisons ne sont pas possibles compte tenu des

groupements chimiques présents sur chaque face. Par exemple, la paire GoG trans Watson-Crick/Sucre n'a pas été observée et la case correspondante dans la matrice reste vide.

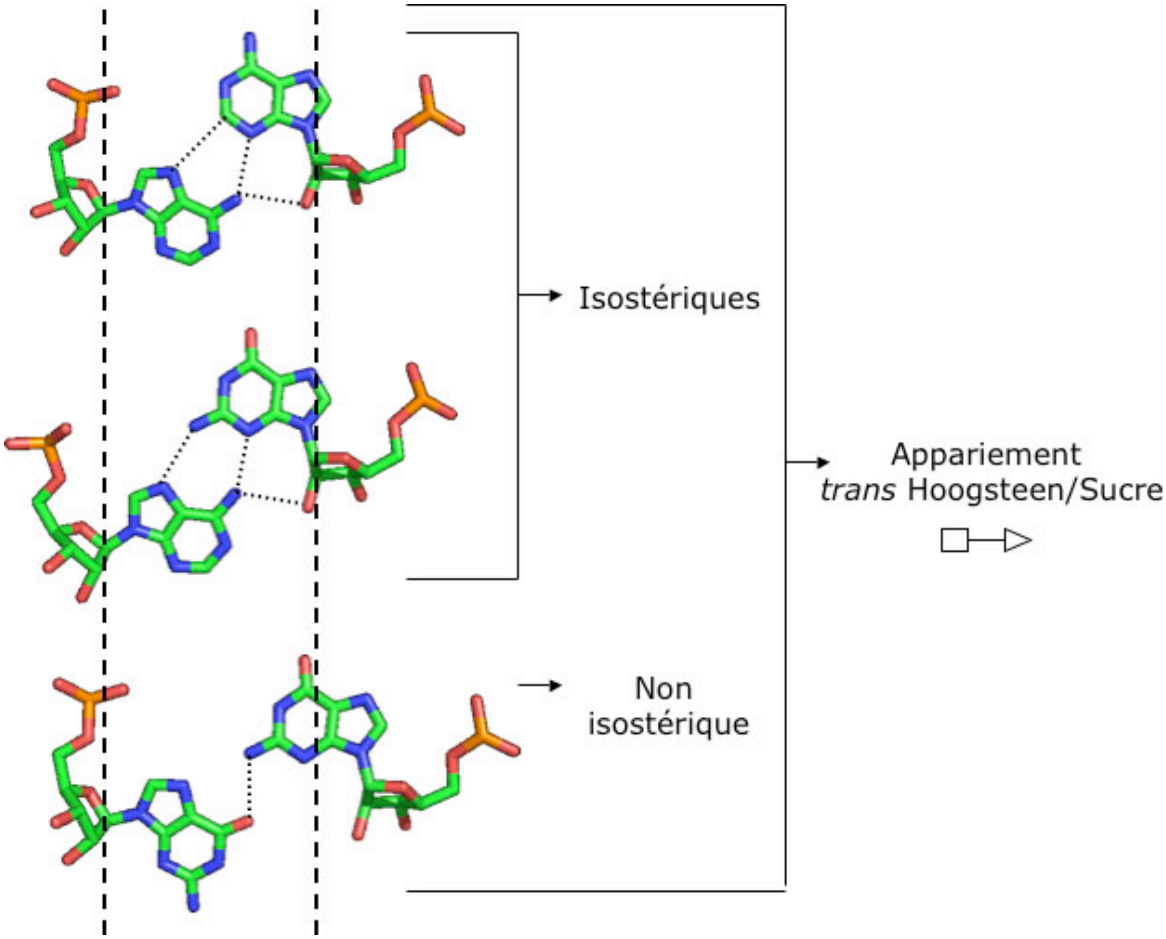
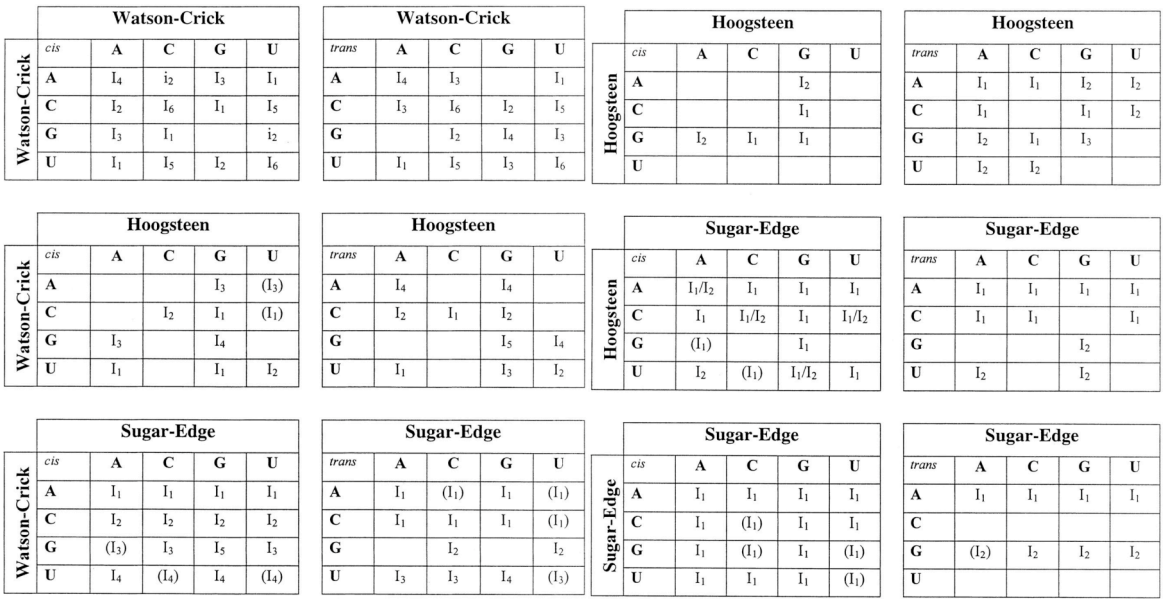


Figure 5: Appariements isostériques et non isostériques appartenant à la famille *trans* Hoogsteen/Sucre. En haut et au milieu les paires AoA et AoG sont isostériques. La distance C1'-C1' est identique dans les deux appariements ; elles appartiennent à la même famille d'isostérie. La troisième paire GoG n'est pas isostérique aux deux autres ; sa distance C1'-C1' est différente des deux autres. Elle appartient à une autre famille d'isostérie.

Alors que les appariements qui définissent la structure secondaire d'une molécule d'ARN n'impliquent que les côtés Watson-Crick des deux bases impliquées, les appariements qui composent un motif peuvent impliquer aussi les côtés Hoogsteen et Sucre des bases qui les composent. Les appariements qui impliquent ces côtés appartiennent aux 12 familles géométriques et possèdent des matrices d'isostérie caractéristiques. Dans un alignement de séquences homologues, comme pour les paires canoniques, les covariations des paires non

Watson-Crick qui composent un motif, doivent être cohérentes avec les matrices d'isostérie correspondantes. Le remplacement d'une paire non Wason-Crick par une autre, ne peut se faire que si les deux sont isostériques. Souvent, une paire non Watson-Crick est impliquée dans un réseau d'interactions avec d'autres bases de la molécule. C'est le cas par exemple dans les motifs ARN qui sont une succession de paires non Watson-Crick. Les règles de covariations d'un appariement du réseau subissent des contraintes additionnelles, et les variations observées dans un alignement seront réduites. L'ensemble des variations de séquence des appariements non Watson-Crick d'un motif est appelé séquence signature du motif. Le procédé de détermination d'une séquence signature sera dévoilé dans le chapitre 2.2 et illustré par deux exemples : le motif tournant K et le motif C.



Watson-Crick	cis	A	C	G	U
	A	I_4	i_2	I_3	I_1
	C	I_2	I_6	I_1	I_5
	G	I_3	I_1		i_2
	U	I_1	I_5	I_2	I_6

Watson-Crick	trans	A	C	G	U
	A	I_4	I_3		I_1
	C	I_3	I_6	I_2	I_5
	G		I_2	I_4	I_3
	U	I_1	I_5	I_3	I_6

Hoogsteen	cis	A	C	G	U
	A			I_2	
	C			I_1	
	G	I_2	I_1	I_1	
	U				

Hoogsteen	trans	A	C	G	U
	A	I_1	I_1	I_2	I_2
	C	I_1		I_1	I_2
	G	I_2	I_1	I_3	
	U	I_2	I_2		

Watson-Crick	cis	A	C	G	U
	A			I_3	(I_3)
	C		I_2	I_1	(I_1)
	G	I_3		I_4	
	U	I_1		I_1	I_2

Watson-Crick	trans	A	C	G	U
	A	I_4		I_4	
	C	I_2	I_1	I_2	
	G			I_5	I_4
	U	I_1		I_3	I_2

Hoogsteen	cis	A	C	G	U
	A	I_1/I_2	I_1	I_1	I_1
	C	I_1	I_1/I_2	I_1	I_1/I_2
	G	(I_1)		I_1	
	U	I_2	(I_1)	I_1/I_2	I_1

Hoogsteen	trans	A	C	G	U
	A	I_1	I_1	I_1	I_1
	C	I_1	I_1		I_1
	G			I_2	
	U	I_2		I_2	

Watson-Crick	cis	A	C	G	U
	A	I_1	I_1	I_1	I_1
	C	I_2	I_2	I_2	I_2
	G	(I_3)	I_3	I_5	I_3
	U	I_4	(I_4)	I_4	(I_4)

Watson-Crick	trans	A	C	G	U
	A	I_1	(I_1)	I_1	(I_1)
	C	I_1	I_1	I_1	(I_1)
	G		I_2		I_2
	U	I_3	I_3	I_4	(I_3)

Sugar-Edge	cis	A	C	G	U
	A	I_1	I_1	I_1	I_1
	C	I_1	(I_1)	I_1	I_1
	G	I_1	(I_1)	I_1	(I_1)
	U	I_1	I_1	I_1	(I_1)

Sugar-Edge	trans	A	C	G	U
	A	I_1	I_1	I_1	I_1
	C				
	G	(I_2)	I_2	I_2	I_2
	U				

Figure 5 : Matrices d'isostérie des 12 familles d'appariements. Figure extraite de l'article de Leontis et collaborateurs (2002).

2. RESULTATS ET DISCUSSIONS

2.1. Les motifs architecturaux de l'ARN

2.1.1. Définitions et outils de classification

Le terme "motif" a été utilisé pour les protéines avant d'être utilisé pour les ARN. En ce qui concerne les protéines, le terme « motif » peut désigner deux concepts différents. Le premier se place au niveau de la structure primaire et représente le motif comme une succession récurrente d'un nombre défini et ordonné d'acides aminés caractéristiques. La séquence consensus met en relief les conservations parmi différents exemples de séquences appartenant à un tel motif. Le deuxième concept propose que des éléments reliés de structure secondaire s'organisent dans l'espace et forment des motifs tridimensionnels simples. Le motif constitue, dans ce cas, une combinaison de quelques éléments de structure secondaire (hélice α, brin β, boucle), possédant un arrangement géométrique spécifique. Les différents motifs apparaissent régulièrement au sein des structures protéiques ; la récurrence est une propriété importante du motif. Certains motifs peuvent être associés à une fonction biologique particulière tandis que d'autres font partie d'ensembles fonctionnels de structures plus importantes. L'assemblage des motifs entre eux, au sein d'un même polypeptide, mène à la structure tertiaire de la protéine.

La signification du "motif ARN" est également ambigüe et il existe dans la littérature plusieurs définitions selon que la séquence nucléotidique ou la structure de l'ARN soit considérée. La séquence primaire de l'ARN peut présenter une séquence motif dont l'arrangement des nucléotides est défini, et qui joue un rôle particulier. Il existe de nombreux exemples de telles séquences comme le site de fixation des protéines Sm sur les petits ARN nucléaires du complexe d'épissage (spliceosome), la séquence Shine-Dalgarno des ARNm bactériens nécessaire au positionnement de la petite sous-unité ribosomique sur le codon initiateur de la traduction ou tout simplement les codons d'un ARNm. Des variations dans la composition nucléotidique de ces séquences sont courantes et une séquence consensus peut être définie. Les ARN structurés sont composés de motifs structuraux récurrents insérés au sein d'éléments hélicoïdaux. Comme

présenté dans l'introduction, la structure 2D d'un ARN est une représentation des appariements Watson-Crick canoniques, c'est-à-dire des éléments hélicoïdaux et des boucles, bulles internes et brins jonctions qui les relient. Dans une structure tridimensionnelle, les nucléotides en simple brin forment des interactions non Watson-Crick et les éléments de structure secondaire constituent, dans la grande majorité des cas, des motifs d'ARN. Le motif d'ARN structuré peut être défini de deux manières différentes selon que l'on considère la conformation de son squelette sucre-phosphate ou la géométrie des paires de bases qui le composent.

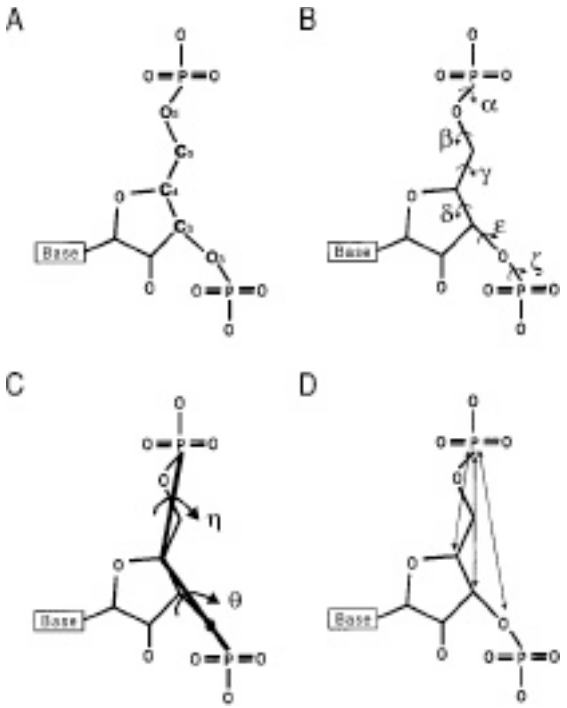


Figure 7 : Résumé des quatre représentations utilisées pour fixer l'information de structure des nucléotides. A :coordonnées cartésiennes, B :angles de torsion, C :pseudoangles de torsion, D :distances. D'après Reijmers et collaborateurs (2001).

La première définition, basée sur la conformation du squelette sucre-phosphate, implique que chaque motif présente une conformation caractéristique de son squelette sucre-phosphate. Elle est à la base de la conception d'outils informatiques utilisés pour analyser et classer les motifs dans des structures tridimensionnelles connues d'ARN. Quelques exemples de tels outils, décrits ci-dessous, mettent en relief différentes approches de traitement du "motif".

Globalement, il existe deux manières d'exploiter les données d'une structure tridimensionnelle : soit l'utilisation des données cartésiennes

(contenues par exemple dans les fichiers pdb) soit l'utilisation de coordonnées internes (angles de torsion, pseudoangles, distances entre atomes) (Figure 7). Reijmers et collaborateurs ont étudié l'influence de la représentation de l'information de structure moléculaire sur les résultats de classements d'ARN trinucléotidiques (Reijmers et al., 2001). Ils considèrent que l'utilisation des coordonnées cartésiennes pour classer différentes structures constitue la plus juste représentation ("gold standard"). Huang et collaborateurs ont développé un logiciel basé sur cette représentation permettant de comparer et de classer les boucles d'ARN (Huang et al., 2005). Des régions de même taille de 15 structures cristallographiques d'ARN ont été superposées et les valeurs de RMSD calculées. Le classement en fonction de la valeur de RMSD a généré des groupes de segments présentant un repliement identique. L'analyse a révélé deux groupes majeurs correspondant aux motifs connus GNRA et UNCG, et permis l'identification de ces tétraboucles au sein de boucles plus larges. Cette technique peut être utilisée pour repérer une boucle qui serait exclue du groupe auquel elle appartient car mal affinée et y remédier. L'utilisation de coordonnées cartésiennes présente toutefois deux inconvénients majeurs : (i) le nombre de variables est très important et (ii) les structures doivent être superposées avant d'être classées.

L'utilisation de coordonnées internes comme les angles de torsion permet de se détacher de ces contraintes. Les résultats de Reijmers et collaborateurs montrent que la représentation par dendrogramme basée sur les angles de torsion diffère de manière importante de la représentation par coordonnées cartésiennes, alors que les représentations par pseudotorsion ou basées sur des distances entre atomes au sein d'un nucléotide en sont plus proches (Reijmers et al., 2001). Récemment, Wadley et Pyle ont proposé une identification de nouveaux motifs récurrents dans les structures cristallographiques d'ARN de la NDB (Nucleic Acid Database ; http://ndbserver.rutgers.edu) par un algorithme automatique appelé COMPADRES (COMParative Algorythm to Discover Recurring Elements of Structure) qui utilise une représentation par pseudotorsion du squelette sucre-phosphate (Duarte & Pyle, 1998; Wadley & Pyle, 2004). Les nouveaux motifs sont définis sur la trajectoire du squelette. Plusieurs fragments de squelette présentant la même succession d'angles θ (P_i - $C4'_i$ - P_{i+1} - $C4'_{i+1}$) et η ($C4'_{i-1}$ - P_i - $C4'_i$ - P_{i+1}) sont superposables et donc récurrents ; ils définissent ainsi un nouveau motif. Ces nouveaux motifs, dont trois étaient inconnus

jusqu'alors, ne présentent pas, selon les auteurs, de structure primaire ou secondaire caractéristique. Pourtant, nous montrerons plus loin que ces « nouveaux » motifs présentent une signature caractéristique d'appariements non Watson-Crick et qu'ensemble avec d'autres brins, ils appartiennent à des motifs structuraux déjà définis.

Hershkovitz propose deux méthodes de reconnaissance et classement des motifs ARN basées sur les angles de torsion (Hershkovitz et al., 2003). Les deux méthodes utilisent un espace de torsion et non pas un espace cartésien, ce qui évite d'utiliser transformation ou superposition pour comparer des fragments d'ARN. Les fragments d'ARN présentant la même conformation d'angles de torsion forment un groupe. La structure obtenue en faisant la moyenne des angles de torsion de ces fragments conduit à une conformation moyenne ou "parent". La conformation parent est ensuite comparée à la librairie de familles de conformations connues (tétraboucle, hélice de type A...) pour être identifiée. Des six angles de torsion définissant le squelette sucre-phosphate d'un nucléotide, deux (β et ε) sont quasiment invariants et centrés autour de 180°. C'est pourquoi les auteurs ont défini quatre angles comme identifiants conformationnels d'un nucléotide. A chaque résidu est assigné une séquence de quatre nombres entiers chacun spécifique d'un intervalle de valeur de l'angle α, γ, δ et ζ (par exemple : α=1 correspond à des angles compris entre 40 et 90°). 37 séquences différentes décrivant un état conformationnel ont été observées dans la grande sous unité ribosomique et les fréquences d'apparition de ces séquences varient. A chaque séquence a été attribué un caractère. Un motif conformationnel, défini comme une succession caractéristique d'états conformationnels, peut être représenté comme une succession de caractères. Il est alors facile de repérer par ordinateur des successions identiques de caractères révélatrices d'un même motif. Le but est d'obtenir une méthode efficace de description et d'analyse de la conformation de l'ARN réduisant la quantité d'information sans pour autant sacrifier l'exactitude. Cette méthode présente l'inconvénient de ne pas considérer les interactions entre résidus mais les auteurs évoquent la possibilité de combiner leur travail avec la nomenclature et la classification basées sur la géométrie des paires de bases proposées par Leontis et Westhof (2001).

La base de données SCOR (Structural Classification Of RNA, version 2.0 ; http://scor.lbl.gov/), équivalent ARN de la base de données de structures

protéiques SCOP, propose une classification des boucles internes et terminales d'une collection de 497 structures ARN résolues par RMN et cristallographie. Les structures sont classées de trois manières différentes : classification structurale, fonctionnelle ou interactions tertiaires (Figure 8). Chaque classe appartenant à une des trois classifications se ramifie en « sous-classes » qui elles-mêmes mèneront vers d'autres groupes (Table 1).

Types de classification (a)	Classes (b)	Nbr mb (c)	Nbr s-c (d)	Sous-classes (e)
Classification structurale	Boucles internes	5350	16	Doucle brin empilé, appariements non-WC Bases vers l'extérieur
	Boucles terminales	2920	18	Tournant U Boucles GNRA
Classification fonctionelle	Fonction moléculaire	480	2	ARN naturel ARN SELEX
	Fonction motif	179	4	Liaison de protéine Empilement d'hélices
	Modèles structuraux	137	14	Duplexes autocomplémentaires Duplexes hybrides ARN-ADN
Interactions tertiaires	Hélices coaxiales	7	6	Hélices I et II Ribozyme tête de marteau Domain (P5a-P5b) intron de groupe I
	Tetraboucle-Recepteur	1	1	GAAA:J6a/6b intron de groupe I
	Motif A mineur	240	4	Motif A-Mineur Type I Motif A-Mineur Type II
	"Kissing Loops"	32	5	Site initiation dimerisation retrovirus ARNr 23S
	BouclesD:T ARNt	7	3	ARNt élongateur ARNt initiateur
	Pseudonoeuds	17	8	Pseudonoeud aptamère B12 Pseudonoeud virus tumeur mammaire de la souris
	"Ribose Zipper"	657	7	Ribose Zipper canonique Ribose Zipper simple

Table 1 : Organisation de la base de données SCOR. Sont indiqués : d'abord les différentes classifications (a), puis pour chaque classe (b) le nombre de membres contenus (c), le nombre de sous-classes (d) et les deux sous-classes contenant le plus de membres (e).

La classification structurale est organisée selon un graphique acyclique dirigé (Directed Acyclic Graph ou DAG) de manière que chaque nœud puisse avoir plusieurs nœuds parents. Ceci est représenté sur la partie droite de la Figure 8. Ainsi, la boucle interne en position 4 de la tige boucle de l'ARN Psi de HIV-1 appartient au groupe « boucles avec plateforme dinucléotidique impliquée dans une triple » qui peut être atteint par deux voies différentes : une passant par le groupe « boucles avec triple » et l'autre par le groupe « boucles avec plateforme dinucléotidique ».

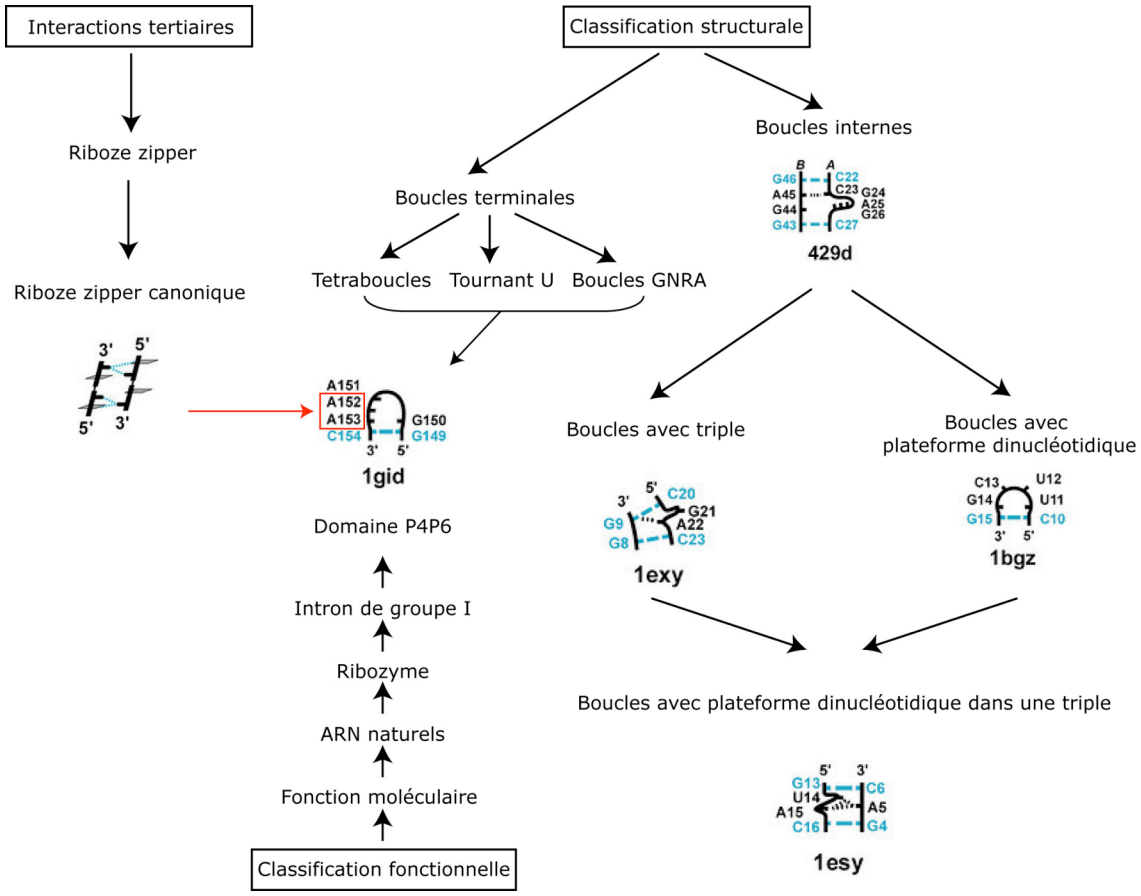


Figure 8 : Exemples montrant l'organisation de la banque de données SCOR. Les trois types de classification sont représentés. L'exemple de la boucle GAAA du domaine P4P6 de l'intron de groupe I (PDB ID : 1gid) montre le type d'informations que donnent les trois classifications. L'exemple de la boucle interne 2 de l'ARN Psi de HIV 1 (PDB ID :1esy) montre que plusieurs chemins permettent d'accéder à une classe ce qui est caractéristique d'une organisation selon un graphique acyclique dirigé.

La classification fonctionnelle présente deux niveaux, un niveau fonction moléculaire et un niveau fonction du motif. Les ARN classés en fonction de leur fonction moléculaire se partagent entre le groupe des ARN naturels et le groupe des ARN sélectionnés SELEX (Systematic Evolution of Ligands by EXponential enrichment). Le premier groupe inclut les ARNt et les ARNr parmi d'autres ARN trouvés dans la nature. Dans le second groupe des structures locales fonctionnelles appartenant à une structure globale d'ARN sont classées en motifs de liaison protéique, de liaison de petits ligands, de reconnaissance de réplicase ou d'empilement d'hélices. Le groupe des modèles structuraux montre des aspects spécifiques de la structure et de la fonction d'ARN. Le motif fonctionnel, plus bas niveau dans cette classification fonctionnelle, a sa place également dans

les classifications structurale ou d'interactions tertiaires. C'est ce qui est représenté sur la Figure 8 où la boucle GAAA en position 150 du domaine P4P6 de l'intron de groupe I de *Tetrahymena thermophila* appartient au domaine fonctionnel P4P6 dans la classification fonctionnelle, aux groupes tetraboucles, tournants U et boucles GNRA de la classification structurale, et, au groupe "riboze zipper" canoniques de la classification des interactions tertiaires.
La classification des interactions tertiaires est divisée en classes incluant un certain nombre des interactions tertiaires responsables du repliement, de la stabilité et/ou de la maintenance de la structure tridimensionnelle d'un ARN comme les interactions boucle/boucle ou les interactions tetraboucle/récepteur. Les motifs en A mineur représentent un groupe très important. Ils seront détaillés dans le prochain sous chapitre. L'interdigitation entre les classifications permet à l'utilisateur d'obtenir une foule d'informations sur un motif donné, sa structure, sa fonction ou d'identifier les motifs qui appartiennent au même groupe. Pourtant, il manque à SCOR un mode de recherche d'information basé sur le critère de structure secondaire ou sur la nature des appariements.

Tout récemment, Laserson et collaborateurs ont proposé et appliqué un protocole informatique identifiant, dans les génomes, de petits motifs fonctionnels basés sur des motifs fonctionnels synthétiques dérivés d'expériences de sélection *in vitro* (Laserson et al., 2005). Cette méthode est basée sur l'idée du groupe de Szostak que la sélection *in vitro*, en « reproduisant » le processus d'évolution, peut recréer des motifs d'ARN actifs identifiés *in vivo*, comme pour le ribozyme à tête de marteau (Wilson & Szostak, 1999). D'un motif sélectionné pour une fonction de liaison d'une petite molécule (ATP, chloramphénicol, streptomycine ou néomycine B) *in vitro*, sont extraites des caractéristiques spécifiques de séquence et de structure secondaire (plus fine si la structure cristallographique ou RMN est connue) qui sont utilisées pour cribler, à l'aide du logiciel *RNAMotif*, des génomes connus et trouver des motifs ARN naturels équivalents dont le repliement est vérifié. Les candidats sont finalement soumis à des tests statistiques de validation (chances de trouver ces caractéristiques par hasard) et de stabilité thermodynamique de manière à éliminer les faux positifs. La validation expérimentale devra être réalisée. Au final, l'objectif est d'augmenter le répertoire d'ARN fonctionnels connus.

Ces outils de recherche de motifs, basés sur des études de conformations récurrentes du squelette sucre-phosphate sont très utiles. Ils permettent, de

manière automatique, de trouver de nouveaux exemples de motifs connus et d'identifier des motifs inconnus jusqu'alors ce qui évite l'exercice laborieux de l'analyse « à l'écran » de grandes structures tridimensionnelles comme comme celles des sous-unités ribosomiques. Tous ces motifs, connus ou nouveaux, sont définis par la conformation caractéristique de leur squelette sucre-phosphate. La limitation majeure de cette définition est qu'elle repose sur la structure tridimensionnelle du motif sans faire de lien avec la séquence. La conséquence est qu'elle n'est pas utilisable pour la prédiction d'un motif à partir de la séquence nucléotidique.

La définition basée sur la géométrie des paires de bases, et c'est là la grande différence, a été proposée dans le but de caractériser des motifs à partir de leur séquence. Le motif est défini comme un ensemble de paires de bases non Watson-Crick ordonnées et orientées qui mènent à un repliement caractéristique du squelette sucre-phosphate. Cette définition, qui décompose le motif en éléments élémentaires (les appariements non canoniques), permet, tout comme la définition "squelette sucre-phosphate" d'identifier, d'analyser et de classer les motifs ARN. Toutes les sections des ARN structurés, de séquence similaire ou différente mais présentant la même succession ordonnée et orientée d'appariements non Watson-Crick, appartiennent au même motif. En d'autres termes, chaque paire non Watson-Crick d'un motif peut être remplacée par une paire impliquant des bases différentes pourvu que la géométrie soit conservée c'est à dire que les deux paires de bases soient isostériques. Ainsi, sous la contrainte de l'isostérie des paires non Watson-Crick d'un motif, différentes séquences nucléotidiques se replient en une structure 3D spécifique. L'ensemble de ces séquences qui se replient en un même motif, forment la "séquence signature" du motif. On peut donc aligner des séquences à des positions qui forment des paires non Watson-Crick en respectant les règles de covariations imposées par les matrices d'isostérie des paires non Watson-Crick concernées. Leontis et Westhof ont analysé les variations observées pour chaque paire non Watson-Crick de la boucle E de l'ARNr 5S chez les bactéries (Leontis & Westhof, 1998). Toutes les substitutions observées étaient isostériques à la paire correspondante de la structure cristallographique. La séquence signature du motif a été ainsi déterminée. L'analyse des variations compensatoires des séquences de la boucle asymétrique en position 581 de l'ARNr 16S

(numérotation d'*E.Coli*), et leur cohérence avec la séquence signature a permis de mettre en évidence une structuration en boucle E. De la même manière, les auteurs ont pu montrer que sur les neuf boucles sarcine-ricine prédites par analyse des covariations non Watson-Crick dans les ARNr 16S et 23S, une seule n'était pas présente dans les structures cristallographiques publiées (Leontis et al., 2002a). Dans le chapitre 2.2, nous avons déterminé de la même manière la séquence signature de deux motifs récurrents : le motif C et le tournant K. Leur séquence signature, tout comme dans le cas de la boucle E, servira dans le futur à identifier ces motifs dans un alignement de séquences homologues.

La définition du motif décrite ci-dessus est basée sur la classification géométrique des paires de bases. Elle permet l'identification de motif au sein d'alignement de séquences ARN homologues. Outre les règles d'isostérie, certains motifs répondent à d'autres contraintes qui limitent les paires autorisées à certaines seulement d'une même sous famille isostérique. C'est ce que l'on observe par exemple dans le cas du motif A mineur qui sera discuté plus loin.

Dans la suite de ce travail, nous considèrerons un motif comme une succession ordonnée et sous contraintes de paires de bases non Watson-Crick. Les descriptions de motifs et les règles de covariation que nous proposons sont basées sur les règles d'isostérie des paires de bases.

2.1.2. Interactions à longue distance

La stratégie des ARN structurés pour se replier de manière stable et fonctionnelle, consiste à former des interactions à longue distance (ou interactions tertiaires ou encore interactions ARN-ARN) entre les éléments périphériques au coeur de la structure. De nouveaux types d'interactions tertiaires ont été observés dès la première structure cristallographique d'ARN dont la taille dépassait l'ARNt : les 160 nucléotides du domaine P4P6 de l'intron de groupe I de *Tetrahymena*. Auparavant, de nombreux travaux avaient montré le rôle des interactions à longue distance entre éléments périphériques dans le cas des ribozymes et plus particulièrement dans le cas des introns de groupe I.

Selon la disposition et le nombre d'éléments de structure secondaire encadrant le cœur catalytique, différents sous-groupes d'introns de groupe I ont été définis (Burke et al., 1987). Ils possèdent un coeur catalytique identique mais des éléments périphériques caractéristiques qui représentent plus de la moitié de leur séquence et dont la morphologie a permis leur classement en quatre groupes et onze sous groupes (Figure 9). Les interactions à longue distance entre éléments périphériques qui contraignent la structure native des ribozymes ont été mises en évidence par analyse comparative de séquences. Au laboratoire, Luc Jaeger a étudié les régions périphériques de plusieurs introns de groupe I et a déterminé les types d'interactions à longue distance impliqués dans le repliement et la stabilisation du coeur (Jaeger, 1995). Il a montré que la majorité de ces interactions étaient des interactions de type pseudonoeud impliquant des paires Watson-Crick mais qu'il existait également des interactions non Watson-Crick entre une boucle de type GNRA et une hélice.

Contrairement au coeur catalytique, les éléments périphériques peuvent être souvent éliminés sans perte majeure de l'activité de l'intron de groupe I dans des expériences *in vitro*. Toutefois, il a été montré récemment qu'en absence de P5abc, l'ARN de *Tetrahymena* se repliait en des formes alternatives aussi stables que la forme native, alors que la présence de P5abc induit une stabilisation très importante de la conformation native de l'ARN de *Tetrahymena* par rapport aux conformations alternatives (Engelhardt et al., 2000; Johnson et al., 2005). D'autre part, une concentration en sels plus élevée est généralement nécessaire pour compenser la perte de repliement autour du coeur catalytique engendrée par l'élimination des régions périphériques et la perte des interactions tertiaires (Doudna & Szostak, 1989a; Joyce et al., 1989; Salvo & Belfort, 1992). Comme nous le verrons dans le chapitre 2.3, le ribozyme à tête de marteau montre lui aussi des interactions tertiaires entre ses éléments périphériques au cœur catalytique, responsables de l'activité du ribozyme à des concentrations physiologiques de Mg^{2+}. Les interactions tertiaires assurent la stabilisation du cœur catalytique de ces ribozymes. Dans le cas des introns de groupe I, chaque sous-groupe possède un lot d'éléments périphériques et d'interactions spécifiques, chacun étant une solution différente et spécifique de stabilisation du coeur catalytique.

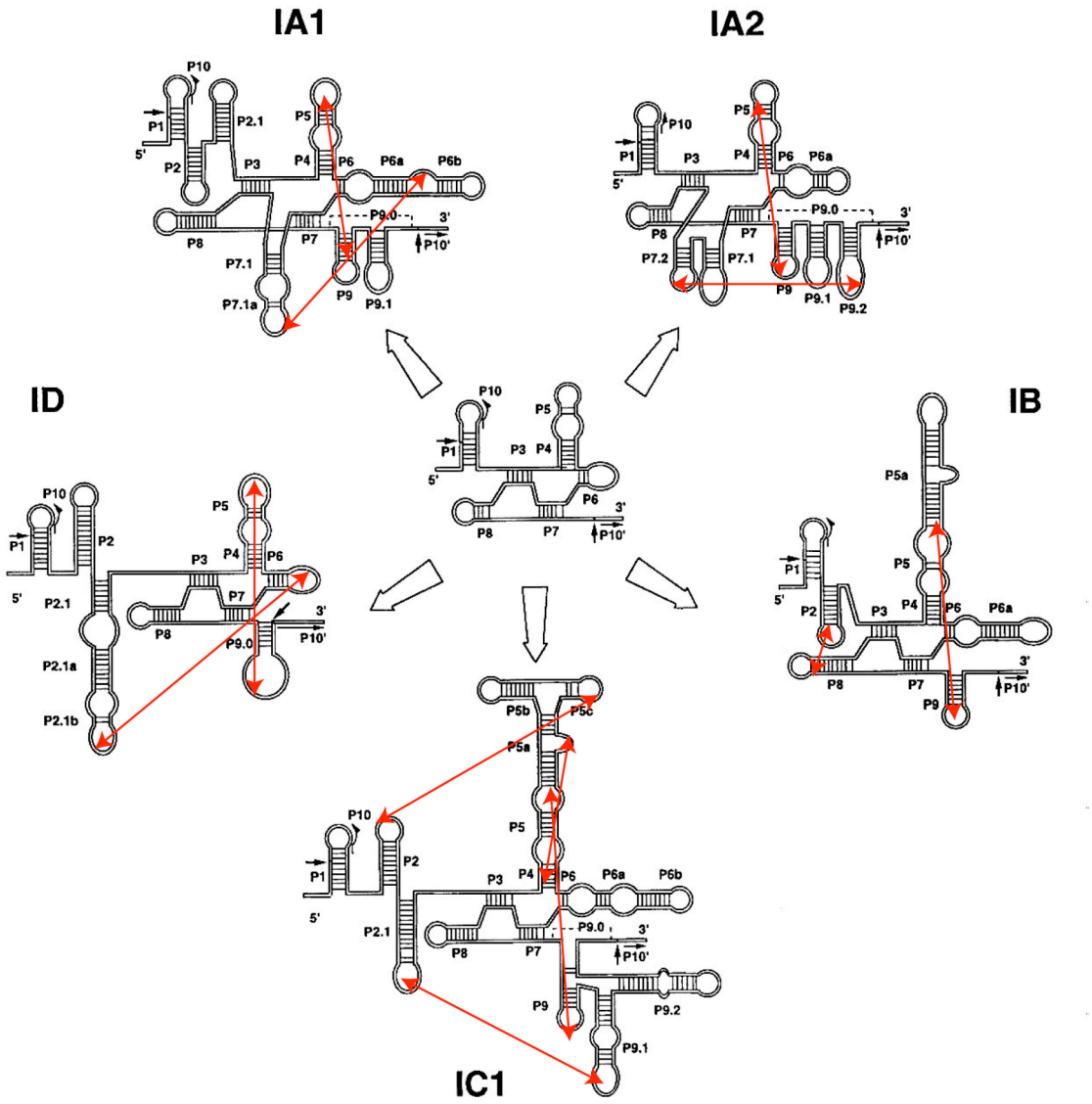


Figure 9 : Diversité des interactions à longue distance dans les différents sous groupes d'introns de groupe I. D'après Jaeger et al (1991)

La ribonucléase P (RNase P) est un ribozyme responsable de la maturation de l'extrémité 5' des ARNt et sa structure secondaire a été établie par analyse comparative de séquences (Haas et al., 1991; Brown et al., 1993; Haas et al., 1994; Haas et al., 1996a; Haas et al., 1996b). Il existe deux types de structures secondaires, type A et type B, dont la structuration du coeur est similaire mais qui diffèrent par leurs éléments périphériques. De manière identique, les RNase P de types A et B s'organisent en deux domaines. Le domaine I, domaine de spécificité, est responsable de la reconnaissance du substrat (pre-ARNt) ; il est constitué des hélices P7 à P12, de la boucle interne L11/12 et des hélices P13 et P14 pour le type A ou P10.1 pour le type B. Le domaine II, composé des hélices

P1 à P5 et des extensions P16/P17/P6 et P18 (Type A) ou P5.1/P15.1/P15.2 (Type B), comporte les nucléotides du site catalytique. La modélisation, les expériences d'empreintes chimiques en solution et l'analyse comparative de séquences ont mis en évidence de nombreuses interactions à longue distance représentées en rouge sur la figure 10 (Massire, 1998). La modélisation avait montré que les deux types de RNase P, bien que présentant des éléments périphériques différents, partagent un coeur central similaire centré autour des interactions à longue distance impliquant les boucles L8 et L9 (Massire et al., 1998). Ces résultats ont été confirmés récemment par les structures cristallographiques du domaine de spécificité des RNases P de types A et B (Krasilnikov et al., 2003; Krasilnikov et al., 2004). Ainsi, comme dans le cas de l'intron de groupe I, malgré les différences d'architecture des régions périphériques, le coeur de l'ARN de la RNase P se replie de la même manière.

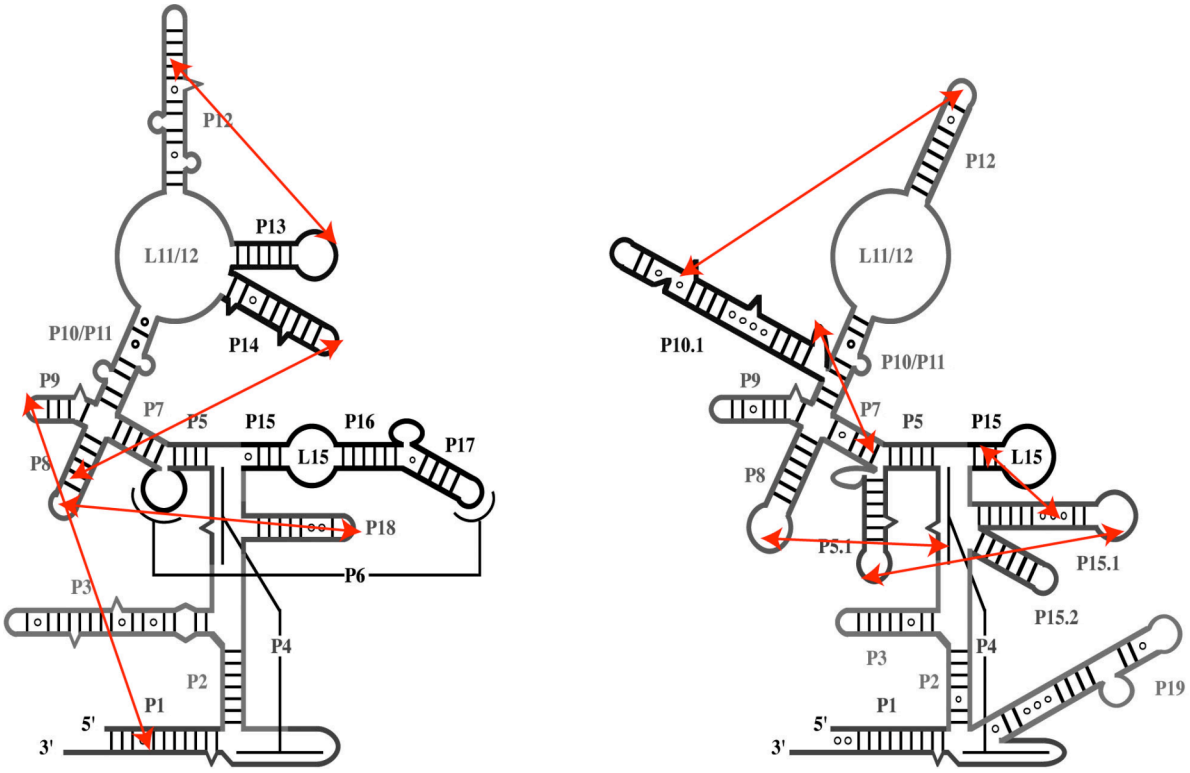


Figure 10 : Représentation des RNases P de types A et B. En rouge sont indiquées les interactions à longue distance entre les éléments périphériques.

Les exemples précédents montrent que différents motifs permettent à des régions éloignées d'interagir entre elles : la boucle GAAA et son récepteur à onze nucléotides (Costa & Michel, 1997), la boucle GNRA et deux paires consécutives d'une hélice Watson-Crick (Michel & Westhof, 1990) ou les adénines d'un simple brin et deux paires canoniques. Les motifs impliqués dans des interactions à longue distance sont variés tout comme la nature sous-jacente des appariements. Ainsi, il est préférable de distinguer le motif d'interaction du protocole d'interaction qu'il implique. Il existe finalement assez peu de motifs d'interaction à longue distance différents ; ils sont répertoriés et définis dans la Table 2. Le faible nombre de motifs ARN-ARN différents illustre la modularité de l'ARN. Les motifs d'interactions à longue distance constituent, comme les motifs insérés dans des hélices, des « briques » utilisées dans différents ARN structurés. Le motif d'interactions à longue distance le plus célèbre est l'association boucle/boucle ("kissing loop") qui est composé essentiellement de paires Watson-Crick. Mais les interactions les plus courantes impliquent les côtés Sucre des bases. Les interactions de ce type les plus souvent observées, impliquent les côtés Sucre d'adénines qui forment des contacts tertiaires en interagissant avec le petit sillon d'une hélice canonique ; elles sont appelées interactions en A mineur. Ces interactions sont impliquées dans la plupart des motifs d'interactions à longue distance comme indiqué dans la Table 2. L'association de différentes interactions en A mineur mène à différents motifs en A mineur. Parce que les motifs en A mineur sont les plus abondants parmi les motifs d'interactions à longue distance et qu'ils jouent des rôles fondamentaux lors de la synthèse protéique, nous avons choisi de les étudier en détail dans la suite (chapitre 2.1.4).

Motifs d'interaction (a)	Protocoles d'interaction (b)	Structure d'ARN (c)	Nomenclature (d)	PDB ID (e)	Référence (f)
"Kissing complex"	Appariements Watson-Crick	Domaine ALU de SRP	L1.2/L2	1E8O	(Weichenrieder et al., 2000)
Intercalation	BoucleT/BoucleD	ARNt	BoucleT/BoucleD		
		RNaseP types A & B	L11/12	1U9S 1NBS	(Krasilnikov et al., 2003; Krasilnikov et al., 2004)
		ARNr 23S		1S72	(Klein et al., 2004)
GNRA/Hélice (/2bp)		Ribozyme tête de marteau	III/Hélice I ou II	1HMH	(Pley et al., 1994b, 1994a)
		ARNr 16S		1J5E	(Wimberly et al., 2000)
		ARNr 23S		1S72	(Klein et al., 2004)
		RnaseP T.m	L18/P8	2A2E	(Torres-Larios et al., 2005)
		Intron I Twort	L2/P8, L9/P5	1YOQ	(Golden et al., 2005)
		RNaseP type A	L14/P8	1U9S	(Krasilnikov et al., 2004)
		Intron I Tetrahymena	P9/P5 J3-4/P6	1X8W	(Guo et al., 2004)
GAAA/Récepteur (/11bp)	▶ Motif A mineur	Intron I Tetrahymena	P5b/P6b	1GID 1X8W	(Cate et al., 1996a) (Guo et al., 2004)
		Intron I Azoarcus	P2/J8-8a, P9/J5-5a	1U6B	(Adams et al., 2004b)
AA/Hélice (/2bp)		Intron I Azo & Twort	J3-4/P6	1U6B 1YOQ	(Adams et al., 2004b; Golden et al., 2005)
		RnaseP type B	P10.1/P10-P7, P8/P4	1NBS	(Krasilnikov et al., 2003)
Boucle E/Hélice (/2bp)		ARNr 16S	A2-A3 (motif générique)	1J5E	(Wimberly et al., 2000; Leontis et al., 2002a; Klein et al., 2004)
		ARNr 23S		1S72	
Boucle/Boucle	Intercalation Sucre/Sucre	Domaine S de SRP	Hélice 8/Hélice 6	1L9A	(Oubridge et al., 2002)
		Riboswitch	L2/L3	1U8D	(Batey et al., 2004)
		L11rRNA	Hélice A/ Hélice C	1QA6	(Conn et al., 1999)

Table 2 : Différents types d'interactions à longue distance. Les différents types d'interactions ARN/ARN (a) font intervenir quatre protocoles d'interaction différents (b). Les structures cristallographiques d'ARN dans lesquelles ces motifs ont été observés sont indiquées (d) ainsi que la nomenclature des régions qui interagissent (d). les deux dernières colonnes indiquent le code PDB et la référence bibliographique de chaque structure.

2.1.3. Diagrammes des réseaux d'interactions

Pendant longtemps notre compréhension de la structuration de l'ARN était limitée à notre seule connaissance de la structure cristallographique de l'ARNt résolue en 1974 (Robertus et al., 1974). La constante amélioration des techniques de synthèse, purification, cristallographie et RMN a permis de résoudre, au niveau atomique, des structures d'ARN de plus en plus grands. En 2000-2001, la résolution des structures des sous-unités ribosomiques a doublé la quantité d'information structurale disponible (Ban et al., 2000; Yusupov et al., 2001; Brodersen et al., 2002). L'analyse initiale de ces structures suggérait que la majorité des éléments structuraux présents étaient déjà connus. Mais des études plus détaillées ont montré qu'il restait une foule de données à extraire dont l'exploitation mènerait à la compréhension des caractéristiques architecturales de ces ARNr mais également des ARN structurés en général. Depuis, les structures d'autres grands ARN fonctionnels ont été déterminées apportant elles ausi une multitude de données structurales. Ainsi, les structures de trois introns de groupe I ont été résolues récemment : un intron de groupe I, appartenant à la bactérie pourpre *Azoarcus* avec ses deux exons en 5' et en 3' (Adams et al., 2004a; Adams et al., 2004b), un complexe du phage *Twort* ribozyme de groupe I/produit (Golden et al., 2005) et le ribozyme d'un intron de groupe I de *Tetrahymena* (Guo et al., 2004). D'autres structures, résolues ces deux dernières années, incluent les domaines de spécificité des ARN des RNaseP de type A et B (Krasilnikov et al., 2003; Krasilnikov et al., 2004). L'ensemble de ces structures tridimensionnelles représente une masse de données extraordinaire et la question de leur traitement suscite de nombreux travaux comme en témoigne le nombre d'articles parus dans la littérature. Une approche consiste à concevoir des outils informatiques capablent d'analyser une structure tridimensionnelle et d'extraire des informations afin d'éviter la fastidieuse examination visuelle de ces structures. Une description succincte de certains de ces outils a été proposée dans le chapitre 2.1.1.

Afin de faciliter l'utilisation de ces structures cristallographiques, une autre approche consiste à exploiter la nomenclature de Leontis et Westhof (2001) pour représenter schématiquement, en deux dimensions, un maximum de données tridimensionnelles. Cette nouvelle représentation bidimensionnelle ne doit pas

être confondue avec la structure 2D d'un ARN. Par définition, la structure 2D d'un ARN représente les éléments de structure secondaire c'est à dire les hélices Watson-Crick reliées par des simples brins (boucles, bulles etc). La structure 2D d'un ARN dont la structure tridimensionnelle est inconnue, peut être obtenue par utilisation de logiciels de minimisation d'énergie, comme Mfold, ou par analyse des covariations révélatrices des paires Watson-Crick dans un alignement de séquences homologues. Dans la représentation traditionnelle de la structure 2D d'un ARN de structure tridimensionnelle connue, la grande majorité des informations 3D sont perdues dont l'empilement des hélices et les appariements non Watson-Crick qui forment les motifs et les interactions à longue distance.

Une alternative, pour conserver l'information d'empilement des hélices, consiste à réaliser une représentation de la structure secondaire (schéma topographique) de la structure. La Figure 11 présente différentes représentations du cœur des introns de groupe I (11A et 11B) et de l'intron de groupe I de *Azoarcus* en entier dont la structure cristallographique a été résolue récemment (11C). La structure 2D (Figure 11A) de différents introns de groupe I dont le cœur est conservé, a été réalisée des années auparavant grâce à l'étude des covariations (Michel et al., 1982). Cette représentation montre les hélices Watson-Crick, les boucles terminales, les bulles internes et les simples brins. Une représentation de la structure secondaire (Figure 11B) a été proposée après modélisation de l'intron, c'est-à-dire une fois qu'une information d'architecture globale était connue (Cech et al., 1994). Les hélices empilées sont dessinées en continuité les unes par rapport aux autres. D'autre part, l'analyse des covariations entre régions éloignées a permis l'identification puis la modélisation de certaines interactions à longue distance impliquant des appariements non Watson-Crick ; elles sont représentés par des lignes pointillées (Jaeger, 1997). Au cours de l'année 2005, trois structures d'intron de groupe I, dont celle d'*Azoarcus* représentée sur la Figure 11C, ont été résolues par cristallographie aux rayons X (Adams et al., 2004b; Guo et al., 2004; Golden et al., 2005). Alors que l'information de structure tridimensionnelle n'est que partiellement indiquée sur une représentation de structure secondaire, le diagramme des réseaux d'interactions réalisé grâce à la nomenclature de Leontis et Westhof (2001), reflète l'ensemble de l'information 3D. Tous les appariements Watson-Crick et non Watson-Crick sont représentés à l'aide des symboles de la nomenclature. L'empilement des hélices ainsi que l'empilement des bases extrudées d'un intérêt

particulier, c'est-à-dire participant à l'organisation tridimensionnelle, sont également indiqués. Le diagramme des réseaux d'interactions d'une structure permet de rassembler les informations de structure tridimensionnelle sur un schéma en deux dimensions.

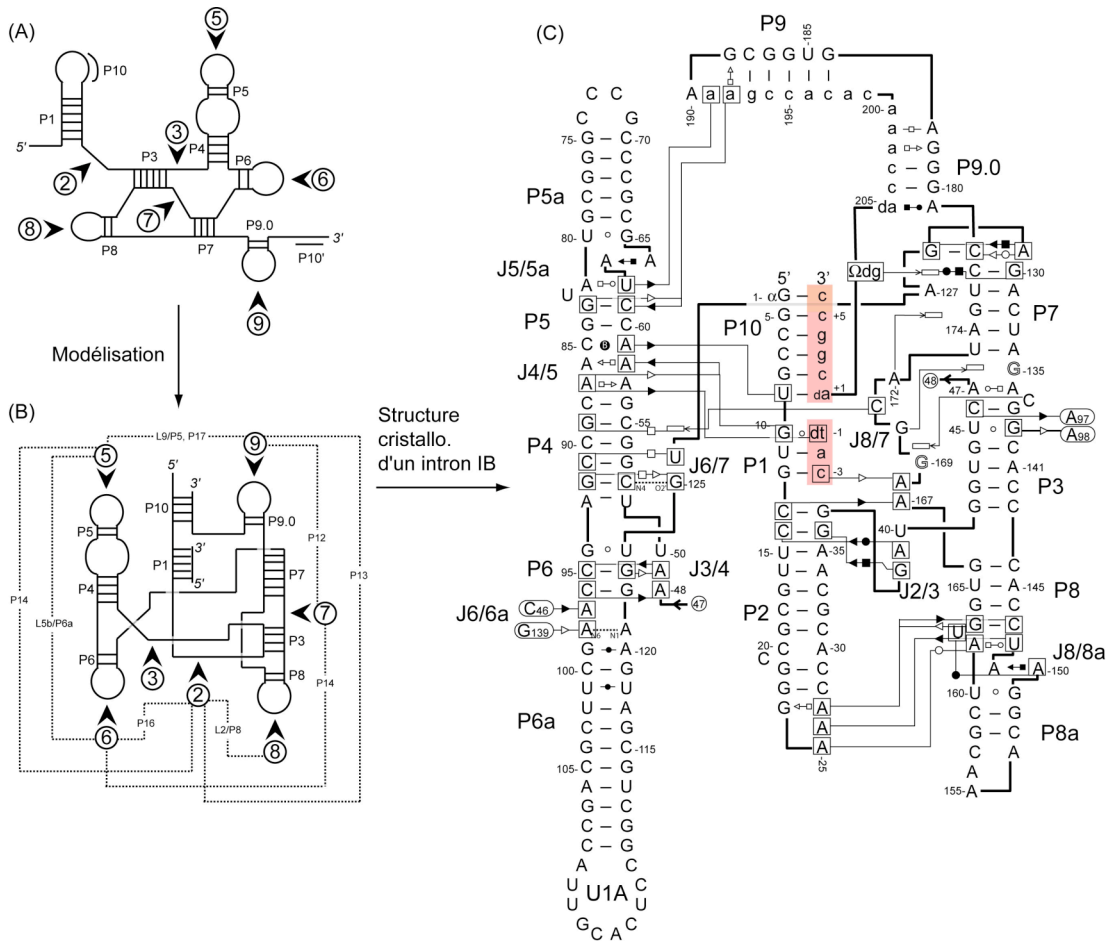


Figure 11 : Différentes représentations d'une structure ARN. La structure 2D du cœur des introns de groupe I est représentée avec seulement les informations d'appariements Watson-Crick (Burke et al., 1987) (A) et, après modélisation, avec les informations d'empilement d'hélices et les interactions à longue distance (Cech et al., 1994) (B). (C) Représentation des réseaux d'interactions 3D de la structure de l'intron de groupe I de *Tetrahymena* (Adams et al., 2004b).

Durant mon travail de thèse, j'ai réalisé un certain nombre de diagrammes de réseaux d'interactions de structures cristallographiques d'ARN structurés dans le but de favoriser leur analyse. Dans une première étape, le logiciel RNAMLview a été utilisé (Yang et al., 2003). Cet outil informatique permet de représenter les structures tridimensionnelles sur un plan en utilisant la nomenclature des appariements proposée par Leontis et Westhof (2001). La représentation

bidimensionnelle est manipulable et peut servir de base pour l'analyse de la structure d'une molécule dans le programme S2S réalisé par la suite au laboratoire (Jossinet & Westhof, 2005). Les représentations obtenues ont montré néanmoins de nombreuses inexactitudes et ont nécessité finalement une détermination nouvelle et non-automatique de la nature des appariements, de l'empilement des hélices ou de l'empilement particulièrement intéressant de certaines bases. Dans les pages suivantes sont représentées les diagrammes des réseaux d'interactions de structures cristallographiques d'ARN de taille et de complexité structurale croissante ; la liste de ces structures est donnée dans la table 4. Ce catalogue montre toutes les interactions tridimensionnelles de ces structures, et constitue une première étape d'un travail d'extraction de l'information structurale de l'ARN qui passera par l'analyse de ces diagrammes. J'ai pu, à ce jour, extraire quelques informations de certaines molécules que nous présentons ci-après, mais une grande partie du travail d'analyse reste à réaliser.

Structures		Organisme	PDB id	Réso. (Å)
Introns groupe I	avec exons	*Azoarcus*	1U6B	3,10
	ribozyme	*Tetrahymena*	1X8W	3,80
	ribozyme/produit	*Twort*	1Y0Q	3,60
ARN RNaseP Type A	Dom. Spéc.	*T. thermophilus*	1US9	2,90
	Entier	*T. maritima*		3,85
ARN RNaseP Type B	Dom. Spéc.	*B. subtilis*	1NBS	3,15
	Entier	*B. stearothermophilus*		3,30
Guanine riboswitch	complexé avec hypoxanthine	*B. subtilis*	1U8D	1,95
	complexé avec guanine		1Y27	2,40
	complexé avec adénine		1Y26	2,10
ARNr 16S		*T. thermophilus*	1J5E	3,05
ARNr 23S	Site de transpeptidation	*H. marismortui*	1S72	2,40
Ribozyme HDV			1DRZ	2,30
Ribozyme Diels-Alder		ARN sélectionné	1YLS	3,00
Ribozyme épingle à cheveux		ARN viroïdes	1HP6	2,40

Table 3 : Structures cristallographiques dont les diagrammes de réseau d'interactions ont été réalisés.

Sur la figure 12 sont représentés les diagrammes de trois ribozymes (ribozyme diels-alder, ribozyme du virus de l'hépatite δ et ribozyme en épingle à cheveux), et du riboswitch répondant à la guanine (Ferre-D'Amare et al., 1998; Rupert & Ferre-D'Amare, 2001; Mandal et al., 2004; Serganov et al., 2005). Les

ribozymes Diels-Alder et HDV présentent deux pseudonoeuds et un double pseudonoeud respectivement. Deux adénines de HDV réalisent des interactions en A mineur que nous avons vu précédemment impliquées dans différents motifs d'interactions à longue distance et qui seront décrites dans le chapitre suivant. Lors de la publication de la structure du riboswitch, nous avons eu l'opportunité de présenter une analyse de la 3D, à l'aide de la représentation en diagramme de réseau, dans un preview (voir revue 3 dans les annexes). Nous avons montré, entre autres, que les deux paires de bases *cis* Watson-Crick/Sucre, qui lient deux nucléotides de la jonction J1/2 et les deux premières paires de base Watson-Crick de l'hélice P1, sont présentes dans un motif C like de l'ARNr 16S. Ces deux paires de bases sont donc à la fois impliquées dans un motif compris dans une hélice et dans le rapprochement d'éléments périphériques. Elles viennent donc s'ajouter aux interactions à longue distance décrites dans le chapitre précédent. Le "riboswitch" présente également une jonction à trois hélices qui a été incluse dans l'étude de classification des jonctions triples du chapitre 2.4.

Le ribozyme en épingle à cheveux présente une jonction à quatre hélices sans nucléotide libre entre les hélices qui joue un rôle fondamental dans l'activité ribozymique (voir chapitre 2.3.1). Les nucléotides en positions 24 et 25 de l'hélice B, et 10 et 11 de l'hélice A interagissent par l'intermédiaire de leur côté Sucre et forme un motif « ribose zipper » qui sera décrit dans le prochain chapitre 2.1.4.

Les diagrammes des trois structures cristallographiques d'introns de groupe I résolues récemment, *Azoarcus*, *Twort* et *Tetrahymena*, sont présentés figure 13 et montrent un certain nombre de similarités (Adams et al., 2004b; Guo et al., 2004; Golden et al., 2005).

Tout d'abord, une conformation identique du site de fixation de la poche fixant la guanine en 3' de l'intron a été identifiée. La guanine en 3' de l'intron, dans les trois cas, est prise en tenaille entre les nucléotides 2 et 3 de la jonction simple brin J6/7 et forme une paire *cis* Watson-Crick/Hoogsteen avec la guanine de la dernière paire G=C de l'hélice P7. Le deuxième nucléotide de J6/7 forme un appariement triple avec la paire Watson-Crick empilée sur l'hélice P7. La poche de fixation constitue ainsi dans les trois structures un réseau complexe de bases empilées et/ou impliquées dans des appariements triples.

Ensuite, bien que les éléments périphériques au cœur catalytique soient différents dans les trois structures cristallographiques, les interactions sont

conservées. C'est le cas de l'interaction entre la boucle GAAA qui ferme P9 et la jonction simple brin J5/5a. P9 appartient à une jonction à trois hélices (*Twort*) ou est situé après un coude qui la sépare de P9.0 (*Azoarcus* et *Tetrahymena*). GAAA interagit avec un récepteur à 11 nucléotides dans le cas de *Azoarcus* ou avec deux paires Watson-Crick dans une hélice dans le cas de *Tetrahymena* ou *Twort*. Dans tous les cas, les deux dernières adénines de la boucle de P9 interagissent avec le petit sillon des paires de bases réceptrices.

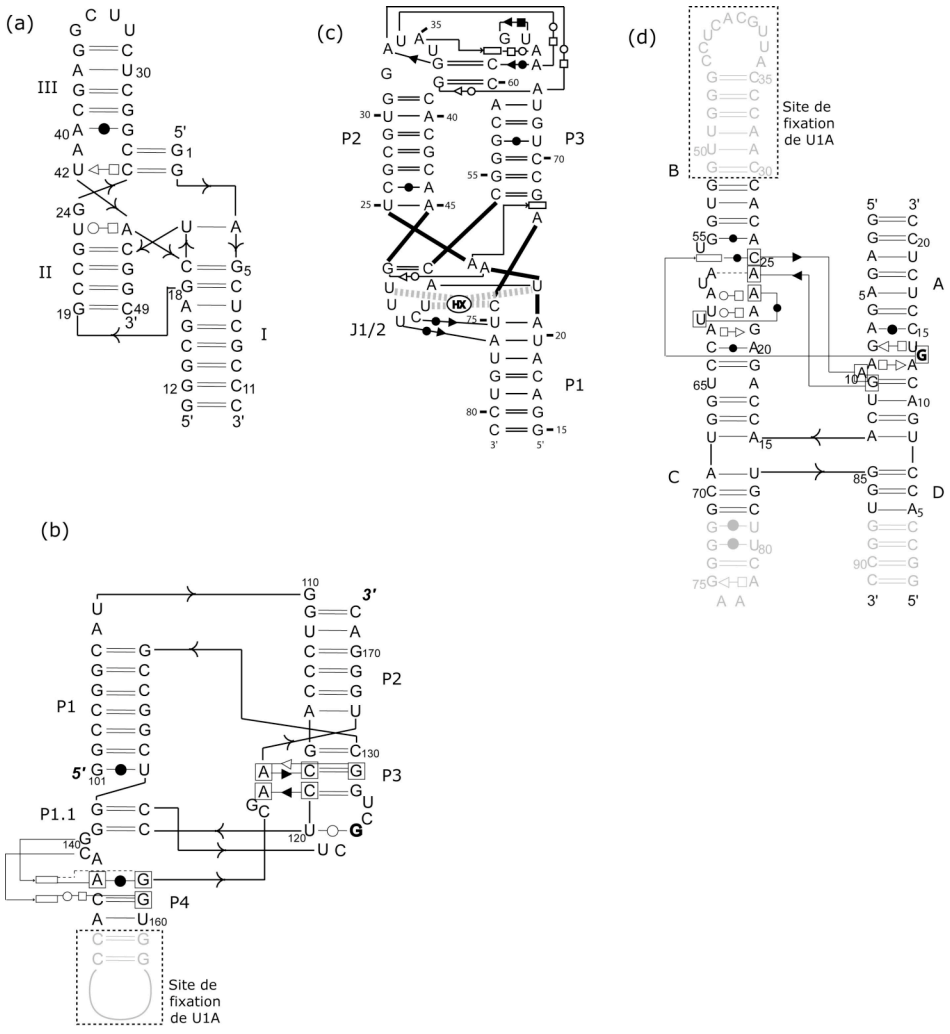


Figure 12 : Diagrammes des structures cristallographiques des ribozymes et d'un riboswitch. (a) Ribozyme Diels-Alder (PDB : 1YLS) (Serganov et al., 2005). (b) Ribozyme du virus de l'hépatite δ (PDB : 1DRZ) (Ferre-D'Amare et al., 1998). (c) Riboswitch répondant à la guanine (PDB :avec guanine 1U8D, avec adénine 1Y26, avec hypoxanthine 1U8D) (Mandal & Breaker, 2004; Mandal et al., 2004). (d) Ribozyme en épingle à cheveux (Pdb : 1HP6) (Rupert & Ferre-D'Amare, 2001).

Enfin, il faut noter, dans les deux structures contenant l'hélice P1 (*Azoarcus* et *Twort*), les interactions entre deux nucléotides de J8/7 et le petit sillon de l'hélice P1. Dans la structure de *Tetrahymena*, où l'hélice P1 est absente, ces mêmes nucléotides sont orientés en *syn* et n'interagissent avec aucun autre élément de la structure. Un changement de conformation de ces nucléotides de *syn* en *anti*, permettant une interaction avec P1 et son immobilisation, pourrait être envisagé.

La complexité des trois structures d'intron de groupe I est importante et leur représentation bidimensionnelle en facilitera l'analyse comme suggéré dans les trois exemples précédents.

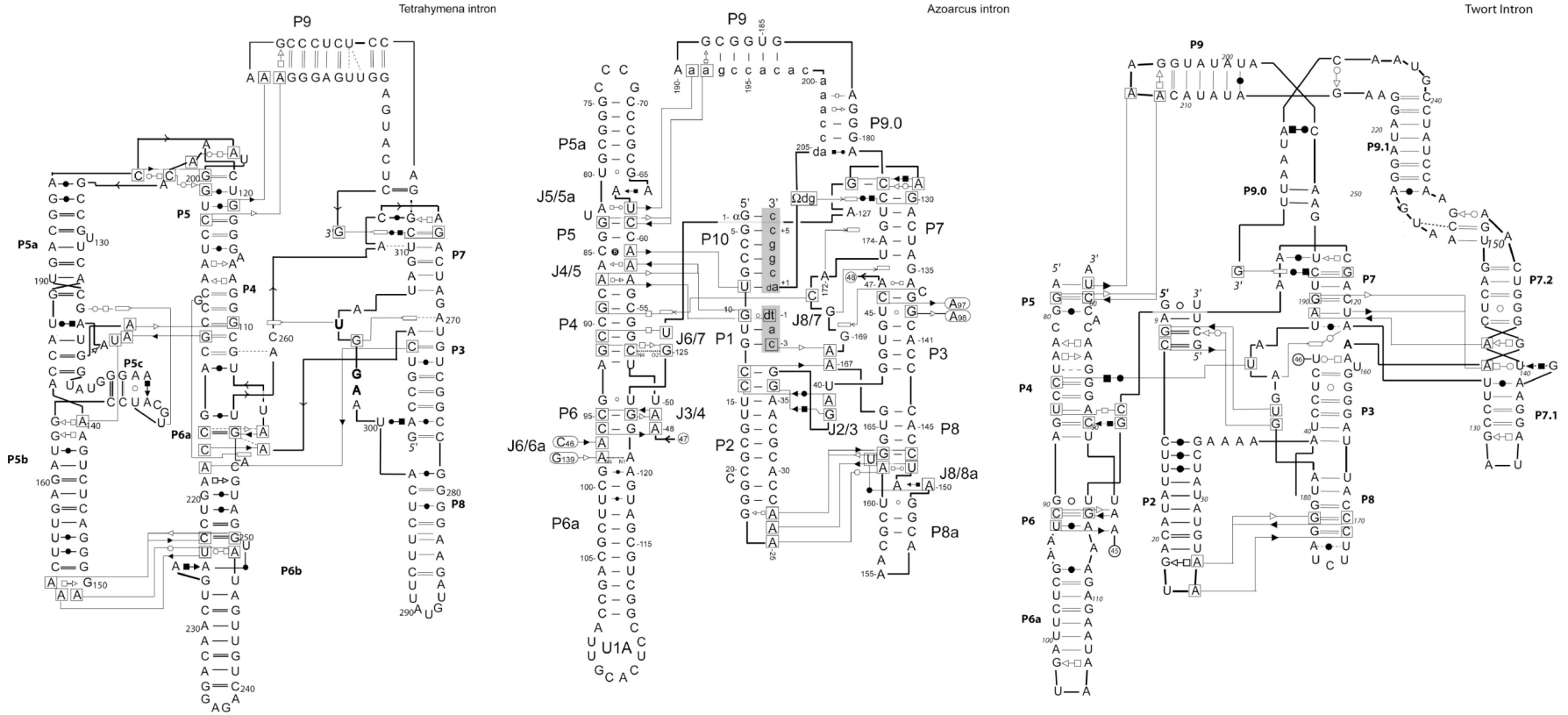


Figure 13 : Diagrammes des réseaux d'interactions des structures cristallographiques des introns de groupe I. A gauche Intron de *Tetrahymena* sans substrat (Guo et al., 2004). Au milieu : Intron de *Azoarcus* complexé aux exons 5' et 3' formant P1 et P10 (Brion & Westhof, 1997; Adams et al., 2004b). La tige-boucle qui fixe la protéine U1A a été ajouté à P6a pour faciliter la cristallisation. A droite : Intron de *Twort* lié à l'exon 5' coupé formant l'hélice P1 (Golden et al., 2005).

La figure 14 montre les diagrammes des domaines de reconnaissance des RNase P de type A et B. Les éléments structuraux périphériques varient d'un type à l'autre mais l'architecture globale de la molécule est maintenue. Ainsi, bien que la région L11/12 appartienne à la jonction entre les hélices P11, P12, P13 et P14 dans le type A, et forme simplement une bulle interne dans le type B, elle se replie, dans les deux cas, en deux motifs boucle T. D'autre part, tout comme les introns de groupe I, les interactions à longue distance entre éléments périphériques varient mais assurent toujours une conservation de l'architecture. Le type B comporte un motif GAAA-récepteur à 11 nucléotides entre P12 et P10.1 et un motif A mineur entre la boucle terminale de P10.1 et P10, tandis que le type A comporte deux motifs A mineur entre P14 et P8 et entre P13 et P12. La structure de l'ARN sans le domaine de reconnaissance de la RNase P de *Bacillus stearothermophilus* a été résolue en 2005 et montre de nombreux contacts à longue distance très intéressants (Figure 15). Les deux jonctions à trois hélices P5/P5.1/P7 et P7/P10.1/P10-11 seront discutées dans le chapitre sur la classification des jonctions triples. La structure de l'ARN de la RNase P de *Thermotoga maritima* a été également résolue mais seule l'information de positionnement des hélices est disponible (Figure 15). Les régions en simple brin sont très désordonnées et n'ont pas pu être observées. Le diagramme de cette structure montre que la boucle terminale de l'hélice P8 a été affinée avec une première adénine en *syn* alors que l'adénine suivante réalise une interaction en motif A mineur avec l'hélice P4 (*cis* Sucre/Sucre impliquant une adénine). Il est fort probable, au vu de la forte présence de motif en A mineur dans les interactions à longues distances de la majorité des structures d'ARN, que ces deux adénines consécutives forment un tel motif. Il faudrait essayer d'affiner la structure en orientant la première adénine en ANTI et regarder si elle peut interagir par son côté Sucre avec l'appariement C=G voisin de la paire contactée par la deuxième adénine. Ainsi, la représentation d'une structure cristallographique en diagramme de réseau d'interactions peut révéler des originalités dues à une inexactitude d'affinement et suggérer de nouvelles solutions de construction.

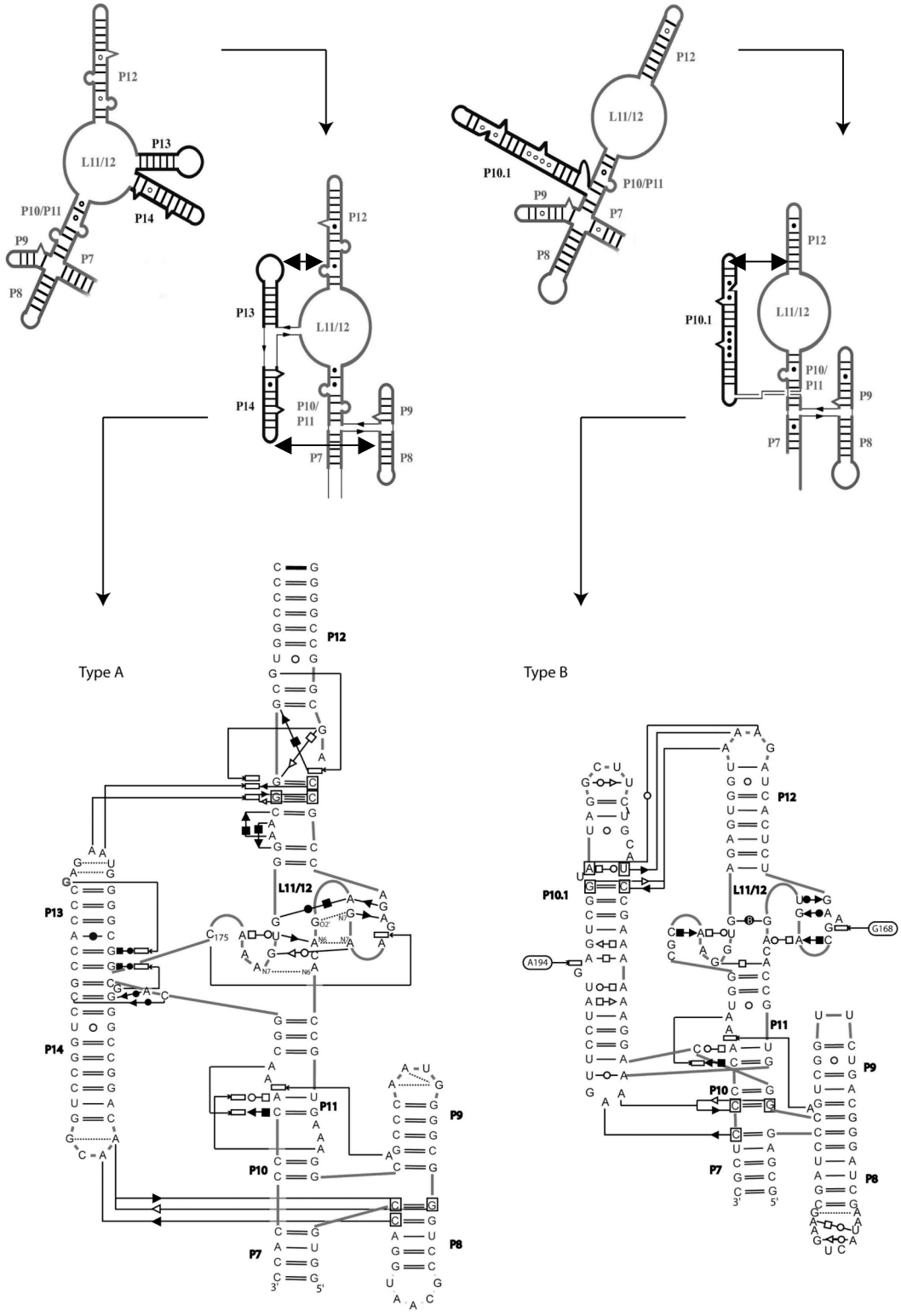


Figure 14 : Représentations du domaine de reconnaissance (domaine I) des RNases P de types A (à gauche) et B (à droite). L'analyse des covariations Watson-Crick d'un alignement de séquences homologues a permis de représenter la structure 2D des domaines (en haut). La modélisation moléculaire a permis de représenter les domaines en rendant compte de l'arrangement dans l'espace des hélices, de leur empilement et de certaines interactions ARN-ARN (au milieu) (Massire et al., 1998). Finalement, les structures cristallographiques des deux domaines ont mené à la représentation en diagrame des réseaux d'interactions donnant toutes les informations des interactions tridimensionnelles de ces structures (en bas) (Krasilnikov et al., 2003; Krasilnikov et al., 2004).

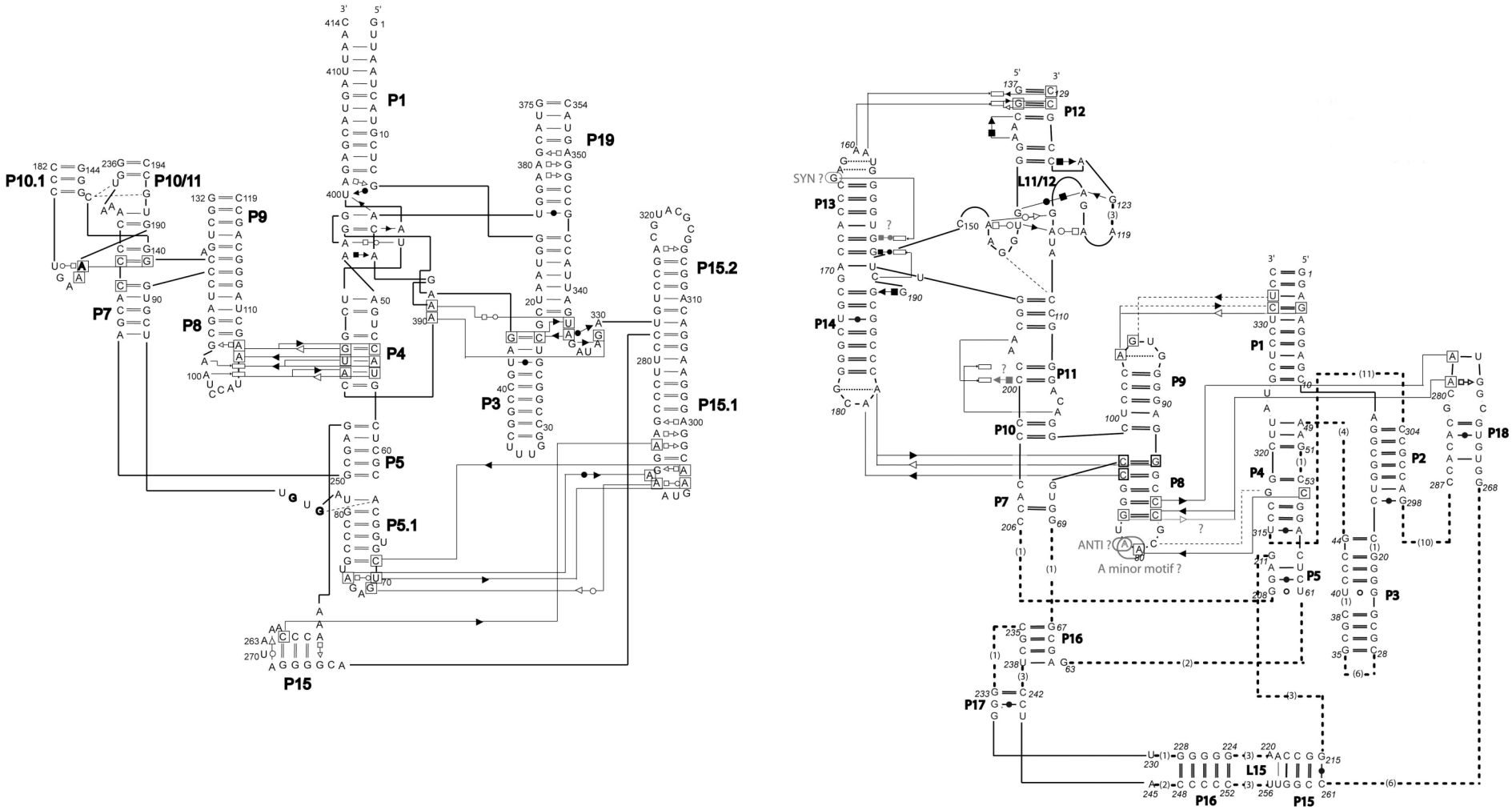


Figure 15: Diagrammes des RNases P de B.stearothermophilus (à gauche) et T.maritima (à droite).

Pour terminer, les diagrammes d'interactions du domaine de transpeptidation de l'ARNr 23S de *Haloarcula marismortui* et de l'ARNr 16S de *Thermus thermophilus* sont présentés sur les figures 16 et 17 (Ban et al., 2000; Wimberly et al., 2000). La figure 16 montre que le domaine de transpeptidation est organisé autour de la jonction à trois hélices H90H91H92 qui sera décrite dans le chapitre sur la classification des jonctions triples. Dans le diagramme de l'ARNr 16S, nous avons choisi de mettre en évidence les contacts interdomaines qui sont, de manière plutôt inattendue, très peu nombreux. La quantité d'informations contenues dans ces structures d'ARNr est évidemment très importante et quoique très étudiées déjà depuis la parution des structures en 2000, nous pensons que l'analyse des diagrammes révèlera de nouveaux éléments sur la structure de ces ARN.

D'une manière générale nous n'avons pas encore exploité en profondeur tous ces diagrammes des réseaux d'interactions. Au vu des exemples sus-cités, il est évident qu'à terme ces diagrammes permettront de mieux comprendre la structure de chaque molécule d'ARN. De plus, la comparaison des diagrammes des différentes structures révèlera les points communs et les différences d'interactions, de motifs, qui seront exploitables pour la prédiction de structures d'ARN.

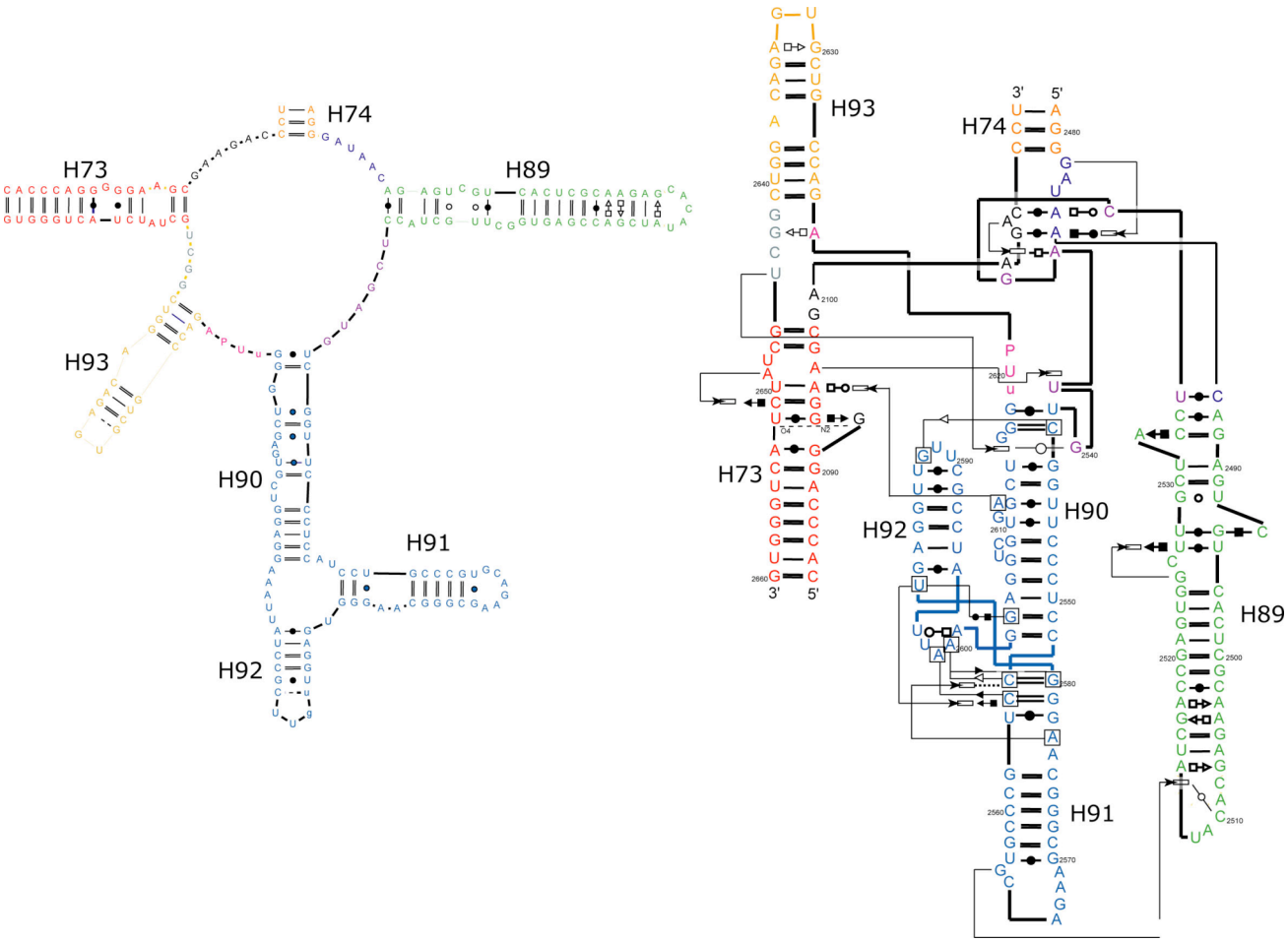


Figure 16: Site de transpeptidation de l'ARNr 23S d'*Haloarcula marismortui* (Ban et al., 2000). A gauche la représentation classique du site ; à droite le diagramme de réseau d'interactions qui montre comment l'architecture du site est organisée autour de la jonction à trois hélices H90H91H92.

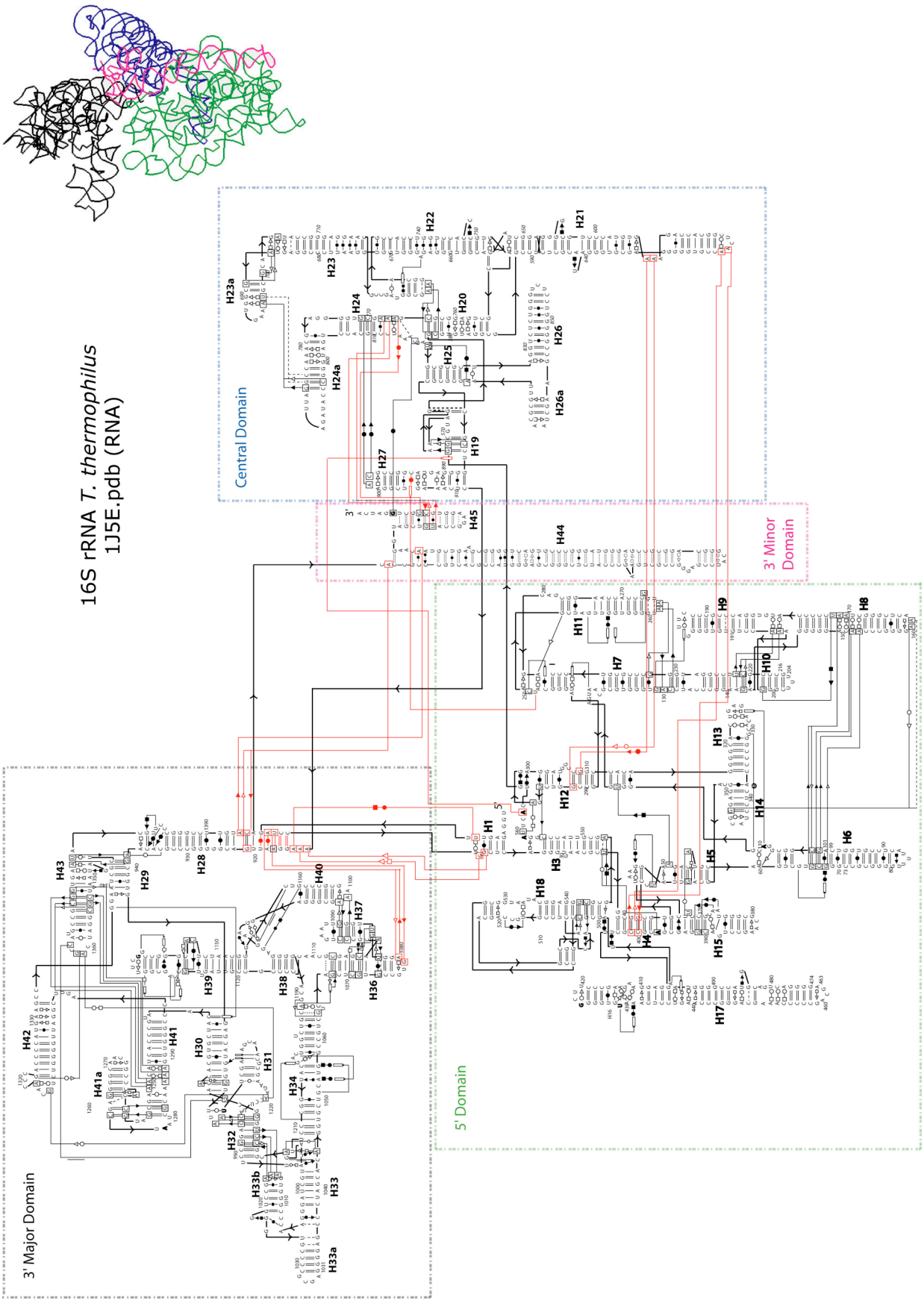


Figure 17: Diagramme de l'ARNr 16S de *Thermus thermophilus* (Wimberly et al., 2000). En haut à droite est représentée la structure cristallographique de l'ARNr 16S. Les domaines sont colorés en accord avec les cadres du diagramme. Les interactions interdomaines sont indiquées en rouge sur le diagramme. Noter le petit nombre d'interactions inter-domaines comparé aux interactions intra-domaines.

2.1.4. Motifs en A mineur

L'activité d'un ARN dépend de sa structure tridimensionnelle, c'est à dire de l'organisation dans l'espace des éléments de structure secondaire maintenus proches les uns des autres par des interactions à longue distance. Comme nous l'avons vu, une majorité des interactions à longue distance fait intervenir les côtés Sucre des bases.

Les appariements Sucre/Sucre (deux nucléotides interagissent par leur côté Sucre) présentent, comme toutes les autres familles d'appariements, deux orientations *cis* et *trans* des liaisons glycosidiques (Figure 18). Pour chaque orientation une matrice d'isostérie qui montre tous les appariements Sucre/Sucre, a été définie (Figure 19.A) (Leontis et al., 2002b). Dans la matrice de l'appariement *trans* Sucre/Sucre seules sont observées les paires dont la base réceptrice du 2'OH, est une purine. Les paires dans lesquelles une adénine ou une guanine est réceptrice, appartiennent à la famille d'isostérie I1 ou I2 respectivement. Par contre, toutes les combinaisons d'appariements *cis* Sucre/Sucre entre les quatre bases sont obervées sauf GoG, et tous les appariements appartiennent à la même famille isostérique I.

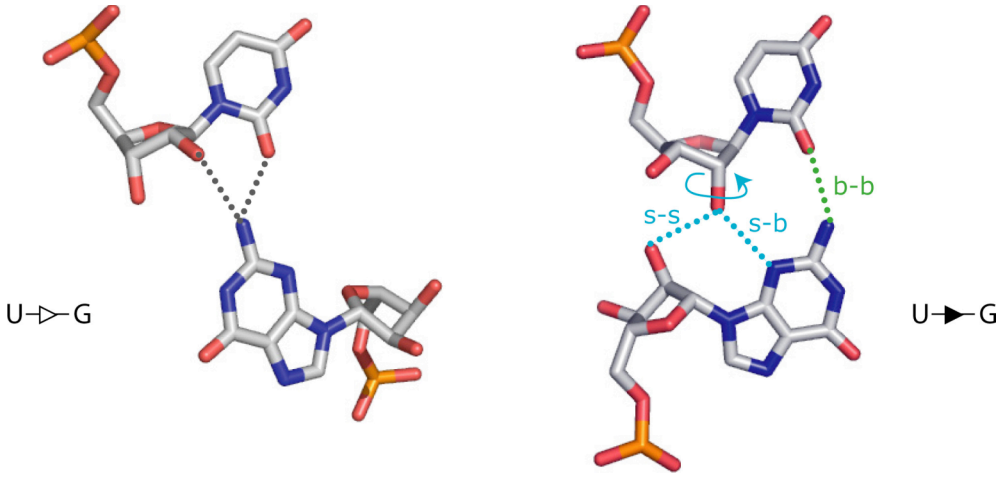


Figure 18 : Deux exemples d'appariement Sucre/Sucre, l'un orienté en *trans* (à gauche) et l'autre en *cis* (à droite). A droite sont indiquées les interactions entre un atome de chaque base (b-b) en vert, le 2'OH d'un nucléotide, et le 2'OH et un atome de la base de l'autre nucléotide en bleu (s-s et s-b). Dans le cas où une cytosine ou une uracile remplace la guanine, seules les laisons s-s et s-b restent et la pyrimidine peut tourner autour de la liaison 2'OH (flèche).

Globalement, dans un appariement *cis* Sucre/Sucre, trois types de contacts entre les faces Sucre des deux nucléotides peuvent être impliqués : contact sucre-sucre (s-s) qui n'implique que les 2'OH des riboses, sucre-base (s-b) qui implique un atome du ribose d'un nucléotide et un ou des atomes de la base de l'autre nucléotide et enfin contact base-base (b-b) qui implique des atomes des deux bases (Figure 18). Dans l'exemple de la paire UoG (Figure 18), tous les contacts sont présents, mais, en remplaçant la guanine par une pyrimidine, le contact b-b est perdu. Lorsqu'une interaction entre deux nucléotides n'implique que des contacts s-s ou s-b, c'est à dire sans contact entre les deux bases, nous appelons l'interaction de type « ribose zipper ». Les différentes associations possibles d'interactions de type « ribose zipper » réalisées par deux nucléotides consécutifs, forment des motifs « ribose zipper » comme défini d'abord par Cate et collaborateurs puis formalisé par Tamura et Holbrook (2002) (Tamura & Holbrook, 2002). Ces derniers ont proposé une classification des « ribose zipper » en fonction de la combinaison des deux interactions « ribose zipper » formées par les deux nucléotides consécutifs (Figure 19). En tout ce sont onze motifs « ribose zipper » que ces auteurs proposent mais seuls sept ont été observés au sein des structures cristallographiques disponibles. Parmi ces sept motifs, nous avons trouvé que cinq d'entre eux sont des dérivés, où l'une ou l'autre liaison hydrogène manque, de deux motifs principaux nommés motif zipper canonique et motif ribose zipper cis. L'appariement de type *cis* Sucre/Sucre est le seul à composer les deux motifs ribose zipper principaux. Dans les deux motifs, chacun des deux nucléotides consécutifs réalise une interaction de type *cis* Sucre/Sucre mais c'est l'orientation d'un des appariements qui varie de l'un à l'autre (Figure 19).

Une interaction « ribose zipper » ne possède pas de géométrie propre puisque les deux nucléotides peuvent « tourner » autour de la liaison C2'-O2' et présenter plusieurs orientations (Figure 18). Par contre, certaines interactions présenteront un contact b-b comme vu plus haut. Ce contact impose une spécificité et restreint la rotation autour du 2'OH d'un des nucléotides. Il apparaît alors nécessaire de clarifier la matrice d'isostérie basée sur les seuls côtés Sucre qui se font face (Figure 20.A) car la présence d'un contact base-base dépend de la nature des bases partenaires. La Figure 20.B montre les nouvelles matrices où les appariements ne présentant pas de contact base-base sont annotés « RZ » tandis que ceux montrant un contact base-base sont indiqués « S ». Les

appariements S, qui montrent une géométrie caractéristique, une matrice d'isostérie spécifique et donc des règles de covariations, peuvent être identifiés dans un alignement de séquences homologues.

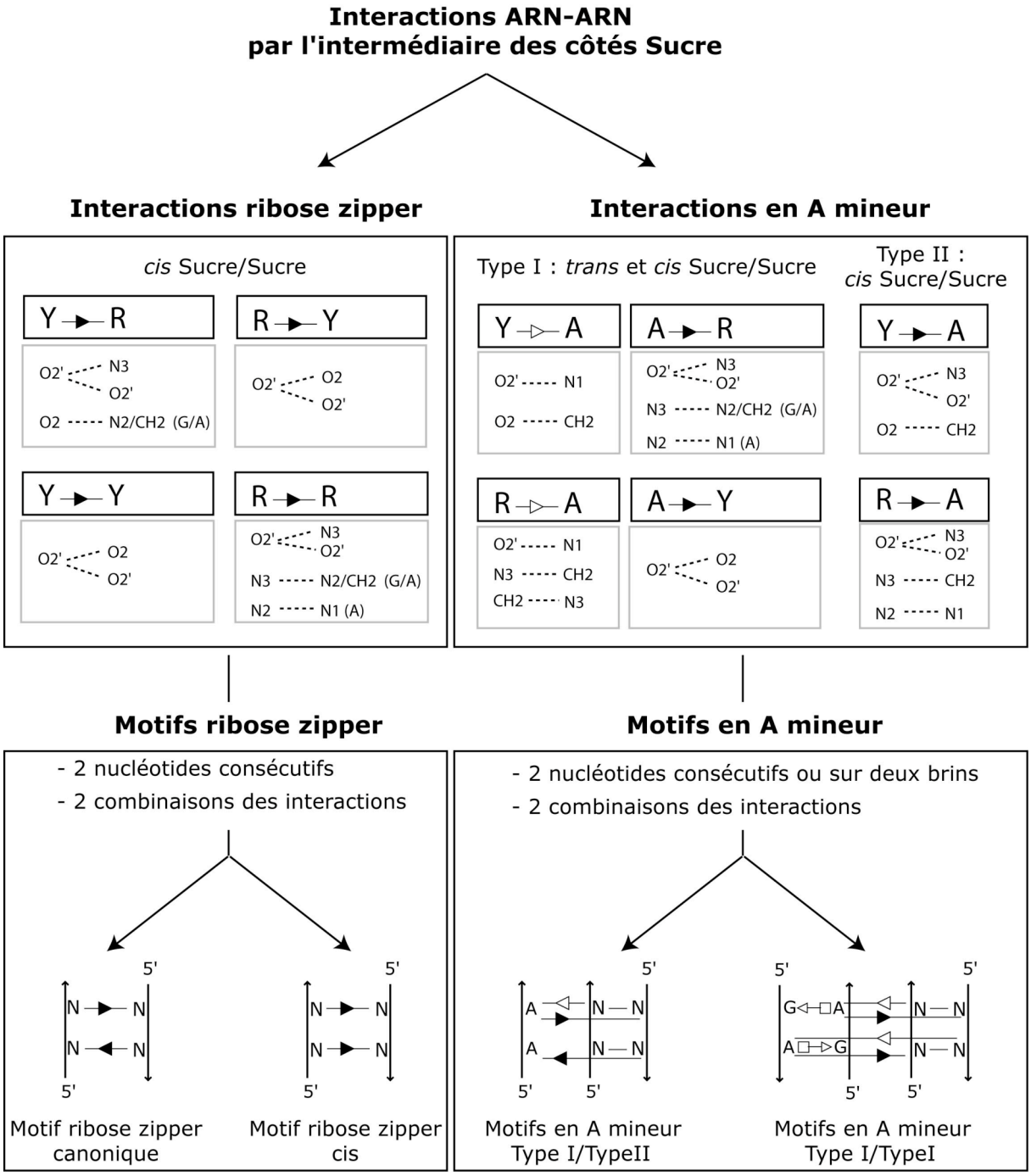


Figure 19: Décomposition des contacts ARN-ARN par l'intermédiaire des côtés sucre des nucléotides en protocoles d'interaction et en motifs résultant de leur assemblage.

Le motif d'interaction à longue distance par l'intermédiaire des côtés Sucre le plus répandu parmi les structures d'ARN connues implique deux adénines. Il a été analysé pour la première fois par Nissen et collaborateur qui l'ont appelé "motif en A mineur" (Nissen et al., 2001). Le motif a été nommé ainsi car il implique l'interaction d'adénines avec le sillon mineur (ou petit sillon) d'une hélice. Il avait été mis en évidence auparavant par Michel et Westhof (1990) dans leur modèle d'intron de groupe I (Michel & Westhof, 1990). Sur la base d'analyse de séquences et des observations faites lors de la construction du modèle, il est apparu que les GNRA présentes étaient proches d'hélices Watson-Crick et que les derniers résidus (RA) pouvaient interagir avec deux paires de bases empilées R-Y et G=C de l'hélice avoisinante.

(A)

▶	A	C	G	U
A	I	I	I	I
C	I	I	I	I
G	I	I		I
U	I	I	I	I

▷	A	C	G	U
A	I1		I2	
C	I1		I2	
G	I1		I2	
U	I1		I2	

(B)

▶	A	C	G	U
A	S	RZ	S	RZ
C	S	RZ	S	RZ
G	S	RZ		RZ
U	S	RZ	S	RZ

▷	A	C	G	U
A	S		S	
C	S		S	
G	S		S	
U	S		S	

Figure 20 : Matrices d'isostérie des appariements *cis* et *trans* Sucre/Sucre.
(A) Matrices d'isostérie générale des appariements impliquant les côtés Sucre des bases. I1 et I2 ne sont pas strictement isostériques(B) Matrices d'isostérie indiquant les appariements spécifiques (s) possédant une liaison b-b et non spécifique (RZ) ne possédant pas de liaison b-b.

D'après la modélisation, seules des interactions par l'intermédiaire des côtés Sucre des bases des deux partenaires pouvaient avoir lieu. Ces interactions ont été confirmées plus tard par empreintes chimiques (Murphy & Cech, 1993, 1994) et mutations (Jaeger et al., 1994) puis finalement observées dans les structures du ribozyme à tête de marteau (Pley et al., 1994b, 1994a) et du

domaine P4P6 de l'intron de groupe I de *Tetrahymena* (Cate et al., 1996a; Cate et al., 1996b). Ces dernières années les structures du ribosome, des RNase P et des introns de groupe I ont établi l'universalité et l'importance du motif en A mineur. Ainsi, la structure cristallographique de l'ARNr 16S de la petite sous-unité ribosomique 30S présente 55 interactions en A mineur observées ou potentielles (Wimberly et al., 2000; Noller et al., 2005).

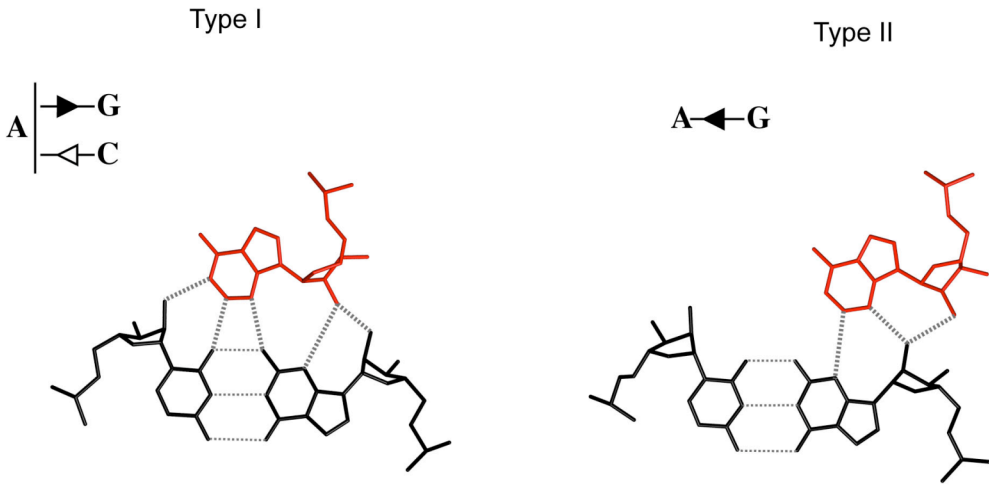


Figure 21 : Exemples d'interactions en A mineur de types I et II. Chaque type est défini par la position du groupement 2'OH de l'adénosine du motif et des groupements 2'OH des paires de bases réceptrices. Dans le type I, l'adénine forme deux interactions sucre/sucre *cis* et *trans* avec les bases de la paire WC. Dans le type II, l'adénine forme une interaction *cis* sucre/sucre mais dans ce cas l'adénine est "réceptrice" du groupement hydroxyle de la guanine impliquée dans l'appariement WC.

Le motif en A mineur est formé par les contacts entre les faces Sucre de deux adénines et les faces Sucre des bases impliquées dans deux appariements Watson-Crick consécutifs. On peut distinguer quatre types d'interactions (de 0 à III) mais seuls les types I et II impliquent exclusivement des adénines (Nissen et al., 2001). Sur la figure 21, un exemple des interactions de types I et II est représenté. L'interaction de type I implique (i) une paire *cis* Sucre/Sucre dans laquelle A (2' OH) contacte G (2'OH et N3) tandis que A (N3) contacte G (N2) et (ii) une paire *trans* Sucre/Sucre dans laquelle A (N1) contacte C (2'OH et O2). L'interaction de type II consiste en une paire *cis* Sucre/Sucre dans laquelle le N3 et le 2'OH de l'adénine sont contactés par le 2'OH de la guanine de la deuxième paire. Dans le motif en A mineur typeI/typeII impliquant deux adénines consécutives, l'adénine en 3' (type I) interagit avec les deux brins de l'hélice

récepteur tandis que l'adénine en 5' (type II) interagit avec un seul brin. L'adénine en 3' forme plus de liaisons hydrogène que l'adénine en 5'.

Selon l'environnement dans lequel il se trouve, le motif en A mineur formé par la combinaison d'une interaction de type I et d'une interaction de type II possède trois degrés de spécificité ; toutefois la nature des interactions (types I et II) reste inchangée. Les trois degrés sont décrits ci-après dans un ordre de spécificité décroissante.

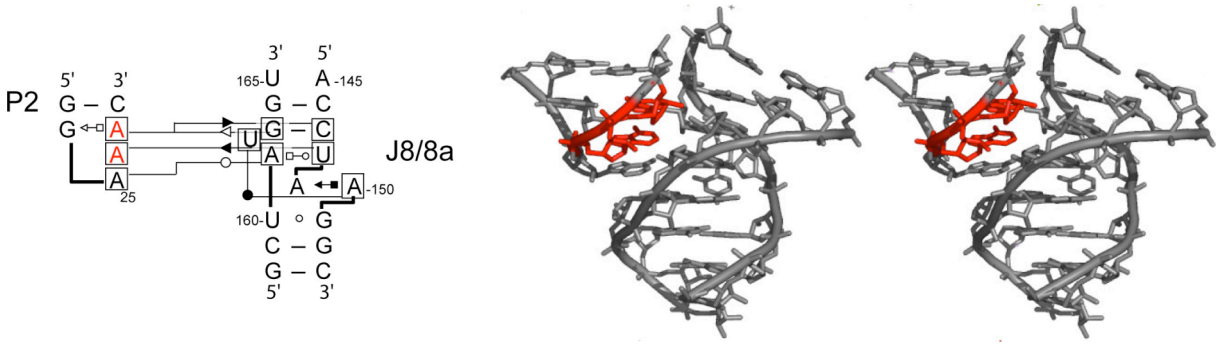


Figure 22 : Motif GAAA/Récepteur à 11 nucléotides P2/J8/8a de l'intron I de *Azoarcus*. A gauche la structure bidimensionnelle montrant les interactions canoniques et non canoniques et à droite la représentation 3D du motif extrait de la structure cristallographique (PDB id : 1U6B) (Adams et al., 2004b). Les deux adénines impliquées dans le motif en A mineur sont représentées en rouge.

Premièrement, les deux adénines aux positions 3 et 4 d'une boucle GAAA interagiront avec un motif récepteur à onze nucléotides séquence spécifique. Ce type d'interaction, d'abord suggéré par SELEX (Costa & Michel, 1997), a été observé pour la première fois dans la structure cristallographique du domaine P4P6 de l'intron de groupe I de *Tetrahymena* (Cate et al., 1996a). Il est également présent dans la structure de l'intron de groupe I de *Azoarcus* entre la boucle terminale P2 GAAA et la bulle interne J8/8a structurée en récepteur à onze nucléotides (Adams et al., 2004b) (Figure 22). Le récepteur à onze nucléotides présente trois paires de bases non canoniques : une paire UoA *trans* Watson-Crick/Hoogsteen, une paire *cis* Watson-Crick/Watson-Crick entre une uridine en bulge et une adénine impliquée dans la paire AoA *cis* Hoogsteen/Sucre formant une plateforme d'adénines. La paire entre l'uridine en bulge et l'une des adénines de la plateforme peut être absente, ce qui est le cas dans le récepteur à onze nucléotides en position J5/5a interagissant avec la boucle terminale GAAA

de P9. Dans le cas de l'interaction GAAA/récepteur, une paire *trans* Watson-Crick/Watson-Crick entre la deuxième adénine de la boucle et l'adénine de la paire *trans* Watson-Crick/Hoogsteen, s'ajoute aux interactions du motif en A mineur.

Deuxièmement, deux adénines, ou, une guanine et une adénine aux positions 3 et 4 d'une boucle GNRA interagiront avec deux paires G=C ou une paire G=C et une paire A-U respectivement (Figure 23 à droite). Les types d'interaction sont conservés mais dans ce cas l'adénine en 3ème position peut être remplacée par une guanine. La Figure 23 montre qu'une guanine en position 3 interagira en *cis* Sucre/Sucre avec l'uracile d'une paire A-U tandis qu'une adénine à cette même position interagira avec la cytosine d'une paire G=C. De telles variations ont été observées dans les séquences homologues d'intron de groupe I et ont permis de mettre en évidence ce type d'appariement et de le modéliser (Michel & Westhof, 1990). Le motif a été observé pour la première fois dans les contacts intermoléculaires du cristal du ribozyme à tête de marteau (Pley et al., 1994a).

Et enfin, troisièmement, deux adénines extrudées, dont l'environnement structural est variable, interagiront avec deux paires canoniques consécutives quelles qu'elles soient, selon les études de Nissen et collaborateurs (Nissen et al., 2001).

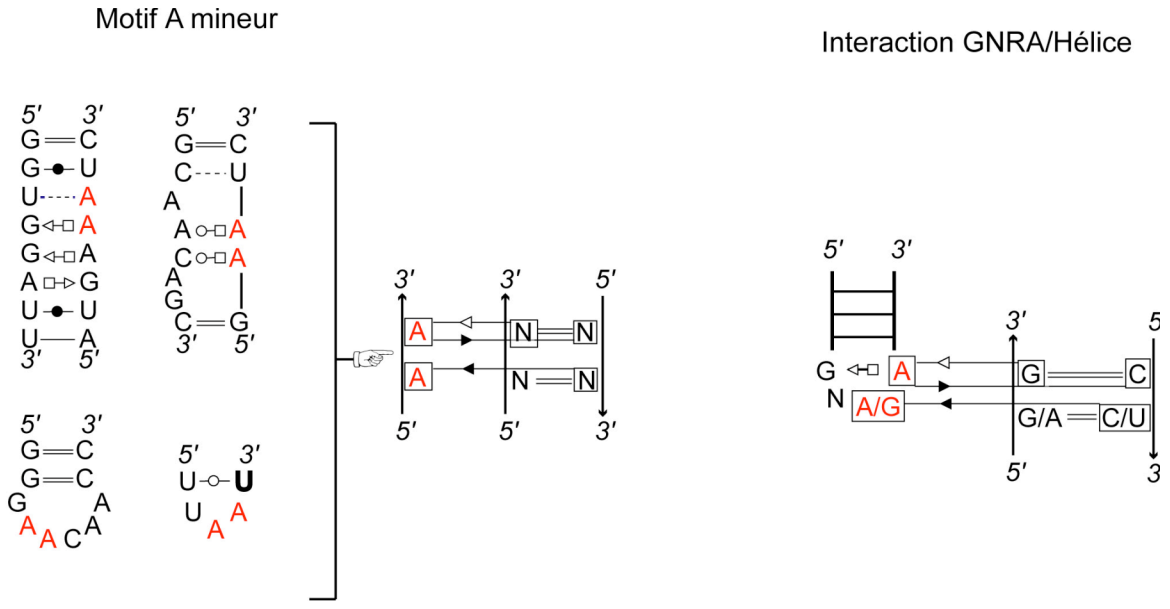


Figure 23 : Motif A mineur. Deux adénines dans une boucle interne ou boucle terminale forment un motif A mineur avec deux paires canoniques adjacentes (à gauche). Deux adénines ou une guanine et une adénine forment un motif A mineur avec des paires GC ou AU et GC (à droite).

Des études de mutations d'une paire Watson-Crick ont montré que les interactions entre adénines et paires canoniques étaient énergétiquement plus favorables que les interactions avec des paires non Watson-Crick mais que le type d'interaction Watson-Crick n'avait pas d'influence sur la capacité des adénines à interagir (Battle & Doudna, 2002). Toutefois, il apparaît également qu'au sein de l'ARNr 23S de *H.marismortui*, les motifs en A mineur ont une préférence nette pour les récepteurs à paires C=G ce qui corrobore une analyse montrant une fréquence très importante de récepteurs de paires de bases C=G, C=G, dans les ARNr 16S et 23S (Tamura & Holbrook, 2002).

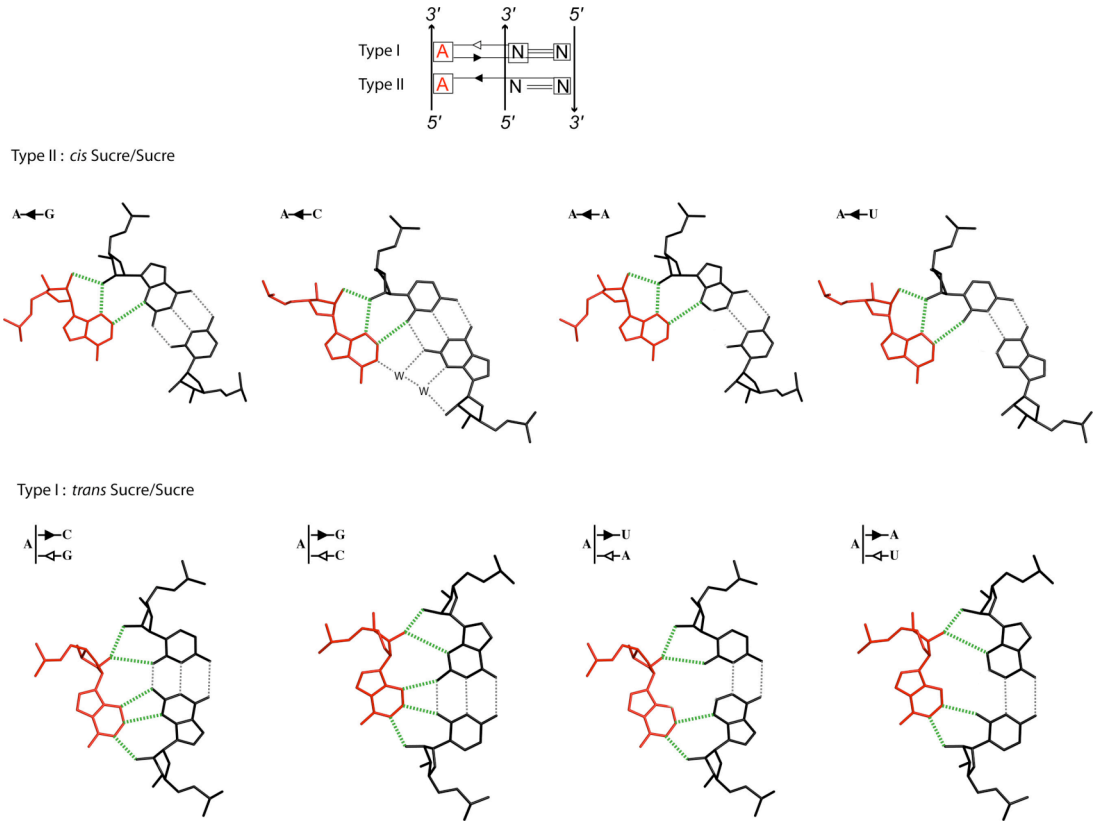


Figure 24 : Interactions des adénines d'un motif en A mineur typeI/typeII avec les paires Watson-Crick canoniques. Dans la plupart des structures cristallographiques, des liaisons hydrogènes additionnelles de la paire AoC *cis* Sucre/Sucre sont médiées par des molécules d'eau. Les interactions de type I avec des paires G=C ou C=G montrent cinq liaisons hydrogène au lieu de quatre avec A-U ou U-A.

Quelques hypothèses pourraient expliquer cette préférence. Tout d'abord, la comparaison des variations d'energie libre (ΔG) entre deux paires Watson-Crick consécutives montre que deux paires C=G consécutives sont très favorables thermodynamiquement (Turner, 1996). Ensuite le nombre de liaisons

hydrogène formées entre les adénines du motif en A mineur et des paires Watson-Crick impliquant des guanines et des cytosines est plus important que dans les interactions impliquant des adénines et des uraciles (Figure 24). Enfin, dans l'appariement de type II *cis* Sucre/Sucre AoC, des liaisons hydrogène médiées par des molécules d'eau liant le N1 de l'adénine et les N2 et 2'OH de la guanine sont souvent observées dans les structures cristallographiques, comme dans le complexe site A/paromomycine ou néomycine ou encore dans l'ARNr 23S de *H.marismortui* (Vicens & Westhof, 2001; Tamura & Holbrook, 2002; Francois et al., 2005). La préférence des A mineur pour un récepteur CG/CG dans la nature peut être comprise mais des expériences de mutations compensatoires des paires Watson-Crick réceptrices de A mineur au sein de P4P6 montrent que toutes les autres combinaisons de paires Watson-Crick sont susceptibles d'interagir avec des adénines extrudées et de former ainsi un A mineur.

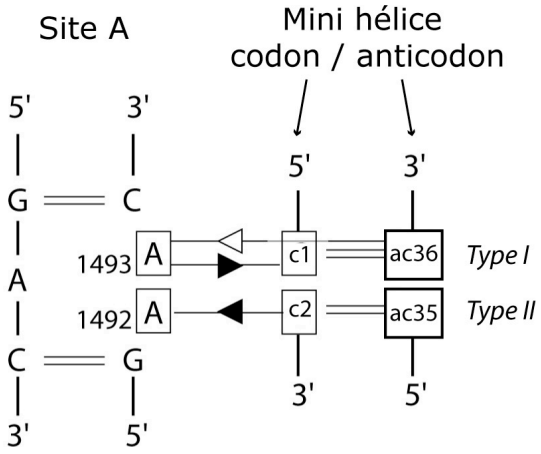


Figure 25 : Schéma du motif en A mineur impliqué dans le processus de décodage de la petite sous-unité ribosomique. Les adénines A1492 et A1493 du site A de l'ARNr 16S forment des interactions de type I et de type II respectivement avec la paire formée par la base en position 35 de l'anticodon et celle en position 2 du codon (c2) et une paire formée par la base en position 36 de l'anticodon et la base en position 1 du codon (Tinoco & Bustamante, 1999; Ogle et al., 2001;

Cette capacité des adénines du motif en A mineur à interagir avec n'importe quelle paire de bases Watson-Crick et non avec des paires non Watson-Crick est exploitée par le site A de la petite sous-unité ribosomique (Ogle et al., 2001). En effet, le site A est responsable du contrôle des appariements de la minihélice formée par les bases appartenant au codon et à l'anticodon et discrimine les appariements non Watson-Crick qui sont révélateurs d'une incohérence entre l'acide aminé codé par le codon et l'aminoacyl-ARNt apparié par son anticodon. Les adénines A1492 et A1493 du site A appartiennent à une boucle interne et se positionnent vers l'extérieur pour reconnaître, par interaction

avec son petit sillon, la minihélice formée par le codon de l'ARNm et l'anticodon de l'ARNt (Figure 25).

Le rôle des interactions en A mineur dans la synthèse protéique ne se limite pas à la validation de la cohérence codon/anticodon au cours du décodage. En effet, les adénines des extrémités CCA des aminoacyl-ARNt positionnées dans les sites A et P du ribosome interagissent avec des nucléotides du site de transpeptidation selon des interactions en A mineur (Figure 26) (Youngman et al., 2004).

Les motifs en A mineur pourraient avoir également une fonction dans les mouvements que réalise le ribosome lors de la traduction (Noller et al., 2005). Des études en cryoEM des états pré- et post-translocation montrent un mouvement de rotation de 6° entre les deux sous-unités ribosomiques, responsable d'un écart de 20 Å entre leurs extrémités. Un interrupteur moléculaire ("switch") au sein du ribosome, c'est à dire un changement dans la structure secondaire qui nécessiterait la rupture de nombreuses interactions et le repliement en une structure alternative présentant d'autres brins complémentaires, n'a pas été trouvé, certainement parce que la rupture d'une hélice demande une énergie d'activation élevée. La formation d'interactions alternatives doit être peu coûteuse d'un point de vue énergétique. Elle doit maintenir des conformations locales et aboutir à des architectures précises. Les motifs en A mineur, nombreux dans l'ensemble des structures ribosomiques et particulièrement impliqués, comme nous l'avons vu plus haut, dans la transpeptidation et le décodage, pourraient être impliqués dans la dynamique du ribosome au cours de la traduction.

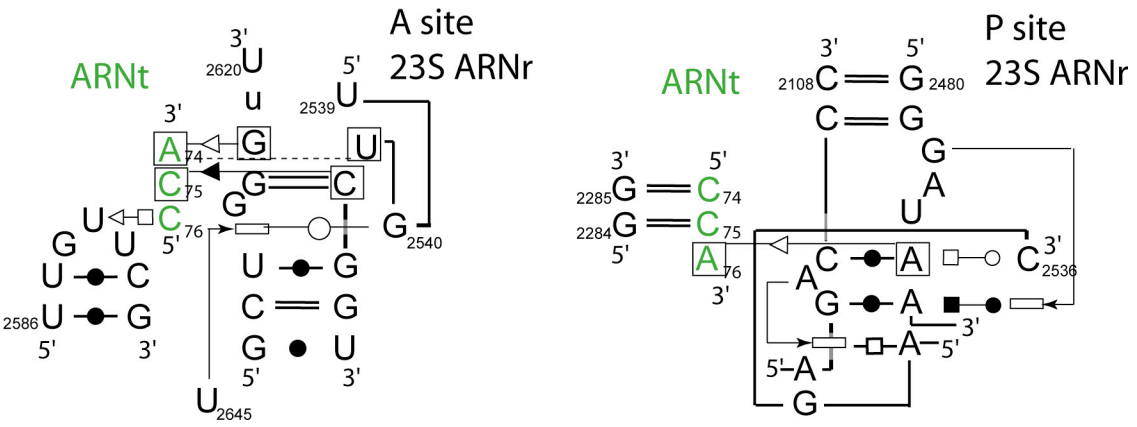


Figure 26 : Interaction des extrémités 3' CCA des ARNt avec les sites A et P du site de transpeptidation de l'ARNr 23S (Youngman et al., 2004).

Un motif en A mineur autre que ceux de type I/type II a été identifié dans les ARNr par Leontis et collaborateurs (Leontis et al., 2002a). Dans ce motif, que nous appelons motif en A mineur typeI/typeI, deux adénines de deux paires *trans* Hoogsteen/Sucre consécutives mais tête-bêche, encadrées par deux paires Watson-Crick, forme deux interactions de type I avec deux paires Watson-Crick consécutives (Figure 27). Le motif en A mineur typeI/typeI, tout comme le motif A type I/type II, pourrait être impliqué dans la dynamique du ribosome. En effet, la structure du ribosome d'*E.coli* montre que les hélices h44 de l'ARN 16S de la petite sous-unité ribosomique et H62 de la grande sous-unité ribosomique interagissent par l'intermédiaire de ce motif A typeI/typeI. L'absence ou la présence de ces interactions pourraient participer aux mouvements des sous-unités l'une par rapport à l'autre. D'un point de vue évolutif, le fait que deux adénines puissent interagir avec n'importe quelle paire Watson-Crick impose peu de contrainte. Le motif en A mineur serait ainsi facile à conserver au cours de l'évolution chez toutes les espèces, ce qui expliquerait les rôles clefs qu'il joue dans les mécanismes de décodage et de transpeptidation et peut-être dans la dynamique globale du ribosome.

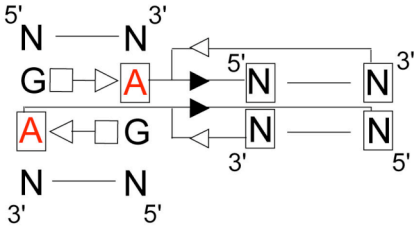


Figure 27 : Motif en A mineur type I/type I. Deux adénines dans des appariement *trans* Hoogsteen/Sucre forment des interactions de type I avec deux paires Watson-Crick.

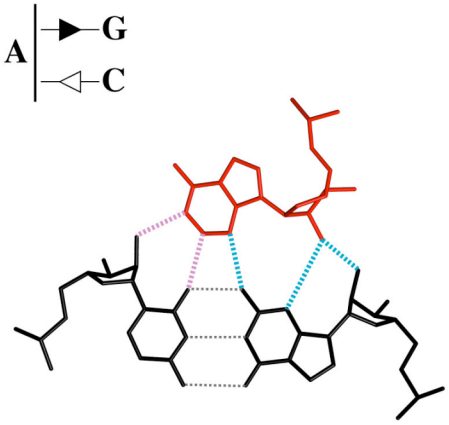
Si on considère les matrices d'isostérie correspondant aux interactions *cis* et *trans* Sucre/Sucre composant le motif en A mineur type I/type II, seules certaines zones de celles-ci peuvent être concernées par d'éventuelles variations de séquences dans un alignement de séquences homologues (Figure 28). Les contraintes supplémentaires imposées par le motif sont les suivantes : c'est toujours une adénine qui est impliquée dans une interaction de type I donc les variations observées se feront dans la première ligne de la matrice *cis* Sucre/Sucre et dans la première colonne de la matrice *trans* Sucre/Sucre. Dans le cas du type II, les variations se feront dans la première et la troisième colonne

de la matrice *cis* Sucre/Sucre selon qu'une adénine ou une guanine est impliquée. Ainsi, le motif A mineur impose un masque à la matrice qui « cache » les appariements base-base impossibles et rend seuls lisibles les appariements autorisés par les contraintes. Cet ensemble d'appariements autorisés constitue la séquence signature du motif A typeI/typeII. La matrice d'isostérie de l'interaction *trans* Sucre/Sucre ne présente que des interactions spécifiques entre bases (liaison b-b). C'est-à-dire qu'il y a toujours une liaison hydrogène entre un atome de chacune des bases. Ce n'est pas le cas de la matrice d'isostérie de l'interaction de type *cis* Sucre/Sucre. Le triangle, symbole de cette interaction, va de la base qui « donne » son hydroxyle en 2' à la base « réceptrice ». Contrairement aux purines, dans le cas d'une pyrimidine réceptrice, les interactions ne sont pas spécifiques. On peut ainsi distinguer au sein de la matrice des interactions spécifiques « S » et non spécifiques « RZ » comme nous l'avons vu plus haut. Ainsi il apparaît que le motif en A mineur dont les interactions impliquent les côtés Sucre des bases de deux paires consécutives, présente une spécificité qui devrait permettre de les identifier dans un alignement structural de séquences homologues.

Toutefois, la caractéristique majeure du motif en A mineur est l'indifférence des paires Watson-Crick contactées par les adénines extrudées. Au cours de l'évolution ou du repliement d'un ARN, tout élément de structure secondaire ayant la possibilité de présenter des adénines extrudées est potentiellement capable d'interagir avec deux plateaux de bases consécutifs d'une hélice quelconque. Afin d'assurer la spécificité du repliement tertiaire et de l'architecture d'un ARN, d'autres motifs structuraux doivent garantir que les motifs en A mineur interagissent entre les partenaires appropriés. Les récents travaux de Schultes et collaborateurs montrent que des séquences arbitraires d'ARN adoptent des repliements compacts et des structures secondaires spécifiques, impliquant que la sélection naturelle n'a pas besoin d'intervenir pour qu'une molécule d'ARN acquière ces caractéristiques structurales (Schultes et al., 2005). Les méthodes employées dans ces travaux ne permettent pas d'interpréter la nature des interactions tridimensionnelles de ces ARN arbitraires et de déterminer s'ils adoptent une structure tridimensionnelle stable ou dynamique. Les auteurs proposent que la stabilisation d'une structure tertiaire spécifique soit imposée par la sélection naturelle. La faible spécificité du motif A mineur laisse envisager son implication dans la structuration de molécules d'ARN

de séquence aléatoire. Au cours de l'évolution, parmi de nombreuses combinaisons, certaines structurations mèneraient à une fonctionnalité de molécules d'ARN, seraient sélectionnées et devraient être maintenues. Ceci implique que les interactions de certains partenaires structuraux au sein de la molécule soient conservées.

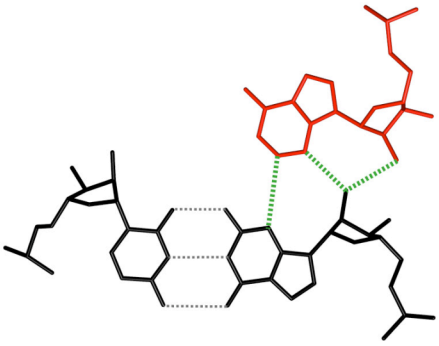
Type I



▶	A	C	G	U
A	S	RZ	S	RZ
C				
G				
U				

▷	A	C	G	U
A	S			
C	S			
G	S			
U	S			

Type II



▶	A	C	G	U
A	S		S	
C	S			
G	S			
U	S		S	

Figure 28 : Séquence signature du motif en A mineur typeI/typeII. En haut : à gauche, exemple d'interaction de type I entre une adénine et une paire C=G. Les liaisons hydrogène appartenant à l'appariement *cis* Sucre/Sucre sont colorées en bleu. Celles appartenant à l'appariement *trans* Sucre/Sucre sont colorées en rose. A droite, les matrices montrent les appariements *cis* et *trans* Sucre/Sucre autorisés dans le motif A mineur typeI/typeII. En bas: à gauche, exemple d'interaction de type II entre une adénine et une paire C=G. Les liaisons hydrogène de l'appariement *cis* Sucre/Sucre sont en vert. A droite, matrice montrant les appariements autorisés dans le motif en A mineur formé par deux adénines sur un simple brin (dans la colonne A) ou par une guanine à la place de l'adénine en 5' et une adénine dans une boucle GNRA (colonne G).

Si le type d'interaction, ici le motif en A mineur, n'assure pas la sélection des partenaires dont l'interaction est nécessaire à la fonction, alors des éléments structuraux doivent assurer leur proximité dans l'espace (Figure 29). C'est pourquoi des études des jonctions entre éléments de structure secondaire, qui maintiennent une architecure fonctionnelle d'un ARN, sont indispensables à la compréhension et à la prévision du repliement des ARN. Au cours de ce travail de thèse, nous avons étudié en vue de leur prévision dans un alignement de séquences homologues le motif tournant K et le motif C, deux motifs induisant respectivement un coude ou un tassement lorsqu'il sont insérés au sein d'une hélice. Nous avons également proposé des règles expliquant la structuration des jonctions à trois hélices, très importantes dans de nombreuses structures comme celle du ribozyme à tête de marteau. Ces motifs et jonctions sont impliqués comme nous le verrons plus loin dans le positionnement des éléments structuraux et contribuent ainsi à la spécificité de leur interaction.

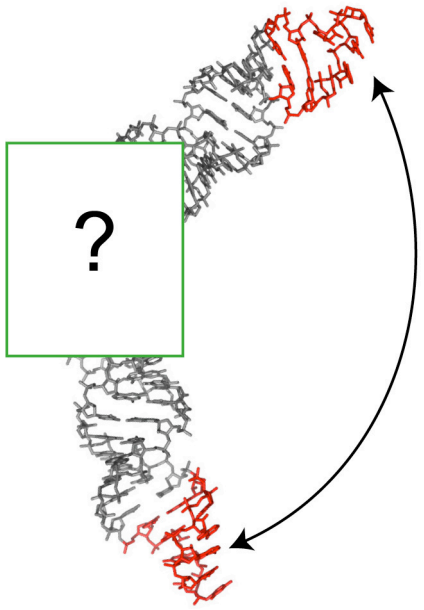


Figure 29 : Différents motifs peuvent assurer la proximité dans l'espace d'éléments capablent d'interagir spécifiquement (ici deux boucles terminales).

2.2. Utilisation des matrices d'isostérie

2.2.1. Introduction

Comme nous l'avons vu dans l'introduction générale, les molécules d'ARN repliées ont une architecture complexe qui engage un grand nombre de bases dans des appariements non Watson-Crick. Ces paires de bases non canoniques forment des motifs impliqués dans des interactions à longue distance, dans la fixation de petites molécules ou dans la liaison de protéines par exemple.

Au sein de molécules d'ARN homologues, l'architecture tridimensionnelle est plus conservée que la structure secondaire, elle même plus conservée que la structure primaire. En effet, différentes structures secondaires peuvent se replier en une même structure tridimensionnelle et différentes séquences peuvent mener à une même structure secondaire. Le phénomène de covariation observé pour la conservation de la structure secondaire implique une isostérie des paires de bases canoniques. Les paires Watson-Crick canoniques A-U, U-A, C=G, G=C sont isostériques et peuvent donc être remplacées les unes par les autres sans que la conformation du squelette sucre-phosphate ne soit perturbé. La prévalence et les rôles centraux des paires non Watson-Crick ont suscité dans de précédents travaux la mise au point de règles étendant le concept d'isostérie aux appariements non canoniques. Ainsi, comme l'étude des covariations de paires Watson-Crick dans un alignement de séquences d'ARN homologues permet de déterminer la structure secondaire de la molécule, les règles de covariation des paires non Watson-Crick permettent d'identifier ces appariements dans un alignement (Leontis & Westhof, 2001; Leontis et al., 2002b).

Notre objectif est de déterminer, sur la base d'alignements de séquences, les règles structurales et d'évolution basées sur l'isostérie auxquelles obéit le repliement de l'ARN. A plus long terme, ces règles devraient permettre de déduire le repliement tridimensionnel d'une molécule ARN à partir d'un alignement de séquences. Une première étape consiste à déterminer qu'elles sont les règles de covariation des motifs ARN connus c'est-à-dire de déterminer la séquence signature de chacun. Nous avons, dans le travail qui suit, déterminé la séquence signature de deux motifs : le motif tournant K et le motif C.

Ces deux motifs ont été définis selon notre concept de motif c'est-à-dire une suite ordonnée de paires de bases non Watson-Crcik qui peuvent être espacées par des bases extrudées. Nous avons considéré les motifs tournant K et les motifs C contenus dans les structures cristallographiques des ARNr 16S et 23S et nous avons analysé, en nous basant sur les matrices d'isostérie, les variations observées aux positions des paires de bases caractéristiques dans l'alignement de séquences correspondant. En moyennant l'ensemble des données pour les motifs C et tournants K nous obtenons une séquence signature caractéristique de chaque motif. Outre l'identification de la séquence signature de ces motifs, ce procédé nous a permis d'améliorer l'alignement structural en repérant les nucléotides mal appariés qui sortent du cadre caractéristique de la matrice correspondante et d'y remédier. Ces séquences signatures valident le travail précédemment fait par Leontis et Westhof (2002) sur les boucles E (voir Introduction générale) c'est-à-dire l'utilisation des matrices d'isostérie non seulement pour l'analyse, la classification et la prédiction de la structure tridimensionnelle de motifs ARN, mais également pour l'amélioration d'un alignement de séquences homologues non plus basé sur la seule structure secondaire mais également sur la structure tridimensionnelle d'une molécule. D'autre part, les règles déterminées pour les motifs C et tournant K sont des informations, qui détectées dans l'alignement des séquences homologues d'un ARN de structure tridimensionnelle inconnue, permettront de conclure quant à l'existence de l'un ou l'autre motif à des positions définies. La détermination des séquences signatures de motifs structuraux entraine l'amélioration d'un alignement structural de séquences homologues et dans le futur permettra d'automatiser le processus d'alignement.

2.2.2. Article 1 : " Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments"

Lescoute A, Leontis NB, Massire C, Westhof E.

Nucleic Acids Res. 2005 Apr 28;33(8):2395-409

Supplementary Material

S1 and S2 : Examples where isostericity matrices were helpful for the local realignment.

S1 : 23S KT-7

KT-7 sequence of *Blochmannia floridanus* was first aligned with 4 nucleotides in the loop (secondary structure on the left) as in *S.cerevisiae* (see figure KT-7 in S3). In this case, as shown in the corresponding matrices (bottom left), bp 3 is an unusual UG pair which belongs to the I2 isosteric family and bp 4 is a GU which is not allowed in a *trans* SE/SE interaction. After realignment (on the right), with five nucleotides in the loop and the resulting slippage, bp3 is an AG pair, bp4 a GA pair which are compatible with the corresponding isostericity matrices.

S2 : 23S C96

C96 sequences of *Sphatidium amphoriforme* and *Euplotes aediculatus* were aligned with three nucleotides in each strand of the internal loop which is closed by an unusual UU *cis* Watson-Crick/Watson-Crick base pair. At bp 4, a UU pair is compatible with I.M.4. However, at pb2, a AU pair is not compatible with the *trans* WC/Hg I.M.. After realignment, the UU pair is replaced by standard AU pair and the *trans* WC/HG pair becomes an usual AA base pair.

S3 : Kink-turn IM Analysis

An Isostericity Matrix analysis is presented for each of the K-turn motif in the ribosome. Each K-turn is discussed critically in order to identify those small set of sequences that either do not conserve the motif or contain a pronounced variant.

16S Kt-11:

BP1 (Orange – *cis* Watson-Crick): For last WC basepair of the C-stem (247/277 in *H.marismortui* numbering, as in the *T. th.* crystal structure, which we use throughout to discuss the 16S motifs), C/G and U/A Watson-Crick base pairs are observed in eukarya (*e*) while archaea (*a*) and bacteria (*b*) are almost exclusively C/G.

BP2 (Red – *trans* H/SE): In the 246-281 base pair, only adenine appears as the Hoogsteen base (246). The SE base is almost exclusively G in *b*, but significant numbers of isosteric substitutions of A and C occur for *a* and of A, C and U for *e*.

BP3 (Purple): This pair is actually *trans* Watson-Crick/Hoogsteen in the crystal structure rather than *trans* SE /Hoogsteen as in canonical K-turns. It is almost invariant as A243-A282 in all three phylogenetic domains.

BP4 (Blue – *trans* SE/SE): Only G appears at position 247 in *a* and *b* while A is more prevalent than G for *e*.

BP5 (Green – *trans* SE/SE): All four bases occur at position 278 in all phylogenetic domains, giving all four isosteric A/N combinations for the 246/278 basepair. Four gaps are observed at position 278 (*Furculomyces boomerangus*, *Adipicola arcuatilis*, *Myrina pacifica* and ~~Tacum~~ *Tamu fisheri*). It seems that 1 – 7 nts are missing in these sequences.

Spironucleus muris, Pelodera punctata and *Pelodera strongyloide* have been removed from the Eukaryotes alignment.

16S Kt-23:

BP1 (Orange – *cis* Watson-Crick): C/G and U/G Watson-Crick base pairs are observed in *a* while *e* and *b* are almost exclusively C/G.

BP2 (Red – *trans* H/SE): Only adenine appears as the Hoogsteen base (687). The SE base (703) is almost exclusively G in *b* and *e*, but significant numbers of isosteric substitutions of A and C occur for *a*.

BP3 (Purple- *trans* Watson-Crick/Hoogsteen): There are significant numbers of GA pairs at positions 686/704 in other sequences. As GA is not possible for *trans* Watson-Crick/Hoogsteen, we propose that for sequences with G at position 686, the interaction reverts to the *trans* H/SE that occurs in classical K-turns.

BP4 (Blue – *trans* SE/SE): In the three kingdoms, GA is almost invariant at positions 688/704.

BP5 (Green – *trans* SE/SE): Three bases, A, G and U, occur at position 700 in *a* and *b* while in *e* CA is present some times in more of the three other pairs. UA is more frequent (>80%) in *a* and *e*.

23S Kt-7:

BP1 (Orange – *cis* Watson-Crick): The Watson-Crick basepair (H.m. C93/G81, orange) shows typical WC covariations.

BP2 (Red – *trans* H/SE): The A/G at positions 80-97 is almost invariant. As seen in the table of values, A/A and A/U are observed only in *b* and belong to the same isosteric family I1.
BP3 (Purple - *trans* H/SE): As for Bp 2, most observed base pair for BP3 is A/G, but A/A and A/U base pairs are also observed.
BP4 (Blue – *trans* SE/SE): As seen for other K-turns, G is preferred at position 81 for 23S Kt-7, while the partner base is preferentially A in both *a* and *b*.

BP5 (Green – *trans* SE/SE): All four bases can be placed at position 94 for *b* but C is not present for *a*. In the two kingdoms the receptor base is, however, always an adenine. This tendency is observed in almost all the K-turn motifs.

Thus, for both *trans* Sugar-Edge/Sugar-Edge pairs (Base pair 4 in blue and base pair 5 in green), the isosteric family I1 is clearly preferred, i.e. adenine is the preferred receptor of 2'-OH of the second partner.

<u>23S Kt-15</u>:

This K-turn belongs to a variable helix present in *H. marismortui* but not present at *b* or *e*. Two interactions present in this motif differ from the classical K-turns described heretofor. Indeed there is one canonical base pair added between an adenine of the internal loop and the uridine of the purple sheared pairing (BP3), and the nature of the *trans* Sugar-edge/Sugar-edge in blue changes to *cis* Sugar-edge/Sugar-edge (U gives its 2'OH to G of the WC/WC pairing).

BP1 (Orange – *cis* Watson-Crick): The Watson-Crick basepair (H.m. C260/G249, orange) shows typical WC covariations.

BP2 (Red – *trans* H/SE): At positions 247/264 A/G is invariant.

BP3 (Purple - *trans* H/SE): The most observed base pair is U/G which belongs to I2. In classical K-turn only base pairs from I1 are observed. The presence of *cis* SE/SE in blue allow U to form *trans* WC/SE base pair what is bannished in the case of *trans* SE/SE base pair.

BP4 (Blue – *trans* SE/SE): As the type of pairing is changed in *cis* SE/SE, the values are not coherent with the isostericity table of *trans* SE/SE pairing. Indeed UC and UU are the most frequent base-pair while they do not exist in a *trans* SE-SE table. In *cis* SE/SE table the all sixteen base pairs belong to the same isosteric family I1. So values are coherent with this table.

BP5 (Green – *trans* SE/SE): The three base pairs , A/A, C/A and G/A belong to the same I1 family.

<u>23S Kt-38</u>:

This K-turn is present only in *a*.

BP1 (Orange – *cis* Watson-Crick): This basepair (1026-940) shows typical WC covariations.

BP2 (Red – *trans* H/SE): At positions 939-1031 A/G is almost invariant.

BP3 (Purple - *trans* H/SE): The base variations of Bp 3 are similar to those of Bp 2. The most observed base pair is A/G but A/U base pair is observed too. They belong to the same isosteric family I1.

BP4 (Blue – *trans* SE/SE): The only pair observed is G/A.

BP5 (Green – *trans* SE/SE): Unlike other K-turns (apart from KT-15), U does not appear at position 261. The three other bases from I1 family are observed at this position.
Thus, for the two *trans* Sugar-Edge/Sugar-Edge pairs (Base pair 4 in blue and base pair 5 in green), the isosteric family I1 is clearly preferred, i.e. adenine is the preferred receptor of 2'OH of the second partner.

23S Kt-42:

BP1 (Orange – *cis* Watson-Crick): This basepair (1147-1216) shows typical WC covariations. While C/G is most frequent in *a* and *b*, UA is the most frequent for *e*.

BP2 (Red – *trans* H/SE): In *a*,1215-1151 is always A/G, but A/A is observed for *b* and *e*, and in addition, in A/U and A/C for *b*. All these base pairs belong to I1 family.

BP3 (Purple - *trans* H/SE): As for BP2, A/A and A/U are observed for BP3, in addition to the more prevalent A/G. What is surprising here is that the bases do not directly H-bond in the x-ray structure (a water molecule mediates the interaction) but the table of variations is coherent with the type of interaction. Could we suppose that contrary to the crystal structure, this pairing is present in a biological context ?

BP4 (Blue – *trans* SE/SE): The G/A base pair is almost the only pairing present for *a* and *b* while AA is the only base pair observed for *e*. However they belong to the same family I1.

BP5 (Green – *trans* SE/SE): All four bases occur at position 1148. In the three superkingdoms the receptor base is always adenine, so all base pairs belong to I1 family. This tendency is observed in almost all the k-turn motifs.

23S Kt-46:

BP1 (Orange – *cis* Watson-Crick): The Watson-Crick basepair shows typical WC covariations but U/G is also present, but infrequent. In one sequence (*Rhodopirellula baltica*) of the *b* (0.1% of 805), the G/G base pair is observed, indicating a different motif.

BP2 (Red – *trans* H/SE): At positions 1341-1316 A/G is almost invariant for *a* and *b* but A/A is significant for *e*. For *b*, the C/C pair belongs to *R. baltica*. However all these base pairs belong to I1 family.

BP3 (Purple - *trans* H/SE): The most observed base pair is A/G but A/A and C/A base pairs are very prevalent for *e*. These three base pairs belong to I1 family. For *b* : G/U was present in *R. baltica* but by realignment an A can replace the G what allows not only the conservation of the nature pairing but still maintains I1 isosteric family. For *e* : One sequence (*Euglena gracilis*) presents the C/G base pair, indicating that all or part of the NC stem seems to be replaced by *cis* WC base-pairs.

BP4 (Blue – *trans* SE/SE): As seen for other K-turns, an A is preferred at position 1317 but its partners are more variable than in other k-turns. For *b* : As for BP3 realignment allows to conserve the pairing by replacing G by A at position 1317. For *e* : In the crystal structure of *H.m* there is only one H-bond between O2' of C1342 and N3 of A1317. So the constraints are more flexible and the number of possible base pairs is larger. Thus, it is not surprising to observe A/C, G/C or C/U base pairs. Indeed an H-bond is possible between the 2'OH of the first base and the O2 of C or U when these are present as the second base.

BP5 (Green – *trans* SE/SE): All four bases can be placed at position 1313 for *a* and *e* but C is not present for *b*. The receptor base is always an adenine apart in one sequence of *b*, *R.baltica*, where AC base pair is observed. This base pair contrary to the others (I1) is not possible in this type of interaction.

23S Kt-58:

This motif is present in *a* but not in *b* or *e*. In *E.coli* the motif is replaced by a helix with two adenines in bulge. In the case of *H. marismortui* almost all the interactions characteristic of K-turn-motif are present. However BP4 (Blue – *trans* SE/SE) is replaced by a water-mediated interaction in the crystal structure.

BP1 (Orange – *cis* Watson-Crick): The Watson-Crick basepair (H.m. G1601/C1593, orange) shows typical WC covariations.

BP2 (Red – *trans* H/SE): At positions 1590/1605 A/G is invariant.

BP3 (Purple - *trans* H/SE): The case of the Bp 3 is the same of Bp 2.

BP4 (Blue – *trans* SE/SE): CA and UA belong to the I1 isosteric family. Surprisingly, the base pairs generally observed in other K-turns, G/A and A/A to a lesser extent, are absent while C/A and U/A which are poorly present in others are the only base pairs observed.

BP5 (Green – *trans* SE/SE): All four bases can be placed at position 1602, giving all four isosteric A/N combinations for the 1602/1590 basepair.

<u>23S Composite Kt-94/99</u>:

This K-turn belongs to the domain VI of the 23S rRNA of *H. marismortui*. It is different from the previous described k-turns because it is composed by three strands as shown on figure 3. All five characteristic tertiary interactions are present. In the case of *E.coli* the formation of this K-turn seems to be more difficult. Indeed there is an added helix at position 2791 in the middle of the single strand from 2789 to 2810 corresponding in *H.marismortui* to the single strand beginning with the nucleotide 2820. Therefore, we have analyzed archaeal sequences. In the archaea, 8 sequences out of 24 also have an added helix on the continuous strand of this K-turn.

BP1 (Orange – *cis* Watson-Crick): This basepair shows typical WC covariations.

BP2 (Red – *trans* H/SE): The A/G base pair which is the most frequent base pair in other k-turns is comparatively poorly represented in this case. Moreover there is no A/A base pair, G/G, U/A and U/G from I2 family are present contrary to the typical tables, where only I1 is represented, and G/A and U/U represent almost 17% -- bases that cannot form *trans* H/SE basepairs.

BP3 (Purple - *trans* H/SE): Almost 17% of sequences present C/G, G/C and G/U at positions 2827-2913 while it is not possible for a *trans* H/SE. Rather, these variations are typical from *cis* WC/WC base pair; thus, in these sequences the sheared base pair appears to be replaced by a canonical base pair.

BP4 (Blue – *trans* SE/SE): The most frequent base pair is G/A but contrary to the other K-turn G/G which belong to I2 and G/C which are not allowed are present too.

BP5 (Green – *trans* SE/SE): U at position 2914 represent almost 42% of the sequences while this base is never a receptor of hydroxyle in *trans* SE-SE base pair. Indeed only A and G can be receptors of the 2' hydroxyle in this type of pairing.

<u>23S Composite Kt-4/5:</u>

This K-turn occurs in domain I of the 23S rRNA of *H. marismortui* , and appears to be present in all three superkingdoms. Like composite K-turn KT94-99, it is composed of three strands (Figure 3). BP3 (purple) which is a sheared base pair in classical K-turns is replaced by a *trans* H/H base pair.

BP1 (Orange – *cis* Watson-Crick): This basepair shows typical WC covariations. In *a* the proportion of G/U base pair is relatively large.

BP2 (Red – *trans* H/SE): As for other K-turns, A/G base pair is the most frequent base pair.

BP3 (Purple - *trans* H/SE): The values are almost invariant so it is difficult to see any difference in the tables between *trans* H/H and *trans* H/SE.

BP4 (Blue – *trans* SE/SE): G/A base pair is the most frequent base pair. A/A base pair which belong to the same family I1 is present too.

BP5 (Green – *trans* SE/SE): G/A seems to be prefered in this case while the four N/A base pairs are equally present in almost all other k-turns. In *e* G/C base pair is observed while it is not allowed.

Trepomonas agilis has been removed from the Eukaryotes alignment.

23S Composite Kt-77/78:

This composite K-turn belongs to the domain V of the 23S rRNA of *H. marismortui*. It is different from the previous composite described k-turns. Indeed, the composed strand is opposite to the composed strand of KT-4/5 and KT-94/99. Only four tertiary interactions are present as BP 2 is absent.

BP1 (Orange – *cis* Watson-Crick): The Watson-Crick basepair shows typical WC covariations. Some U/G base pairs are present.
BP2 (Red – *trans* H/SE): This basepair is not present in the structure, as the conserved A is unpaired.

BP3 (Purple - *trans* H/SE): The most observed base pair is A/G for all three superkingdoms. All the base pairs belong to the same I1 family. There are gaps at position 2218.

BP4 (Blue – *trans* SE/SE): A/G is preferred at position 2168. The partner base of this pairing is preferentially an A. There are gaps at position 2218.

BP5 (Green – *trans* SE/SE): All the four bases can be placed at position 2207 for *e* but C is not present for *a* and *b*, nor U for *a*. In the three superkingdoms the receptor base is always an adenine. All the base pairs belong to I1 family apart G/G base pair for *e* which belong to I2 family.

S4 : C-Loop IM Analysis

An Isostericity Matrix analysis is presented for each of the C-Loop motif in the ribosome. Each C-Loop is discussed critically in order to identify those small set of sequences that either do not conserve the motif or contain a pronounced variant.

23S C-38:

BP 1 (Red – *cis* Watson-Crick) : the principal pairing adopted is a canonical CG base pair with the C in this basepair forming a *cis* Watson-Crick/Sugar-Edge interaction with a C in the opposite strand.

BP 2 (Purple - *trans* Watson-Crick/Hoogsteen) : C/A pairing (I2) is the most frequent for the three superkingdoms. For *a* some base pairs belong to I1 family while for *b* and *e,* some belong to I4 family.

BP 3 (Blue - *cis* Watson-Crick/Sugar-Edge)The most frequent basepair for *a* and *b* is C/C, which belongs to the I2 family. However, there are four other possible isosteric pairing combinations : C/A, C/C, C/G and C/U. Why is this base pair in C-38 constrained? Some nucleotides in the same strand as C of G/C interact with protein L10E. The cytosine O1P H-bonds with Arg16 NH2 of protein L10E, but this interaction is not specific. For *e* the principal pairing is A/C (I1).

BP 4 (Green - *cis* Watson-Crick) : The canonical A/U pairing is very conserved in *b* and *e* but to a lesser extent for *a*. For *a* A/G and especially A/C also occur.

23S C-50

BP 1 (Red – *cis* Watson-Crick) : The *cis* WC/WC in red is absolutely conserved in *a*. Why this pairing is constrained is not clear. For *b* and *e,* the base pairs are more distributed among the I1 isosteric family and in others for *e*.

BP 2 (Purple - *trans* Watson-Crick/Hoogsteen) : The covariation matrix shows that a significant proportion of ~~C/G~~ C/G occur, however this base pair is not expected for this type of interaction. It appears therefore that an alternative structure is forming in some organisms. The simplest possibility is the motif occur as in *H.marismortui* but without the *trans* Watson-Crick/Hoogsteen pairing. In addition, in *a* and *e* the protein L39E binds the motif C-50. Protein L39E is present in eukaryotes and archaea, but not in bacteria where the *trans* Watson-Crick/Hoogsteen base pair is absent.

BP 3 (Blue - *cis* Watson-Crick/Sugar-Edge) : Majority of base pairs for *a*, *b* or *c* belong to I1 family.

BP 4 (Green - *cis* Watson-Crick) : The Watson-Crick basepair shows typical WC covariations. For *a* some sequences present A/C base pair.

23S C-96.

BP 1 (Red – *cis* Watson-Crick) : The G/C base pair is absolutely conserved in *a*. For *b,* the significant proportion of A/U also is observed. For *e,* the base pairs are distributed between G/C, A/U and A/C.

BP 2 (Purple - *trans* Watson-Crick/Hoogsteen) : The covariation matrix shows that C/A is preferred for *a* and *b* while A/A is preferred for *e*. In 1.8% of sequences A/U was observed – which is not possible in this type of interaction but after realignment another solution was found with A in the place of U at position 2761.

BP 3 (Blue - *cis* Watson-Crick/Sugar-Edge) : The majority of base pairs for *a* and *b* belong to I2 family while for *e* it is more distributed.

BP 4 (Green - *cis* Watson-Crick) : The A/U base pair is the most frequent in the three superkingdoms.

23S C-15.

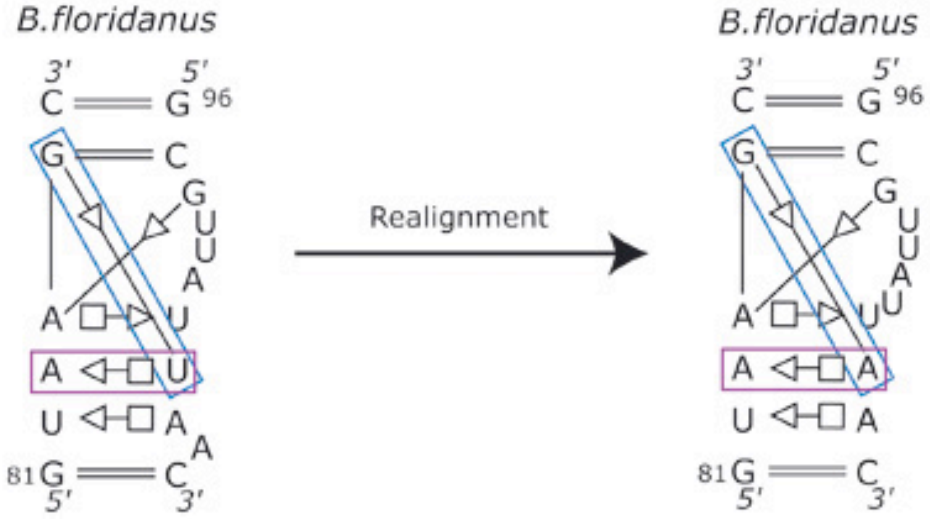
BP 1 (Red – *cis* Watson-Crick) : The Watson-Crick basepair shows typical WC covariations with some nearly isosteric wobble A/C and G/U present for *e* ; these pairs belong to the same isosteric family, i2.

BP 2 (Purple - *trans* Watson-Crick/Hoogsteen) : The covariation matrix shows that C/A base pair is preferred for *a*, *b* and *e*. For *e,* a few other base pairs are represented in low frequency.

BP 3 (Blue - *cis* Watson-Crick/Sugar-Edge) The majority of base pairs for *a* and *b* belong to I1 family while for *e* the C/G base pair is also present.

BP 4 (Green - *cis* Watson-Crick) : The A/U base pair is the most frequent in the three kingdoms.

23S KT-7



3 SE 83 — Hg 103

trans	A	C	G	U	-
A	15.4	-	71.3	12.9	-
C	-	-	-	-	-
G	-	-	-	-	-
U	-	-	0.1	-	-
-	-	0.1	-	-	-

3 SE 83 — Hg 104

trans	A	C	G	U	-
A	15.4	-	71.4	12.9	-
C	-	-	-	-	-
G	-	-	-	-	-
U	-	-	-	-	-
-	-	0.1	-	-	-

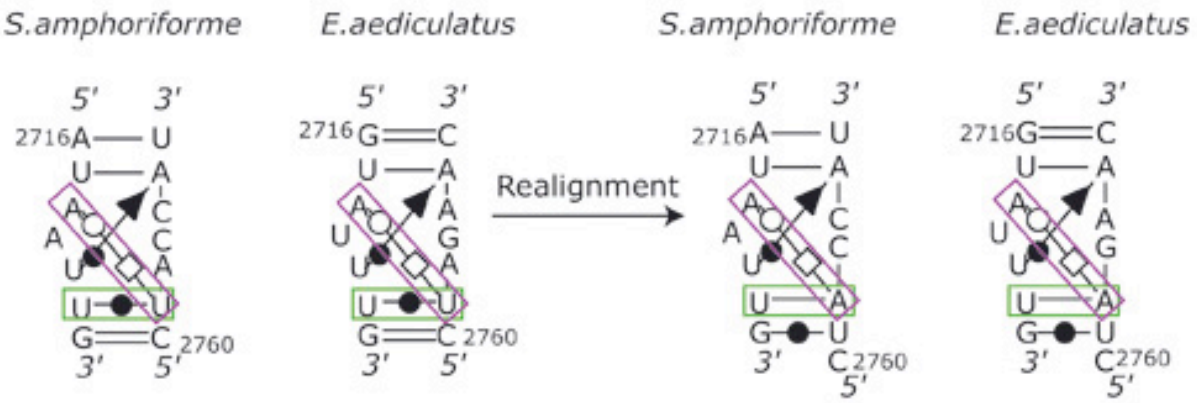
4 SE 103 — SE 85

trans	A	C	G	U	-
A	0.7	-	-	-	-
C	-	-	-	-	-
G	98.1	-	-	0.1	0.1
U	0.9	-	-	-	-
-	-	-	-	-	-

4 SE 104 — SE 85

trans	A	C	G	U	-
A	0.7	-	-	-	-
C	-	-	-	-	-
G	98.3	-	-	-	0.1
U	0.9	-	-	-	-
-	-	-	-	-	-

23S C96

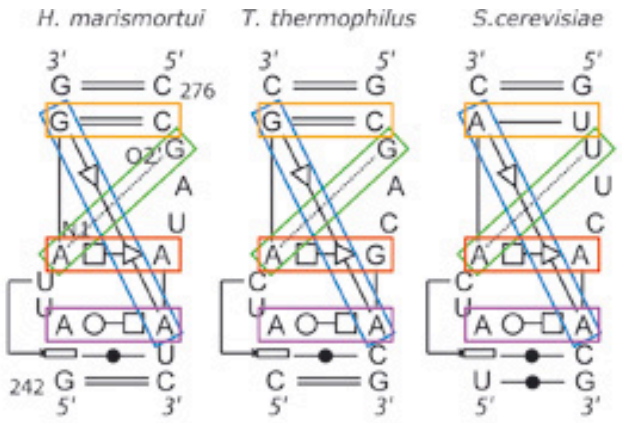


2	H 2761					
WC 2718	trans	A	C	G	U	-
	A	77.0	-	-	1.8	-
	C	17.7	-	1.8	-	-
	G	0.9	-	0.9	-	-
	U	-	-	-	-	-
	-	-	-	-	-	-

2	H 2761					
WC 2718	trans	A	C	G	U	-
	A	78.8	-	-	-	-
	C	17.7	-	1.8	-	-
	G	0.9	-	0.9	-	-
	U	-	-	-	-	-
	-	-	-	-	-	-

4	WC 2721					
WC 2761	cis	A	C	G	U	-
	A	0.9	-	-	94.7	-
	C	-	-	-	-	-
	G	-	2.7	-	-	-
	U	-	-	-	1.8	-
	-	-	-	-	-	-

4	WC 2721					
WC 2761	cis	A	C	G	U	-
	A	0.9	-	-	96.5	-
	C	-	-	-	-	-
	G	-	2.7	-	-	-
	U	-	-	-	-	-
	-	-	-	-	-	-



16S KT-11

1 WC 247

WC 277 \ cis	A	C	G	U	-
A	- - 0.1	-	- - <	- - 0.2	- - <
C	- 0.1	- - 0.1	100.0 99.9 16.9	- - 0.1	- - <
G	- - <	- - 0.3	-	-	-
U	- - 81.8	- - <	- 0.1 0.2	- - <	-
-	- - <	-	-	-	-

3 H 282

WC 243 \ trans	A	C	G	U	-
A	100.0 100.0 99.3	- - 0.1	- - <	- - <	- - <
C	-	-	-	-	-
G	- - 0.1	-	-	-	-
U	- - 0.2	-	-	-	-
-	-	-	-	-	-

2 SE 281

H 246 \ trans	A	C	G	U	-
A	37.5 - 56.3	8.3 - 6.3	54.2 100.0 32.0	- - 4.9	- - <
C	- - <	- - <	-	-	-
G	-	- - <	- - <	-	-
U	-	-	-	-	-
-	- - <	-	- - <	-	-

3 SE 243

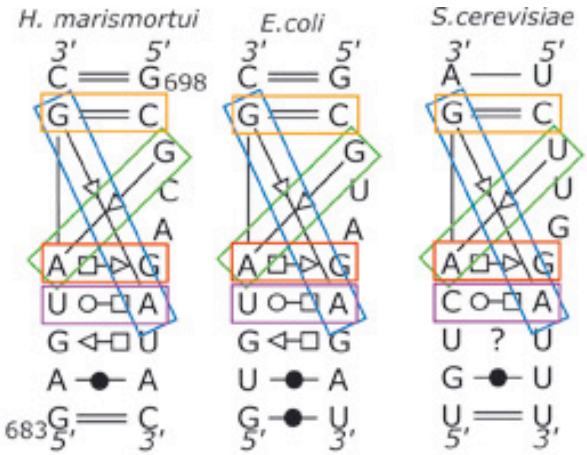
H 282 \ trans	A	C	G	U	-
A	100.0 100.0 99.3	-	- - 0.1	- - 0.2	-
C	- - 0.1	-	-	-	-
G	- - <	-	-	-	-
U	- - <	-	-	-	-
-	- - <	-	-	-	-

4 SE 282

SE 247 \ trans	A	C	G	U	-
A	- - 81.7	- - 0.1	- - <	- - <	- - <
C	- - 0.4	-	-	-	-
G	100.0 100.0 17.0	-	- - <	-	-
U	- - 0.3	-	-	-	-
-	- - <	-	-	-	-

5 SE 246

SE 278 \ trans	A	C	G	U	-
A	12.5 21.3 9.7	- - <	-	-	-
C	4.2 0.6 4.5	-	- - <	-	-
G	58.3 67.9 40.3	- - <	- - <	-	- - <
U	25.0 10.3 45.1	-	-	-	- - <
-	- - 0.1	-	-	-	-



16S KT-23

1 — WC 688 (WC 699)

cis	A	C	G	U	-
A	-	-	- / <	- / <	-
C	- / - / 0.1	-	75.0 / 98.3 / 97.8	- / 0.1	- / - / <
G	-	- / 0.1 / 0.2	- / - / 0.2	-	-
U	- / 0.9 / 0.3	-	25.0 / 0.8 / 1.0	- / - / 0.2	-
-	-	-	-	-	-

2 — SE 703 (H 687)

trans	A	C	G	U	-
A	83.3 / 0.4 / 0.5	4.2 / 0.1 / <	12.5 / 99.3 / 99.0	- / - / 0.2	- / - / <
C	-	-	-	-	-
G	-	-	- / - / <	-	-
U	-	- / - / <	-	-	-
-	-	-	- / - / <	-	-

3 — SE 686 (H 704)

trans	A	C	G	U	-
A	- / 0.3 / 0.5	- / - / 60.0	8.3 / 1.5 / 9.8	91.7 / 98.3 / 29.0	- / - / <
C	-	- / - / <	-	- / - / <	-
G	-	-	-	- / - / <	-
U	- / 0.1	-	- / - / <	- / - / 0.1	-
-	-	- / - / <	- / - / 0.1	- / - / <	-

3 — H 704 (WC 686)

trans	A	C	G	U	-
A	- / 0.3 / 0.5	-	-	- / 0.1	-
C	- / - / 60.0	- / - / <	-	-	- / - / <
G	8.3 / 1.5 / 9.8	-	-	- / <	- / - / 0.1
U	91.7 / 98.3 / 29.0	- / - / <	- / - / <	- / - / 0.1	- / - / <
-	- / - / <	-	-	- / -	-

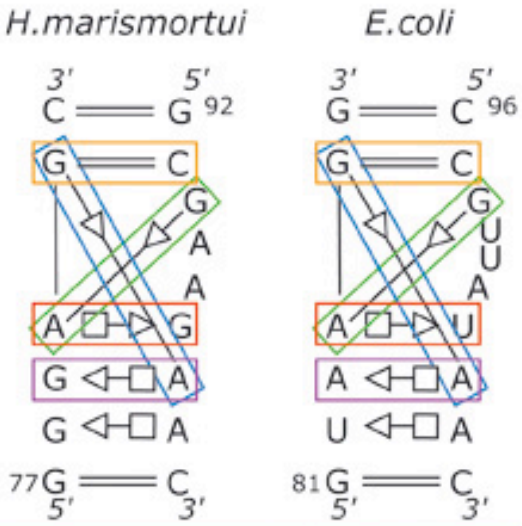
4 — SE 704 (SE 688)

trans	A	C	G	U	-
A	- / 0.9 / 0.4	-	-	-	- / - / <
C	- / 0.1 / 0.2	-	-	-	-
G	100.0 / 99.0 / 98.7	- / 0.1	- / - / <	- / 0.1	- / - / 0.1
U	- / 0.1	-	-	- / 0.2	-
-	- / - / <	-	-	-	-

5 — SE 687 (SE 700)

trans	A	C	G	U	-
A	4.2 / 0.9 / 1.8	-	-	-	-
C	- / - / 0.5	-	-	- / - / <	-
G	12.5 / 97.1 / 3.6	-	-	-	-
U	83.3 / 2.0 / 93.9	-	- / - / <	-	- / - / <
-	- / - / 0.1	-	-	-	-

23S KT-7



1 WC 81 — WC 93

cis	A	C	G	U	-
A	-	-	-	- 0.9	-
C	-	-	100.0 97.0	-	-
G	-	-	-	-	-
U	- 0.7	-	- 1.4	-	-
-	-	-	-	-	-

2 SE 97 — H 80

trans	A	C	G	U	-
A	- 0.5	-	100.0 81.6	- 17.8	-
C	-	-	-	-	-
G	-	-	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

3 SE 79 — H 98

trans	A	C	G	U	-
A	- 15.4	-	62.5 71.4	37.5 12.9	-
C	-	-	-	-	-
G	-	-	-	-	-
U	-	-	-	-	-
-	-	- 0.1	-	-	-

4 SE 98 — SE 81

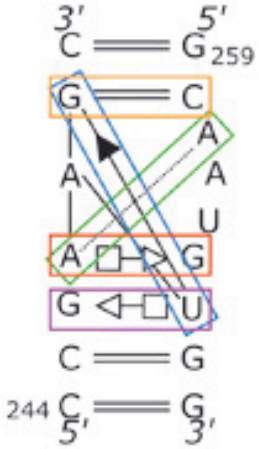
trans	A	C	G	U	-
A	- 0.7	-	-	-	-
C	-	-	-	-	-
G	100.0 98.3	-	-	-	- 0.1
U	- 0.9	-	-	-	-
-	-	-	-	-	-

5 SE 80 — SE 94

trans	A	C	G	U	-
A	29.2 12.9	-	-	-	-
C	- 2.6	-	-	-	-
G	41.7 36.8	-	-	-	-
U	29.2 47.6	-	-	-	-
-	-	-	-	-	-

H.marismortui

23S KT-15



1 WC 249

WC 260 \ cis	A	C	G	U	-
A	-	-	-	54.2	-
C	-	-	12.5	-	-
G	-	33.3	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

2 SE 264

H 247 \ trans	A	C	G	U	-
A	-	-	100.0	-	-
C	-	-	-	-	-
G	-	-	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

3 SE 246

H 265 \ trans	A	C	G	U	-
A	4.2	-	-	-	-
C	-	-	-	-	-
G	-	-	-	-	-
U	12.5	-	83.3	-	-
-	-	-	-	-	-

4 SE 265

SE 249 \ trans	A	C	G	U	-
A	-	-	-	-	-
C	4.2	-	-	29.2	-
G	-	-	-	12.5	-
U	-	-	-	54.2	-
-	-	-	-	-	-

4 SE 249

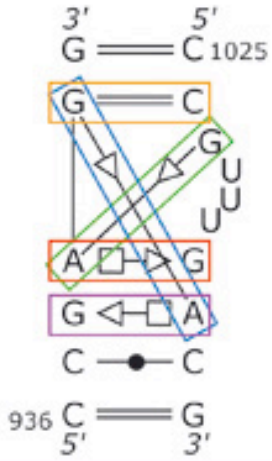
SE 265 \ cis	A	C	G	U	-
A	-	4.2	-	-	-
C	-	-	-	-	-
G	-	-	-	-	-
U	-	29.2	12.5	54.2	-
-	-	-	-	-	-

5 SE 247

SE 261 \ trans	A	C	G	U	-
A	29.2	-	-	-	-
C	12.5	-	-	-	-
G	58.3	-	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

23S KT-38

H.marismortui



3 — H 1032 × SE 938

H 1032 \ SE 938	trans	A	C	G	U	-
	A	-	-	95.8	4.2	-
	C	-	-	-	-	-
	G	-	-	-	-	-
	U	-	-	-	-	-
	-	-	-	-	-	-

1 — WC 1026 × WC 940

WC 1026 \ WC 940	cis	A	C	G	U	-
	A	-	-	-	-	-
	C	-	-	87.5	-	-
	G	-	-	-	-	-
	U	-	-	12.5	-	-
	-	-	-	-	-	-

4 — SE 940 × SE 1032

SE 940 \ SE 1032	trans	A	C	G	U	-
	A	-	-	-	-	-
	C	-	-	-	-	-
	G	100.0	-	-	-	-
	U	-	-	-	-	-
	-	-	-	-	-	-

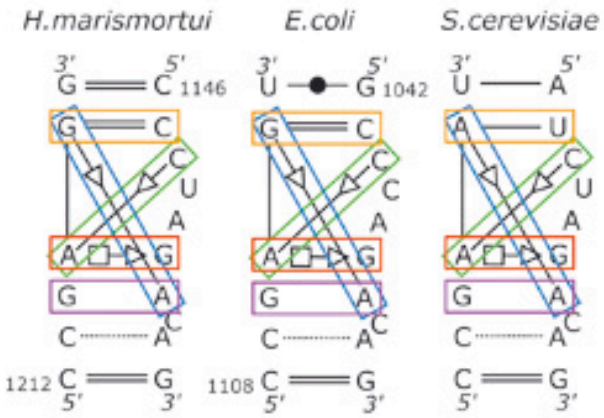
2 — H 939 × SE 1031

H 939 \ SE 1031	trans	A	C	G	U	-
	A	-	-	95.8	4.2	-
	C	-	-	-	-	-
	G	-	-	-	-	-
	U	-	-	-	-	-
	-	-	-	-	-	-

5 — SE 1027 × SE 939

SE 1027 \ SE 939	trans	A	C	G	U	-
	A	12.5	-	-	-	-
	C	12.5	-	-	-	-
	G	75.0	-	-	-	-
	U	-	-	-	-	-
	-	-	-	-	-	-

23S KT-42



1 — WC 1216 (WC 1147)

cis	A	C	G	U	-
A	- - 0.8	-	-	-	-
C	-	-	100.0 99.1 -	-	-
G	-	-	-	-	-
U	- 0.9 99.2	-	-	-	-
-	-	-	-	-	-

2 — SE 1151 (H 1215)

trans	A	C	G	U	-
A	- 11.6 3.0	- 0.1 -	100.0 87.5 97.0	- 0.7 -	-
C	-	-	-	-	-
G	-	-	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

3 — SE 1214 (H 1152)

trans	A	C	G	U	-
A	- 8.4 1.5	-	100.0 80.8 98.5	- 10.8 -	-
C	-	-	-	-	-
G	-	-	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

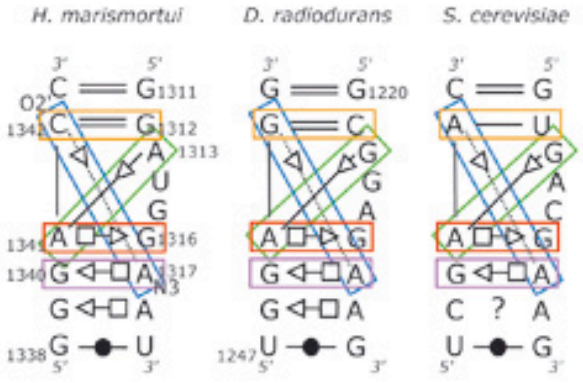
4 — SE 1152 (SE 1216)

trans	A	C	G	U	-
A	- 0.9 100.0	-	-	-	-
C	-	-	-	-	-
G	100.0 99.1 -	-	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

5 — SE 1215 (SE 1148)

trans	A	C	G	U	-
A	8.0 30.8 28.6	-	-	-	-
C	48.0 27.6 28.6	-	-	-	-
G	8.0 5.6 3.8	-	-	-	-
U	36.0 36.0 39.1	-	-	-	-
-	-	-	-	-	-

23S KT-46



1 — WC 1342 — WC 1312

cis	A	C	G	U	-
A	-	-	-	- 0.4 0.8	-
C	-	- 0.1 -	29.2 67.8 42.1	-	-
G	-	70.8 20.7 6.8	- 0.1 -	- 0.2 -	-
U	- 7.5 45.1	- - 1.5	- 3.1 3.8	-	-
-	-	-	-	-	-

2 — SE 1316 — Hg 1341

trans	A	C	G	U	-
A	- 0.1 15.8	- - 1.5	100.0 99.8 81.2	- - 0.8	-
C	-	- 0.1 -	-	-	-
G	-	-	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

3 — SE 1340 — Hg 1317

trans	A	C	G	U	-
A	- - 5.3	-	100.0 99.9 81.2	- 0.1 -	-
C	- - 12.0	-	- - 0.8	-	-
G	-	-	-	-	-
U	-	-	- - 0.8	-	-
-	-	-	-	-	-

4 — SE 1317 — SE 1342

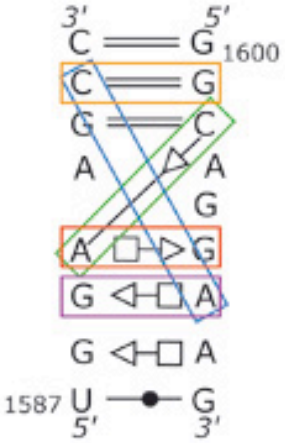
trans	A	C	G	U	-
A	- 7.5 33.8	- - 11.3	-	-	-
C	70.8 20.9 7.5	-	-	- - 0.8	-
G	29.2 71.1 44.4	- - 1.5	-	-	-
U	- 0.6 0.8	-	-	-	-
-	-	-	-	-	-

5 — SE 1341 — SE 1313

trans	A	C	G	U	-
A	12.5 11.1 16.5	- 0.1 -	-	-	-
C	4.2 - 7.5	-	-	-	-
G	79.2 64.3 54.9	-	-	-	-
U	4.2 24.5 21.1	-	-	-	-
-	-	-	-	-	-

H.marismortui

23S KT-58



3 — SE 1589 — H 1606

trans	A	C	G	U	-
A	-	-	100.0	-	-
C	-	-	-	-	-
G	-	-	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

1 — WC 1593 — WC 1601

cis	A	C	G	U	-
A	-	-	-	8.3	-
C	-	-	-	-	-
G	-	91.7	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

4 — SE 1606 — SE 1593

trans	A	C	G	U	-
A	-	-	-	-	-
C	91.7	-	-	-	-
G	-	-	-	-	-
U	8.3	-	-	-	-
-	-	-	-	-	-

2 — SE 1605 — H 1590

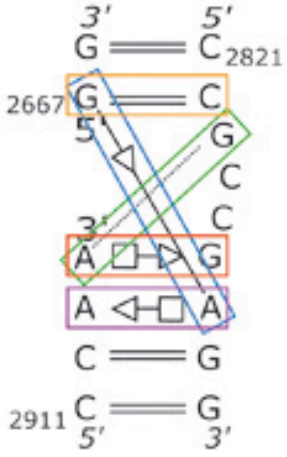
trans	A	C	G	U	-
A	-	-	100.0	-	-
C	-	-	-	-	-
G	-	-	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

5 — SE 1590 — SE 1602

trans	A	C	G	U	-
A	4.2	-	-	-	-
C	83.3	-	-	-	-
G	4.2	-	-	-	-
U	8.3	-	-	-	-
-	-	-	-	-	-

H.marismortui

23S KT-94/99



3 — SE 2913

H 2827 \ trans	A	C	G	U	-
A	25.0	4.2	54.2	-	-
C	-	-	4.2	-	-
G	-	4.2	-	8.3	-
U	-	-	-	-	-
-	-	-	-	-	-

1 — WC 2667

WC 2822 \ cis	A	C	G	U	-
A	-	-	-	-	-
C	-	-	87.5	-	-
G	-	-	-	-	-
U	12.5	-	-	-	-
-	-	-	-	-	-

4 — SE 2827

SE 2667 \ trans	A	C	G	U	-
A	12.5	-	-	-	-
C	-	-	-	-	-
G	70.8	4.2	12.5	-	-
U	-	-	-	-	-
-	-	-	-	-	-

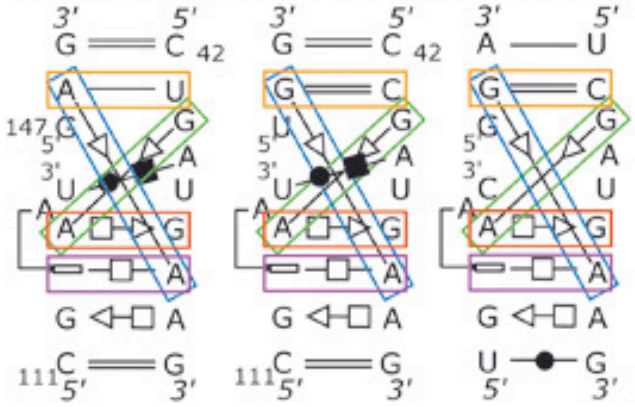
2 — SE 2826

H 2914 \ trans	A	C	G	U	-
A	-	4.2	37.5	4.2	-
C	-	-	-	-	-
G	4.2	-	8.3	-	-
U	25.0	-	4.2	12.5	-
-	-	-	-	-	-

5 — SE 2914

SE 2823 \ trans	A	C	G	U	-
A	4.2	-	-	-	-
C	25.0	-	12.5	4.2	-
G	16.7	-	-	4.2	-
U	-	-	-	33.3	-
-	-	-	-	-	-

H.marismortui D.radiodurans S.cerevisiae



23S KT-4/5

1 WC 148

WC 43 \ cis	A	C	G	U	-
A	-	-	-	-	-
C	- 0.2 -	-	54.2 98.6 98.6	-	-
G	-	-	-	-	-
U	8.3 1.0 1.4	-	37.5 - -	-	-
-	- 0.1 -	-	-	-	-

3 SE 114

H 48 \ trans	A	C	G	U	-
A	100.0 99.9 100.0	- 0.1 -	-	-	-
C	-	-	-	-	-
G	-	-	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

2 SE 47

WC 113 \ trans	A	C	G	U	-
A	- 0.1 -	-	100.0 99.9 98.6	-	-
C	-	-	- - 1.4	-	-
G	-	-	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

3 H 48

H 114 \ trans	A	C	G	U	-
A	100.0 99.9 100.0	-	-	-	-
C	- 0.1 -	-	-	-	-
G	-	-	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

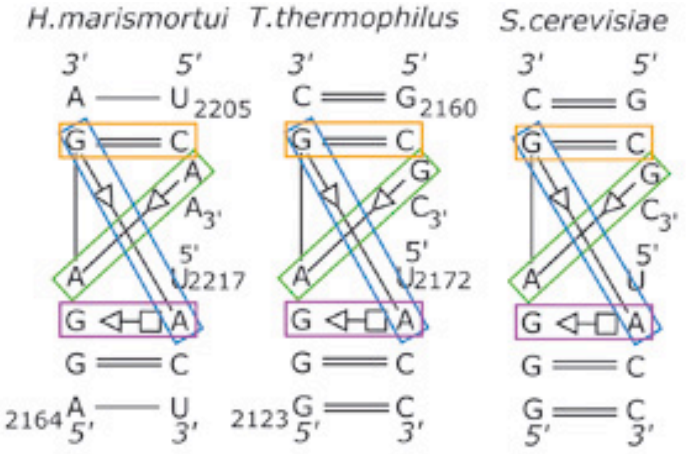
4 SE 48

SE 148 \ trans	A	C	G	U	-
A	8.3 1.4 1.4	-	-	-	-
C	-	-	-	-	-
G	91.7 98.6 98.6	-	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

5 SE 113

SE 44 \ trans	A	C	G	U	-
A	- 0.4 -	-	-	-	-
C	- 0.1 -	-	-	-	-
G	87.5 99.5 98.6	- - 1.4	-	-	-
U	12.5 - -	-	-	-	-
-	-	-	-	-	-

23S KT-77/78



1 — WC 2168 × WC 2206

cis	A	C	G	U	-
A	-	-	-	-	-
C	-	-	100.0 99.1 94.7	-	-
G	-	- - 0.8	-	-	-
U	- 0.7 0.8	-	- 0.1 3.8	-	-
-	-	-	-	-	-

4 — SE 2218 × SE 2168

trans	A	C	G	U	-
A	- 0.7 0.8	-	-	-	-
C	- - 0.8	-	-	-	-
G	100.0 99.3 97.0	-	-	-	- - 1.5
U	-	-	-	-	-
-	- -	-	-	-	-

3 — SE 2166 × H 2218

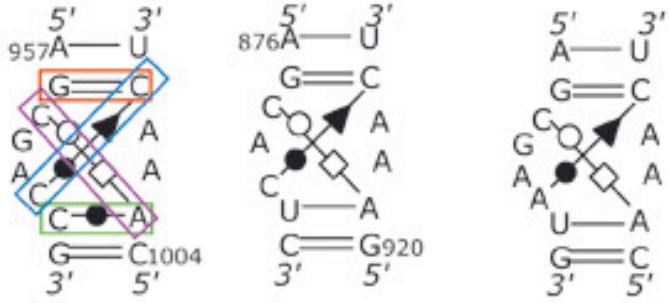
trans	A	C	G	U	-
A	- 0.1 2.3	- - 0.8	100.0 96.2 95.5	- 3.7 -	-
C	-	-	-	-	-
G	-	-	-	-	-
U	-	-	-	-	-
-	-	-	- - 1.5	-	-

5 — SE 2167 × SE 2207

trans	A	C	G	U	-
A	25.0 37.5 8.3	-	-	-	-
C	- - 2.3	-	-	-	-
G	75.0 62.0 86.5	-	- - 0.8	-	-
U	- 0.5 2.3	-	-	-	-
-	-	-	-	-	-

23S C38

H.marismortui *D.radiodurans* *S.cerevisiae*



1 WC 958 — WC 1008

cis	A	C	G	U	-
A	-	-	-	-	-
C	- - 0.8	-	100.0 100.0 97.7	-	-
G	-	-	-	-	-
U	- - 1.5	-	-	-	-
-	-	-	-	-	-

2 H 1005 — WC 959

trans	A	C	G	U	-
A	- 0.1 6.0	-	-	-	-
C	87.5 98.3 91.7	4.2 - -	- 1.6 2.3	-	-
G	-	-	-	-	-
U	4.2 - -	-	-	4.2 - -	-
-	-	-	-	-	-

3 SE 1008 — WC 962

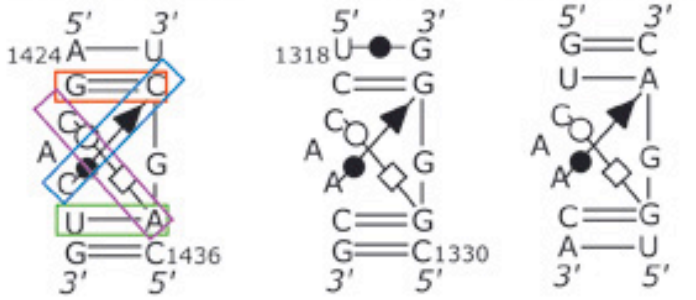
cis	A	C	G	U	-
A	-	- - 95.5	-	-	-
C	-	100.0 100.0 2.3	-	- - 1.5	-
G	-	- - 0.8	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

4 WC 963 — WC 1005

cis	A	C	G	U	-
A	-	20.8 - -	4.2 - -	66.7 98.4 97.7	-
C	-	-	4.2 - -	-	-
G	-	- 1.6 2.3	-	-	-
U	4.2 - -	-	-	-	-
-	-	-	-	-	-

23S C50

H.marismortui *E.coli* *S.cerevisiae*



1 WC 1425

WC 1439 \ cis	A	C	G	U	-
A	- 0.1 0.8	- - 3.0	- - 1.5	- 6.0 64.7	-
C	-	- 0.1 -	100.0 68.6 3.8	- - 0.8	-
G	-	- 24.6 15.0	-	-	-
U	- 0.6 8.3	-	- - 0.8	- - 1.5	-
-	-	-	-	-	-

2 H 1437

WC 1426 \ trans	A	C	G	U	-
A	-	-	- - 2.3	-	-
C	95.8 4.0 1.5	- 4.2 0.8	4.2 37.1 13.5	-	-
G	-	- 39.8 -	- 0.1 65.4	- 0.1 0.8	-
U	-	-	- - 14.3	- 14.7 -	- - 1.5
-	-	-	-	-	-

3 SE 1439

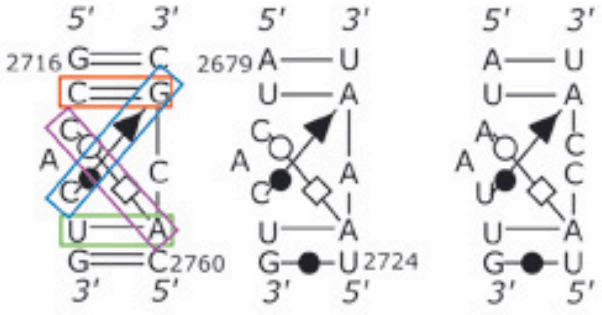
WC 1428 \ cis	A	C	G	U	-
A	- 6.1 66.2	91.7 68.4 3.0	- 24.6 15.0	- 0.6 10.5	-
C	- - 0.8	8.3 0.1 1.5	-	-	-
G	- - 0.8	- 0.1 -	-	-	-
U	- - 2.3	-	-	-	-
-	-	-	-	-	-

4 WC 1429

WC 1437 \ cis	A	C	G	U	-
A	-	4.2 - -	-	91.7 4.0 1.5	-
C	-	-	- 44.0 0.8	-	-
G	-	4.2 37.3 95.5	-	-	-
U	- 14.7 -	-	- 0.1 0.8	-	-
-	-	- - 1.5	-	-	-

23S C96

H. marismortui *E. coli* *S. cerevisiae*



1 WC 2717

WC 2763 \ cis	A	C	G	U	-
A	-	- - 20.4	-	- 17.4 41.6	-
C	-	-	- 0.4 -	-	-
G	-	100.0 82.2 38.1	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

2 H 2761

WC 2718 \ trans	A	C	G	U	-
A	- - 78.8	-	-	-	-
C	95.8 99.8 17.7	-	4.2 0.3 1.8	-	-
G	- - 0.9	-	- - 0.9	-	-
U	-	-	-	-	-
-	-	-	-	-	-

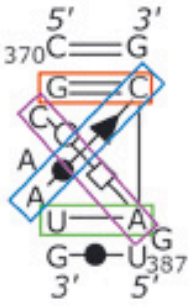
3 SE 2763

WC 2720 \ cis	A	C	G	U	-
A	-	-	- - 2.4 3.5	-	-
C	- 17.4 2.7	- 0.4 -	100.0 79.9 10.6	-	-
G	- - 0.9	-	- - 0.9	-	-
U	- - 58.4	-	- - 23.0	-	-
-	-	-	-	-	-

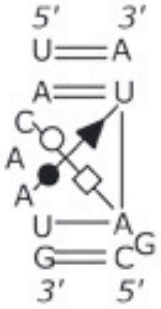
4 WC 2721

WC 2761 \ cis	A	C	G	U	-
A	- - 0.9	- - 0.2 -	-	95.8 99.5 96.5	-
C	-	-	-	-	-
G	-	4.2 0.2 2.7	-	-	-
U	-	-	-	-	-
-	-	-	-	-	-

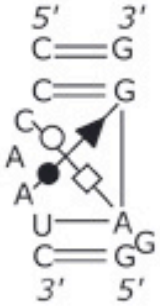
T.thermophilus



H.marismortui



S.cerevisiae



16S C15

1 WC 371

WC 390 \ cis	A	C	G	U	-
A	-	- - 0.4	-	- - 3.6 0.3	-
C	- - <	- - <	54.2 41.4 0.3	-	-
G	- - <	8.3 - 98.2	-	- - 0.1	-
U	37.5 54.8 0.1	- - 0.1	-	-	-
-	-	-	-	-	-

2 H 389

WC 372 \ trans	A	C	G	U	-
A	- - <	- - <	-	-	-
C	100.0 100.0 98.8	-	- - 0.3	- - 0.1	-
G	- - <	- - 0.4	-	-	-
U	- - 0.2	-	-	-	-
-	- - 0.1	-	-	-	-

3 SE 390

WC 374 \ cis	A	C	G	U	-
A	- 3.9 0.4	54.2 41.4 0.4	8.3 - 73.9	25.0 54.8 0.1	-
C	- - 0.2	-	- - 23.3	12.5 - <	-
G	-	-	- - 0.1	-	-
U	-	-	- - 1.0	-	-
-	-	-	-	-	-

4 WC 375

WC 389 \ cis	A	C	G	U	-
A	-	- - 0.1	-	100.0 100.0 99.2	- - <
C	-	-	- - 0.4	- - <	-
G	-	-	-	- - 0.3	-
U	- - <	-	-	- - <	-
-	-	-	-	-	-

2.2.3. Discussion

Un certain nombre de notions sur la structure générale de l'ARN ont été décrites dans l'article précédent. Nous avons, avant de procéder à l'analyse des motifs dans les alignements d'ARN, proposé une définition des paires de bases non Watson-Crick et de leur classification, de l'isostérie et des matrices d'isostérie, de la séquence signature d'un motif et du motif ARN. Comme nous l'avons vu dans l'introduction générale, il existe différents concepts définissant un motif ARN dont un se basant sur la conformation du squelette sucre-phosphate. En 1980, Olson et collaborateurs, dans la suite du concept de nucléotide rigide de Sundaralingam, ont proposé une représentation du squelette sucre-phosphate sur la base d'angles de torsion ω et ω' coupant ainsi le nucléotide en deux blocs (C4'-C5'-O5'-P et C4'-C3'-O3'-P) facilitant l'analyse de molécules complexes comme l'ARNt (Sundaralingam, 1969; Olson, 1980). L'utilisation des ces deux pseudoangles de torsion a permis la représentation des conformations de nucléotides dans un diagramme à deux dimensions ressemblant au diagramme de Ramachandran chez les protéines.

C'est sur ce concept de pseudoangle de torsion que se base l'identificateur automatique de motifs COMPADRES (Comparative Algorithm to Discover Recurring Elements of Structure) de Wadey et Pyle (2004). Ce logiciel se base sur les angles η ($C4'_{i-1}$-P_i-$C4'_i$-P_{i+1}) et θ (P_i- $C4'_i$-P_{i+1} - $C4'_{i+1}$), similaires aux angles ω et ω', pour trouver et classer les "nouveaux motifs" d'ARN d'une structure 3D d'ARN (Wadley & Pyle, 2004). Selon Wadey et Pyle, un motif ARN doit répondre à l'une des conditions suivantes (i) il est superposable à un autre fragment de structure d'ARN (ii) il interagit de manière particulière avec d'autres éléments de la structure d'ARN. C'est la première condition qui a été utilisée dans leur recherche de nouveaux motifs à l'aide du logiciel COMPADRES. Quoique très utile pour l'analyse de structures tridimensionnelles existantes, cette définition du motif ARN qui ne relie pas la structure tridimensionnelle à la séquence, ne permet pas la déduction de structure à partir d'un alignement de séquences homologues. Les auteurs expliquent que la conformation du squelette sucre-phosphate des nouveaux motifs trouvés est conservée sans qu'il y ait conservation de séquence ou de structure secondaire. Cette conception du motif est pour nous difficile à accepter. En effet, la structure du squelette sucre-

phosphate d'un motif prend le chemin imposé par la géométrie des paires de bases qui le composent donc à un chemin donné correspond une succession particulière de paires non Watson-Crick.

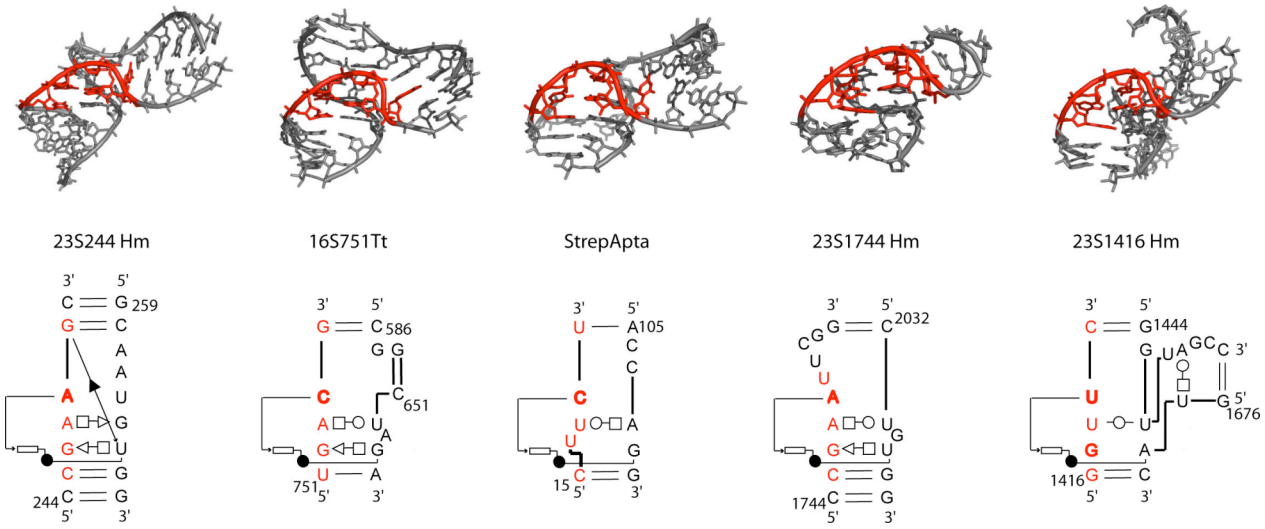


Figure 30 : Les cinq exemples du nouveau motif Ω proposé par Wadley et Pyle (2004). En haut, la représentation tridimensionnelle met en évidence la conservation de la conformation d'une portion de cinq nucléotides (en rouge). En bas, représentation à l'aide des symboles géométriques des interactions dans lesquelles les nucléotides du motif (en rouge) sont impliqués.

Afin de repérer le réseau potentiel de paires non Watson-Crick caractéristique de deux nouveaux motifs (motifs π et Ω) proposés dans cet article, nous avons examiné les structures tridimensionnelles auxquelles ils appartiennent. Nous avons pris en compte les appariements formés par les cinq nucléotides du brin défini comme motif. Le tournant Ω est une portion de brin, formé de cinq bases successives, dont le squelette sucre-phosphate adopte une conformation ressemblant à la lettre « Ω » (Figure 30). Tous les exemples de tournant Ω présentés dans l'article montrent les mêmes caractéristiques géométriques : la première base forme toujours une paire Watson-Crick, la deuxième base est impliquée dans une paire *trans* Hoogsteen/Sucre dans trois cas sur quatre, la quatrième base, toujours en *syn*, est extrudée de manière à former une paire *cis* Watson-Crick avec l'avant dernière base du brin complémentaire au brin motif. La troisième base forme des appariements variables mais tous sont orientés en *trans*. Ainsi, une succession de paires non

Watson-Crick particulières est commune aux différents fragments présentant un motif tournant Ω. Mais le tournant Ω peut être apréhendé comme un sous motif car il appartient parfois à un tournant K (KT-11) dans le cas de 23S244Hm ou une jonction triple ou une bulle interne. Nous avons trouvé que les deux nouveaux motifs possèdent, à part une conformation identique du squelette, des paires de bases non Watson-Crick caractéristiques. Cette succession ordonnée de paires non Watson-Crick est en accord avec notre définition opérationnelle du motif ARN. Elle permettra la détermination de la séquence signature de chacun des motifs et à terme leur utilisation pour repérer ces motifs dans un alignement de séquences homologues. La représentation sous la forme d'un simple brin ne permet pas de réaliser que le motif défini est en fait une portion d'un motif plus grand. Nous pensons que la définition d'un motif à partir de la conformation d'un seul des brins qui le composent ne donne que des informations partielles et biaisées sur les contraintes agissant et ne peut pas être utilisée à des fins prédictives. Le concept d'isostérie au contraire peut servir de base pour extraire les contraintes qui imposent une structure à un motif et les utiliser pour déduire de la séquence une structure tridimensionnelle.

En fait, on perçoit ici l'importance de la définition d'un motif. Le but n'est pas de partir à la chasse aux motifs dans les structures cristallographiques publiées pour, ensuite, les classer et leur choisir un nom. Le but est de trouver une définition opérationnelle du motif qui permette de relier séquence et repliement. C'est à dire une définition qui prend en compte des éléments identifiables et spécifiques du motif dans les structures tridimensionnelles existantes, qui pourront être utilisés pour repérer dans une séquence d'ARN de structure inconnue, le même motif. Quoi d'autre que les appariements des bases d'une structure pourrait dicter le chemin que prend le squelette sucre-phosphate ? Les appariements non Watson-Crick dont la nature, le nombre, l'orientation et l'ordre sont spécifiques à chaque motif, et qui possèdent tous une matrice d'isostérie spécifique, constituent l'élément identifiable et spécifique qu'il faut exploiter.

Le motif tournant K qui est ubiquitaire, a été initialement identifié dans la tige 5' du petit ARN nucléaire U4 et dans l'ARNr 23S (Nottrott et al., 1999; Reuter et al., 1999; Ban et al., 2000; Klein et al., 2001; Klein et al., 2004). Dans l'article précédent nous définissons ce motif comme la succession de cinq paires

de bases : un appariement Watson-Crick, deux appariements *trans* Hoogsteen/Sucre et deux appariements *trans* Sucre/Sucre. Dans de nombreux articles le motif est décrit comme deux hélices, l'une canonique l'autre non canonique, séparées par une boucle interne asymétrique de trois nucléotides qui impose un angle de 90° entre les deux hélices. Cette définition qui décrit grossièrement la conformation du motif est source d'erreur quant à l'identification de tournants K dans de nouvelles structures. Ainsi, Strobel et collaborateurs (2004) proposent que le coude présent dans la structure de l'intron de groupe I de Azoarcus entre les hélices P9.0 et P9 soit une nouvelle sorte de tournant K tournant à droite (Figure 31). Il propose de l'appeler tournant K inverse en opposition au K-turn classique qui tourne à gauche. Cette appellation nous pose problème car la succession des paires de bases qui définit pour nous un tournant K n'est pas présente dans ce tournant inverse. La première paire *trans* Hoogsteen/Sucre est remplacée par un appariement *trans* Hoogsteen/Hoogsteen et les deux appariements *trans* Sucre/Sucre sont absents. Ce sont justement ces deux derniers appariements qui assurent le positionnement de l'hélice non canonique (NC) vers le petit sillon de l'hélice canonique (C). Dans le tournant K inverse, qui ne présente pas ces interactions, l'hélice NC est orientée vers le grand sillon de l'hélice C.

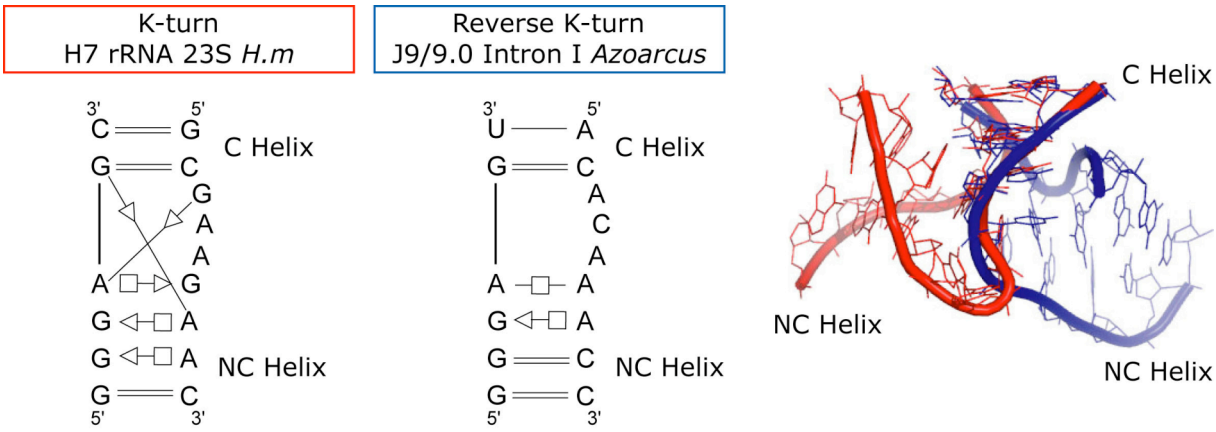


Figure 31 : Le tournant K et le tournant K inverse. A gauche sont représentées à l'aide de la nomenclature, les structures du tournant K de l'hélice 7 (H7) de l'ARNr 23S et de la bulle interne qui sépare les hélices P9 et P9.0 dans l'intron de groupe I de *Azoarcus*. A droite les hélices C (Canoniques) des deux structures 3D du tournant K (orange) et tournant K inverse (bleu) ont été superposées.

Nous avons identifié dans les structures crisallographiques des ARNr 16S et 23S d'autres tournants présentant glogalement la même conformation que le tournant K inverse de l'intron I de *Azoarcus*. L'appariement *trans* Hoogsteen/Sucre présent dans ces structures est le seul point commun avec les tournants K classiques. Les tournants inverses montrent des contraintes différentes des tournants K classiques. Nous pensons que l'utilisation de la même appellation "tournant K" pour deux motifs sous-tendus par des paires différentes rend la définition des motifs inutilement confuse. Il serait plus juste de parler d'analogies entre motifs.

La définition d'un motif sur sa conformation globale pose un autre problème. En effet, un motif peut présenter les appariements caractéristiques d'un motif connu mais avoir une conformation globale différente. Nous avons vu plus haut que dans la définition "conformationnelle" le tournant K est défini comme deux hélices C et NC séparées par une boucle asymétrique. Or, récemment, il a été montré que la boucle de la tige-boucle terminale d'un ARN guide à boîte C/D d'archae qui fixe la protéine L7Ae pouvait se replier en réalisant les appariements caractéristiques du tournant K (Nolivos et al., 2005). Nous avons conforté cette hypothèse par la réalisation d'un modèle présentant les appariements caractéristiques du tournant K (Figure 32). Le tournant K ne peut donc pas être défini en fonction des hélices NC et C qui l'entourent car elles ne sont pas systématiquement présentes.

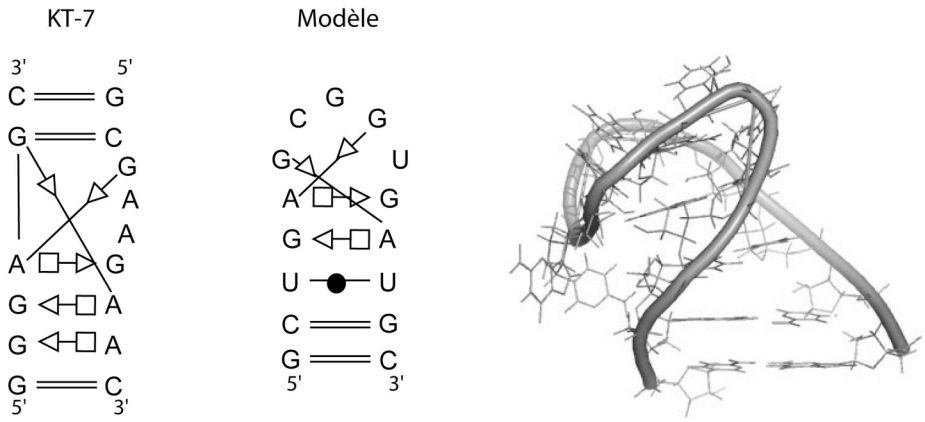


Figure 32: Modèle de la tige-boucle terminale présentant les appariements du tournant K. A gauche, représentations bidimensionnelles du tournant K standard de l'hélice 7 de l'ARNr 23S de *H.marismortui* (KT-7) et du modèle de la tige boucle de l'ARN non codant fixant la protéine L7AE (Modèle) (Klein et al., 2004). A droite, représentation 3D du modèle.

Un exemple de motif qui présente des similarités avec le tournant K est le motif en A mineur typeI/typeI que nous avons décrit dans la partie sur les motifs ARN. Figure 33 sont représentés un motif tournant K standard (KT7) et le consensus d'un motif A mineur typeI/typeI. Les appariements *trans* Hoogsteen/Sucre et *trans* Sucre/Sucre sont conservés dans les deux motifs. Il manque dans le motif tournant K les appariements *cis* Sucre/Sucre et la première paire Watson-Crick qui interagit avec l'adénine en 3' de la paire *trans* Hoogsteen/Sucre. Nous avons, dans la partie sur les motifs en A mineur, proposé des matrices de séquences signatures du motif A mineur typeI/typeII. La matrice correspondant à l'appariement *trans* Sucre/Sucre, dans lequelle est impliquée l'adénine en 3', n'autorise que des appariements où une adénine est impliquée. Cette caractéristique est retrouvée dans les variations consensus de l'appariement *trans* Sucre/Sucre du motif tournant K. La détermination des variations de séquences à partir de l'alignement nous a permis néanmoins de déterminer la séquence signature de l'appariement *trans* Sucre/Sucre dans le cadre du tournant K. Toutes les variations de séquences qui apparaissent pourront être utilisées lors de la prédiction de motif dans un alignement de séquences ou dans une séquence ARN.

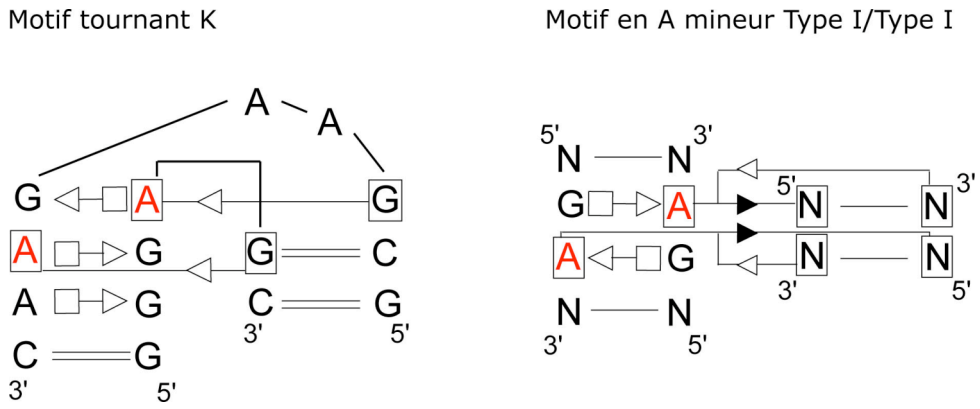


Figure 33 : Tournant K et Motif en A mineur Type I/Type I. Les deux adénines impliquées dans les interaction trans Sucre/Sucre sont colorées en rouge.

2.2.4. Revue 1 : "The building blocks and motifs of RNA architecture"

Neocles Leontis, Aurélie Lescoute & Eric Westhof

Current Opinion in Structural Biology

Soumis

Title: **The Building Blocks and Motifs of RNA Architecture**

Short Title: **RNA 3D Building Blocks**

Authors: Neocles B. Leontis,[1] Aurelie Lescoute,[2] and Eric Westhof[+2]

[+]Corresponding author

Affiliations:
[1]Department of Chemistry and Center for Biomolecular Sciences, Bowling Green State University, Bowling Green, OH 43402, USA.

[2]Institut de Biologie Moléculaire et Cellulaire du CNRS, UPR 'Architecture et réactivité de l'ARN', Université Louis Pasteur, 15 rue René Descartes, 67084 Strasbourg, France.

Emails: Neocles Leontis: leontis@bgnet.bgsu.edu
Aurelie Lescoute: a.lescoute@ibmc.u-strasbg.fr
Eric Westhof: e.westhof@ibmc.u-strasbg.fr

Keywords: RNA Motifs, RNA 3D Structure, Conformation, Base-pairing

Summary

RNA 3D motifs can be defined broadly as recurrent structural elements, observed in RNA atomic-resolution structures and comprising multiple intra-molecular RNA-RNA interactions. They constitute the modular building blocks of RNA architecture, which is organized hierarchically. Recent work focuses on analyzing RNA backbone conformations to identify, define, and search for new instances of recurrent motifs in 3D structures. While some authors propose that recurrent RNA strand segments having a characteristic backbone configuration qualify as independent motifs, we show that to identify modular motifs one must take into account the larger structural context of such chain segments. This is consistent with the biologically relevant motivation, which is to identify RNA structural characteristics that are subject to sequence constraints and which, thus, relate RNA architectures with sequences.

Introduction

What is an RNA Motif?

No single definition exists for RNA motifs, since they can be proposed and analyzed at different levels of RNA structure. As discussed in a previous review, RNA motifs can be broadly defined as recurrent structural elements, subject to constraints [1]. The present review is complementary to a recent review of new high-resolution RNA structures that exhaustively catalogued new and recurrent 3D motifs appearing in these structures [2]. Therefore, we do not attempt to comprehensively discuss each newly reported motif, but rather aim to critically review evolving notions of recurrent RNA motifs in the context of RNA function and evolution and how to identify, find and classify them.

Types of RNA Motifs

We can distinguish two main classes of motifs – those that operate at the level of RNA sequence and those that entail a specific 3D structure, characterized by a set of three-dimensional coordinates. An example of a sequence motif is the Shine-Dalgarno sequence of bacterial mRNAs or the Sm binding sites in some eukaryotic non-coding RNAs [3]. At an intermediate level of analysis, the secondary structure (2D) of an RNA is prominent because it can be calculated quite accurately from sequence information, usually through a combination of thermodynamic and comparative sequence analyses [4]. At the level of secondary structure, the RNA double helix is the fundamental motif. Once helices are specified, other motifs become apparent, which at the level of secondary structure are classified as hairpin (or terminal) loops, internal loops (including bulges), and multi-helix or junction loops. However, this description of RNA structure is incomplete as it takes no account of non-Watson-Crick basepairing and most tertiary interactions that stabilize the 3D fold.

Secondary Structure Motifs

How much can we retrieve from analysis of secondary (2D) structures? Zorn et al. calculate and compare frequency distributions for Watson-Crick (WC) paired nucleotides in helical stems and for nominally unpaired nucleotides in hairpin, internal and junction (multi-helix) loops, as they appear in the secondary structures of the 16S and 23S ribosomal RNAs [5]. Thus, they treat all bases in "loops" as unpaired, even though a large fraction of these form non-WC basepairs, as is evident from high-resolution 3D structures, which have been available since 2000 [6-8]. Such an approach restricts the definition of RNA motif to secondary structure, and is not necessarily connected to RNA function and evolution as 3D structure is. A further point is that in such an analysis, all junctions, regardless of the number of helices, are lumped together as the same motif. In fact, it is well-known that stable 4-way junctions can be constructed without "unpaired" bases, whereas stable, geometrically defined 3-way junctions require "extra" bases forming stabilizing non-WC basepairs [9,10].

Motifs considered at the level of secondary structure are the focus of a recent study which employs RNAMotif, a widely used secondary structure definition and search algorithm [11], to search for RNA aptamers in genomic sequences [12]. While these

authors claim to use the available 3D information for the search, having chosen aptamers for which the corresponding x-ray crystal structures have been solved to test their search algorithm, in fact most of the 3D information in these structures is ignored by the search models employed so that the search is conducted essentially at the level of 2D motifs.

The ability to calculate the probabilities of functional motifs occurring in libraries of random sequence RNA molecules as a function of library size, sequence length, and base composition is useful for planning *in vitro* selection (SELEX) experiments and in theoretical considerations of the role of RNA molecules in the origin of life. Knight and co-workers have approached these issues computationally [13-15]. The outcome of such calculations depends critically on how motifs are defined. While in their most recent contribution, the presence of supporting double helices is taken into account, the actual substrate binding or catalytic motifs are treated as single-stranded motifs, subject to independent sequence constraints. The presence in such motifs of non-Watson-Crick basepairs is not taken into account and it is difficult at this time to assess the effects of such approximations on the statistical outcome.

Graph theory has been used to represent RNA secondary structure in various ways for some time [16,17]. In a series of recent articles, Schlick and co-workers have promoted the use of two types of graphs, tree graphs and dual graphs to represent RNA 2D structure [18-21]. In tree graphs, edges represent helices and vertices hairpin, internal and junction loops. In dual graphs this is reversed. Dual graphs can also represent pseudoknots, which tree graphs cannot do. For each known RNA, both representations are available on the RNA-As-Graphs website [19], which catalogues these graphs according to the number of vertices (V) and the topological complexity, which is identified with the 2^{nd} smallest eigenvalue of the Laplacian matrix of the graph. In addition, graph theory is used to enumerate possible graphs having the same value of V, to systematically catalog possible RNA secondary structures. However, only graphs with the same V can be directly compared for topological complexity using the second eigenvalue. As homologous RNA molecules, especially large functional RNAs like Group I introns and ribosomal RNAs, vary widely in the number of stems and loops, and therefore in their V-values, they cannot be compared using this approach.

Karklin et al. introduced a labeled dual graph representation of RNA secondary structures and developed a similarity measure to compare and distinguish RNA molecules belonging to different families of homologs [22]. In such graphs, helices are represented as nodes labeled with the number of Watson-Crick basepairs while edges are the nominally single-stranded regions that connect helices to each other (hairpin loops, internal loops, multi-helix loops), labeled with the number of nucleotides they comprise. As the authors point out the accuracy of this approach depends on the accuracy of the secondary structures. However, a further implicit assumption is made in this approach, that homologous RNA molecules are conserved fundamentally at the level of secondary structure. In fact, 2D structure is less conserved than 3D structure, and it is the 3D structure of an RNA molecule that is subject to natural selection. A dramatic example of this was recently revealed with the publication of x-ray crystal structures for the

specificity (S) domain of A and B type RNase P molecules [23-25]. The A and B architectures present similar features, but the two secondary structures display significant differences [26]. Furthermore, detailed structural analyses of internal and hairpin loop motifs shows that motifs with different numbers of nucleotides can adopt similar 3D structures, except for variations in the number of looped out bases. Examples include T-loops [27] and simple internal loops consisting of a single *trans* Hoogsteen/Sugar-edge (sheared) basepair and one to three unpaired, looped out bases [28].

Representations of RNA 3D Structure

Different representations can be used to describe molecular structure information [29]. The most basic representation, used by the 3D structure databases, is the Cartesian coordinates of individual atoms. Other representations can be derived from this representation. However, the large number of variables makes this representation awkward when comparing structures or searching for recurrent motifs. Internal coordinates (torsion angles) significantly reduce the number of variables and remove the need to align structures to a common coordinate system, and therefore their use has found favor among a number of workers (see below). A further simplification is the use of pseudo-torsion angles [30]. Another approach using distance matrices has also been applied [29]. Finally, symbolic representations involving higher levels of abstraction have been described [31,32]. These approaches seek to capture the biologically most relevant structural information at the appropriate granularity, by averaging over the minor variations of structure typical for non-covalent interactions such as hydrogen-bonding, to identify features that connect 3D structure explicitly to sequence data [33].

As pointed out by Reijmers et al. (2001), the outcome of clustering experiments depends largely on the way the data are represented. Huang et al. used the 3D coordinates of 15 atoms per RNA residue, including 3 base atoms and all the backbone and sugar atoms, to calculate the RMSD distance between two RNA fragments of the same length after they are superposed in 3D [34]. They applied the method to compare all hairpin loops of fixed size in a set of RNA 3D structures including the large rRNAs. The RMSD distances between all pairs of sequence segments were used to cluster the motifs using UPGMA to produce dendograms of the hairpin loop structures. Because the algorithm is limited in its ability to find motifs involving different chain segments (composite motifs) or insertions, the authors also clustered subsets of nucleotides belonging to longer hairpin loops and recovered GNRA or UNCG tetra-loops with inserted nucleotides, i.e. “penta-“ or “hexa-loops” that are GNRA or UNCG hairpins with insertions in characteristic positions.

Harrison and co-workers describe a reduced vectorial representation of RNA 3D structure designed to convert the problem of searching for recurrent 3D motifs to the subgraph isomorphism problem, for which algorithms are known from graph theory [35]. These methods were first developed for searching substructures in libraries of structures of small molecules, then applied to proteins and carbohydrates, and recently to RNA [36]. For RNA structure searching, two pairs of pseudo atoms forming two vectors represent each base. These vector pairs compose the nodes of labeled graphs, one node per base. The relative positions of the bases in the 3D structure are captured

by edges connecting the nodes and labeled with the distances between the start and end points of the vectors comprising each node. The problem of searching for a 3D motif is thus reduced to the problem of finding subgraph isomorphisms of graphs representing query motifs in graphs representing structures in the RNA database. Harrison et al. (2003) use their approach to search for non-Watson-Crick basepairs and other small motifs. As the Ullman algorithm they use scales with n factorial (n!), where n is the number of nodes in the query motif (subgraph), it is not clear how practical this approach is for searching larger motifs representing entire hairpin or internal loops.

Classifying Backbone Conformations

Local motifs (i.e. hairpin loops and internal loops) result in distinct and reproducible backbone conformations. Therefore, several groups have focused on analyzing and classifying RNA backbone conformations to identify new motifs and search for recurrent motifs in complex high-resolution RNA 3D structures.

Schneider and co-workers analyzed and classified the backbone conformations of the 5S and 23S rRNA of the 50S ribosomal subunit of *H. marismortui* using Fourier averaging of the six 3D distributions of torsion angles [37]. They identified 18 non-A-type conformations and 14 A-RNA related conformations and determined their corresponding torsion angles. Hershkovitz and co-workers binned the continuous torsional information into a limited number of discrete values and used pattern-recognition methods to find structural recurrences. They found they could represent backbone conformations using a small alphabet as four torsion angles contain the bulk of the structure information [38]. Richardson and co-workers applied quality-filtering techniques to reduce noise levels in the backbone torsion angle distributions from an 8,636-residue RNA database. The signal that emerged for half-residue torsion angle distributions for alpha-beta-gamma and for delta-epsilon-zeta was plotted and contoured in 3D. About a dozen distinct peaks were observed in the distributions and combined in pairs to define complete RNA backbone conformers. The RNA backbone conformations were reparsed into base-to-base "suites" comprising seven variables, with sugar pucker specified at both ends. Their analysis produced a small library of 42 RNA backbone conformer [39,40]. Thus, all three of these independent methods of analysis show that the torsional angles of the RNA backbone are quite constrained as to the number of distinct conformations that can result without steric clashes, a concept that was already apparent in the early days of nucleic acid stereochemistry [41,42].

The use of the pseudo torsion angles η ($C4'_{i-1}$-P_i-$C4'_i$-P_{i+1}) and θ (P_i-$C4'_i$-P_{i+1}-$C4'_{i+1}$) – which in older notation were designated ω_v and ω_v' -- makes possible a reduced representation of an RNA molecule's backbone configuration – the "RNA worm" – a three-dimensional trajectory described by η, θ and the position of each nucleotide in the sequence as the coordinates [30,43]. Two-dimensional η-θ plots correspond formally to Ramachandran plots of ϕ–ψ torsion angles used to analyze protein conformation. The program *Primos* was written to search RNA 3D structures for recurrent RNA worms [43]. Szép et al. used *Primos* to identify additional occurrences of an RNA strand segment having a sharp turn that they observed an oligonucleotide x-ray structure and which they called "hook-turn" [44]. All hook-turns identified in large

RNA structures occur where the strands of a duplex separate so that they can interact with other RNA regions. One strand doubles back to interact with itself in the 5'-helical region. A careful analysis of hook-turns reveals other characteristic structural elements involving base-base or base-sugar interactions. A detailed analysis of these motifs and others will be the subject of a future report.

A new computer program, COMPADRES (Comparative Algorithm to Discover Recurring Elements of Structure), implements a novel algorithm, based on the RNA-worm representation of the backbone, to identify new recurrent backbone conformations of RNA molecules in the structural database without prior knowledge [45]. The algorithm compares all short RNA worms in the structure database against each other to discover recurrences within user-supplied tolerances. Applying this algorithm, these authors identified four new recurrent backbone conformations comprising five or more nucleotides, which they named for their shapes: π-turns (Type 1 and Type 2), Ω-turns, α-loops, and C2'-*endo* mediated flipped adenosine motifs. The authors claim no common primary or secondary structure features exists among the strand segments assigned to three of these types, the π-turns, the Ω-turns, or the α-loops. Figure 1 displays the structural annotation for each Ω-turn reported by Wadley and Pyle [45] in the context of its interactions with other RNA strand segments. To prepare Figure 1, each Ω-turn was visually inspected in its structural context and annotated for base-pair [31] and base-stacking [46] interactions. This analysis clearly shows that motifs comprising Ω-turns share other common structural characteristics and are subject to sequence constraints. Thus, the first base of each Ω-turn forms a Watson-Crick basepair. In three out of the four cases reported, the second base forms a *trans* Hoogsteen/Sugar-edge (sheared) pair [45]. In the fourth case, the corresponding base (G1417 in *H. m.* 23S rRNA) is in the *syn* glycosidic configuration, and, were it to rotate back to the more common *anti* configuration, it would form the same type of basepair with A1678. For each Ω-turn, the fourth base is in the *syn* glycosidic configuration and is extruded from the helix formed by the preceding nucleotides so as to form a *cis* Watson-Crick basepair with the base belonging to the other strand that pairs with the second base of the Ω-turn strand. The third base also basepairs, in a variable fashion, but always forming a *trans* basepair. Figure 1 shows that Ω-turns, viewed in their contexts, also form characteristic ordered arrays of non-Watson-Crick basepairs, even though embedded in various kinds of motifs, including a K-turn, an internal loop, or within three-way junctions. The inescapable conclusion is that none of the RNA chain segments exhibiting an Ω-turn conformation forms an independent structural unit, in the way a GNRA hairpin loop does, for example. While at this stage, there is no agreement in the field as to where to draw the line between a motif and a submotif, with some workers maintaining that any substructure that occurs more than once in the structure database, and is "large enough to be interesting", qualifies as an RNA motif in its own right [45]. The present analysis clearly shows that Ω-turns are best considered in a larger context, and are thus effectively sub-motifs. This example supports the notion that a criterion of independence and modularity should be applied to distinguish motifs from submotifs [1].

Motifs defined by global features

In the striking crystal structure of the Azoarcus Group I intron [46,47], Strobel and coworkers noticed a sharp bend between two helical segments, which they named "Reverse Kink-turn" [48]. Such a naming implies a close relationship with the previously named Kink-turn [49]. In fact, the only common feature of these two motifs is that they can produce a sharp bend or kink between two double-stranded elements, but otherwise the name is misleading because the similarity stops here. The Kink-turn bends toward the minor/shallow groove and is stabilized by base-base A-minor motifs. The Reverse Kink-turn bends toward the major/deep groove and is not stabilized by base-base interactions. Each is, thus, characterized by a different set of non-Watson-Crick basepairs involving bases in the internal loop. Annotated drawings comparing the structures of representative kink-turns and reverse kink-turns are shown in Figure 2. This shows that these motifs are fundamentally different.

The SCOR classification

The SCOR database aims to comprehensively classify local RNA motifs appearing in 3D structures – i.e. internal and hairpin loops [50,51]. While SCOR features 3D data, it is in fact organized using categories defined by secondary structure motifs (http://scor.lbl.gov/scor.html). Thus, only local versions of motifs are provided, omitting structurally similar composite motifs that share the same core of basepairing and stacking interactions. Strikingly, much 3D information is ignored; thus all non-Watson-Crick basepairs are represented in schematic diagrams the same way using a single dashed line and the geometric type of each non-Watson-Crick basepair is ignored in classifying the motifs. For example, internal loops containing n non-Watson-Crick basepairs, where n = 1, 2, 3, …, are all classified together for a given value of n, regardless of the nature or the order of the component non-Watson-Crick basepairs. The result is that quite heterogeneous motifs are grouped together while *bona fide* similarities are overlooked.

Basepairing Patterns and RNA Motifs

Keeping in mind that sequences are the more fundamental biological data, other workers have focused attention on base-pairing patterns and their symbolic representation, and have pointed out that defined backbone configurations are necessary to form ordered arrays of non-Watson-Crick [1]. In 2001 systematic geometry-driven nomenclature was proposed for non-Watson-Crick RNA basepairs, along with easy to remember annotations for drawing schematic diagrams [31]. Using the nomenclature, all observed and chemically allowed basepairs can be classified in geometric families and isosteric subfamilies that identify those base combinations that can substitute during evolution while preserving 3D structure [52]. Moreover, this approach makes it possible to write computer programs to identify automatically and classify basepairs in 3D structures [53-55]. Alternative classifications use names that are not related directly to the pairing geometry, do not provide ways to annotate 2D diagrams or to automate basepair identification in 3D structures, and neglect H-bonds involving the 2'-hydroxyl group [56,57].

It is now apparent from crystal structures that the driving force for folding RNA molecules is the continuous stacking of bases and some arrays of non-Watson-Crick base

pairs are prevalent because of favorable stacking patterns coupled with standard and energetically satisfying sugar-backbone conformations. Significant efforts are still necessary to reconcile these alternative approaches so as to define and classify the major types of recurrent backbone conformations and associate them with specific occurrences of non-Watson-Crick basepairs.

The 3D structure of the RNA double helices is very regular compared to DNA helices. Sequence dependent differences are more subtle for RNA than for DNA, and are due primarily to near- or non-isosteric basepair substitutions, including wobble pairs (G/U or A/C), homo-purine (A/G or A/A) and homo-pyrimidine (C/U, U/U, and C/C) pairs, which significantly distort the backbone conformation. Just as RNA helices require the stacking of two or more adjacent Watson-Crick basepairs, RNA motifs result from combinations of two or more, usually stacked, non-Watson-Crick basepairs. In this view, individual non-Watson-Crick basepairs constitute the building blocks of RNA motifs, but without forming by themselves integral motifs.

At the level of 3D motifs, attention is therefore directed to the single-stranded regions of RNA molecules, hairpin or internal loops and multi-helix junctions loops, more so than to the helical regions. We now have sufficient numbers of high-resolution structures to conclude that most of the bases in these "loop" regions of structured RNA molecules are paired in non-Watson-Crick geometries and stacked to form specific structures – 3D motifs. Because such a large proportion of bases in "loops" are base-paired, a useful definition for RNA motifs at this level is "an ordered array of non-Watson-Crick basepairs under constraints." It is worth noting here that, depending on the crystallographic resolution, the presence or absence of individual H-bonds may be difficult to ascertain. It may therefore be difficult to distinguish and classify motifs on the sole basis of H-bonds. As 3D motifs may be local or composite, both types should be included in searches and classifications. Local motifs involve exclusively nucleotides that are close to each other in the secondary structure, that is, they belong to the same hairpin or internal loop. Composite motifs are formed when three or more strands converge to form an ordered array of non-Watson-Crick basepairs.

Consensus sequences and motif signatures:
Consensus sequences are used frequently to describe protein and RNA motifs. Known motifs from different sources are aligned and the frequency of each residue type is calculated for each column of the alignment and displayed as a sequence logo [58]. This is only appropriate for RNA motifs that are strictly single-stranded. To better describe RNA motifs that include Watson-Crick basepairs, Gorodkin and co-workers introduced the RNA structure logo that includes mutual information for paired positions [59]. The non-Watson-Crick basepairs that compose RNA 3D motifs are also subject to pairwise sequence constraints. A more complete description of a recurrent RNA 3D motif, the sequence signature, includes information about the base-pairs that can substitute at paired positions and the positions at which insertions and deletions occur. A recent study of two recurrent 3D motifs, the kink-turn and the C-loop, analyzes the sequence variations of all occurrences of these motifs known from crystal structures and derives sequence signatures of this type for each motif [33]. This paper demonstrates the utility of

Isostericity Matrices [52] for analyzing RNA motifs comprising non-Watson-Crick basepairs and outlines the steps for productively iterating motif analysis and sequence alignment. The flow chart shown in Figure 3 illustrates the role of Isostericity Matrices in 3D structural analysis, 3D motif identification and classification, sequence analysis to produce accurate structure-based sequence alignments, and 3D modeling, all with the goal of increasing understanding of RNA function and evolution.

Conclusions

Biological data are fundamentally sequence data. Given the ease of obtaining sequence data and the difficulty of determining 3D structures at high-resolution, there will always be more sequence data than structural data. The key challenge for RNA structural and computational biologists and bio-informaticians is to fully integrate these two types of data with a common ontology [60]. The fact that structured RNA molecules are mosaics of recurrent modular motifs means that high-resolution 3D information about one molecule may be useful in analyzing the sequences of another molecule, whether or not the molecules are homologous. For each modular 3D motif identified, we need to define the sequence constraints, as these allow one to identify the motif in other sequences. In this respect, a classification based on an analysis of secondary structures is of limited use.

With proper integration, the accumulated knowledge of 3D structures can be maximally applied to the fundamental problems of searching for non-coding RNA genes in genomic sequences and constructing 3D models for RNA molecules based on known sequences. We conclude by observing that the ultimate purposes of classification must be borne in mind to avoid unnecessary proliferation of confusing jargon that will certainly result if every recurrent element of structure at every possible degree of granularity is given a distinct name, without regard to the ways in which these elements combine to create integral structural and functional units or modules.

Acknowledgments

NBL acknowledges grant support NIH 2 R15 GM055898-03 and PRF# 42357 -AC 4 from the American Chemical Society.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

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

••2. Holbrook SR: **RNA structure: the long and the short of it**. *Curr Opin Struct Biol* 2005, **15**:302-308.

Recent crystallographic structures are reviewed. All new and recurrent RNA motifs in recent structure are identified and described.

3. Khusial P, Plaag R, Zieve GW: **LSm proteins form heptameric rings that bind to RNA via repeating motifs**. *Trends Biochem Sci* 2005, **30**:522-528.
4. Mathews DM, Zuker M: **Predictive Methods Using RNA Sequences**. In *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. Edited by Baxevanis AD, Ouellette BFF: John Wilen & Sons; 2005:144-171.
5. Zorn J, Gan HH, Shiffeldrim N, Schlick T: **Structural motifs in ribosomal RNAs: implications for RNA design and genomics**. *Biopolymers* 2004, **73**:340-347.
6. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA: **The complete atomic structure of the large ribosomal subunit at 2.4 A resolution**. *Science* 2000, **289**:905-920.
7. Schuwirth BS, Borovinskaya MA, Hau CW, Zhang W, Vila-Sanjurjo A, Holton JM, Cate JH: **Structures of the bacterial ribosome at 3.5 A resolution**. *Science* 2005, **310**:827-834.
8. Wimberly BT, Brodersen DE, Clemons WM, Jr., Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V: **Structure of the 30S ribosomal subunit**. *Nature* 2000, **407**:327-339.
9. Lilley DM: **Structures of helical junctions in nucleic acids**. *Q Rev Biophys* 2000, **33**:109-159.
10. Lescoute A, Westhof E: **Topology of three-way junctions in folded RNAs**. *Rna* 2006, **12**:83-93.
11. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R: **RNAMotif, an RNA secondary structure definition and search algorithm**. *Nucleic Acids Res* 2001, **29**:4724-4735.

•12. Laserson U, Gan HH, Schlick T: **Predicting candidate genomic sequences that correspond to synthetic functional RNA motifs**. *Nucleic Acids Res* 2005, **33**:6057-6069.

Structural features of RNA aptamers were used to search genomic sequences RNAMotif.

••13. Knight R, De Sterck H, Markel R, Smit S, Oshmyansky A, Yarus M: **Abundance of correctly folded RNA motifs in sequence space, calculated on computational grids**. *Nucleic Acids Res* 2005, **33**:5924-5935.

The probabilities of functional motifs occurring in libraries of random sequence RNA molecules are calculated at the level of secondary structure as a function of library size, sequence length, and base composition.

14. Legiewicz M, Lozupone C, Knight R, Yarus M: **Size, constant sequences, and optimal selection**. *Rna* 2005, **11**:1701-1709.
15. Knight R, Yarus M: **Finding specific RNA motifs: function in a zeptomole world?** *Rna* 2003, **9**:218-230.
16. Benedetti G, Morosetti S: **A graph-topological approach to recognition of pattern and similarity in RNA secondary structures**. *Biophys Chem* 1996, **59**:179-184.
17. Le SY, Nussinov R, Maizel JV: **Tree graphs of RNA secondary structures and their comparisons**. *Comput Biomed Res* 1989, **22**:461-473.
18. Kim N, Shiffeldrim N, Gan HH, Schlick T: **Candidates for novel RNA topologies**. *J Mol Biol* 2004, **341**:1129-1144.
19. Fera D, Kim N, Shiffeldrim N, Zorn J, Laserson U, Gan HH, Schlick T: **RAG: RNA-As-Graphs web resource**. *BMC Bioinformatics* 2004, **5**:88.
20. Gan HH, Fera D, Zorn J, Shiffeldrim N, Tang M, Laserson U, Kim N, Schlick T: **RAG: RNA-As-Graphs database--concepts, analysis, and features**. *Bioinformatics* 2004, **20**:1285-1291.
21. Gan HH, Pasquali S, Schlick T: **Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design**. *Nucleic Acids Res* 2003, **31**:2926-2943.
•22. Karklin Y, Meraz RF, Holbrook SR: **Classification of non-coding RNA using graph representations of secondary structure**. *Pac Symp Biocomput* 2005:4-15.
Labeled dual graphs are introduced to compare RNA 2D structures distinguish RNA molecules belonging to different families of homologues.

23. Krasilnikov AS, Xiao Y, Pan T, Mondragon A: **Basis for structural diversity in homologous RNAs**. *Science* 2004, **306**:104-107.
24. Krasilnikov AS, Yang X, Pan T, Mondragon A: **Crystal structure of the specificity domain of ribonuclease P**. *Nature* 2003, **421**:760-764.
25. Torres-Larios A, Swinger KK, Krasilnikov AS, Pan T, Mondragon A: **Crystal structure of the RNA component of bacterial ribonuclease P**. *Nature* 2005, **437**:584-587.
26. Westhof E, Massire C: **Structural biology. Evolution of RNA architecture**. *Science* 2004, **306**:62-63.
27. Nagaswamy U, Fox GE: **Frequent occurrence of the T-loop RNA folding motif in ribosomal RNAs**. *Rna* 2002, **8**:1112-1119.
28. Leontis NB, Stombaugh J, Westhof E: **Motif prediction in ribosomal RNAs Lessons and prospects for automated motif prediction in homologous RNA molecules**. *Biochimie* 2002, **84**:961-973.
29. Reijmers TH, Wehrens R, Buydens LM: **The influence of different structure representations on the clustering of an RNA nucleotides data set**. *J Chem Inf Comput Sci* 2001, **41**:1388-1394.
30. Olson WK: **Configuration statistics of polynucleotide chains**. *Macromolecules* 1980, **13**:721-728.

31. Leontis NB, Westhof E: **Geometric nomenclature and classification of RNA base pairs**. *Rna* 2001, **7**:499-512.
32. Gendron P, Lemieux S, Major F: **Quantitative analysis of nucleic acid three-dimensional structures**. *J Mol Biol* 2001, **308**:919-936.
••33. Lescoute A, Leontis NB, Massire C, Westhof E: **Recurrent structural RNA motifs, Isostericity Matrices and sequence alignments**. *Nucleic Acids Res* 2005, **33**:2395-2409.
Isostericity Matrices for non-Watson-Crick basepairs were used to analyze structures and sequence alignments of two recurrent RNA motifs, kink-turns and C-loops. Sequence signatures were derived for each motif to use to identify and align motifs in other RNA sequences.

••34. Huang HC, Nagaswamy U, Fox GE: **The application of cluster analysis in the intercomparison of loop structures in RNA**. *RNA* 2005, **11**:412-423.
Hairpin tetraloops in 3D RNA structures are clustered according to their geometric similarity using an RMSD measure calculated using fifteen backbone and base atoms per nucleotide. Major clusters included the GNRA and UNCG type hairpin loops.

35. Harrison AM, South DR, Willett P, Artymiuk PJ: **Representation, searching and discovery of patterns of bases in complex RNA structures**. *J Comput Aided Mol Des* 2003, **17**:537-549.
36. Artymiuk PJ, Spriggs RV, Willett P: **Graph Theoretic Methods for the Analysis of Structural Relationships in Biological Macromolecules**. *J. Am. Soc. Inf. Sci. Tech.* 2005, **56**:518-528.
••37. Schneider B, Moravek Z, Berman HM: **RNA conformational classes**. *Nucleic Acids Res* 2004, **32**:1666-1677.
Fourier averaging of the six 3D distributions of torsion angles followed by clustering identified 14 A-type (helical) and18 non-A-type RNA conformations and their torsion angles.

38. Hershkovitz E, Tannenbaum E, Howerton SB, Sheth A, Tannenbaum A, Williams LD: **Automated identification of RNA conformational motifs: theory and application to the HM LSU 23S rRNA**. *Nucleic Acids Res* 2003, **31**:6249-6257.
•39. Murray LJ, Richardson JS, Arendall WB, Richardson DC: **RNA backbone rotamers--finding your way in seven dimensions**. *Biochem Soc Trans* 2005, **33**:485-487.
Quality-filtering techniques are applied to RNA backbone dihedral angle distributions within sugar-to-sugar "suites." A small library of RNA backbone rotamers is identified which describes almost all RNA backbones in experimental structures.

40. Murray LJ, Arendall WB, 3rd, Richardson DC, Richardson JS: **RNA backbone is rotameric**. *Proc Natl Acad Sci U S A* 2003, **100**:13904-13909.
41. Sundaralingam M: **Stereochemistry of nucleic acids and their constituents**. *Biopolymers* 1969, **7**:821-860.
42. Sundaralingam M, Mizuno H, Stout CD, Rao ST, Liedman M, Yathindra N: **Mechanisms of chain folding in nucleic acids. The (omega, omega) plot and**

its correlation to the nucleotide geometry in yeast tRNAPhe1. *Nucleic Acids Res* 1976, **3**:2471-2484.
43. Duarte CM, Wadley LM, Pyle AM: **RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space**. *Nucleic Acids Res* 2003, **31**:4755-4761.
44. Szep S, Wang J, Moore PB: **The crystal structure of a 26-nucleotide RNA containing a hook-turn**. *Rna* 2003, **9**:44-51.
••45. Wadley LM, Pyle AM: **The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery**. *Nucleic Acids Res* 2004, **32**:6650-6659.
New software is described that identifies new recurrent backbone conformations in RNA structures without prior knowledge. Four new conformations of five or more nucleotides are defined.

46. Adams PL, Stahley MR, Gill ML, Kosek AB, Wang J, Strobel SA: **Crystal structure of a group I intron splicing intermediate**. *Rna* 2004, **10**:1867-1887.
47. Adams PL, Stahley MR, Kosek AB, Wang J, Strobel SA: **Crystal structure of a self-splicing group I intron with both exons**. *Nature* 2004, **430**:45-50.
•48. Strobel SA, Adams PL, Stahley MR, Wang J: **RNA kink turns to the left and to the right**. *Rna* 2004, **10**:1852-1854.
A new motif featuring a sharp bend toward the major (deep) groove of the RNA helix is described.

49. Klein DJ, Schmeing TM, Moore PB, Steitz TA: **The kink-turn: a new RNA secondary structure motif**. *Embo J* 2001, **20**:4214-4221.
50. Tamura M, Hendrix DK, Klosterman PS, Schimmelman NR, Brenner SE, Holbrook SR: **SCOR: Structural Classification of RNA, version 2.0**. *Nucleic Acids Res* 2004, **32**:D182-184.
••51. Klosterman PS, Hendrix DK, Tamura M, Holbrook SR, Brenner SE: **Three-dimensional motifs from the SCOR, structural classification of RNA database: extruded strands, base triples, tetraloops and U-turns**. *Nucleic Acids Res* 2004, **32**:2342-2352.
The SCOR database is a compilation and classification of RNA hairpin and internal loop motifs. Several new motifs are described in this article.

52. Leontis NB, Stombaugh J, Westhof E: **The non-Watson-Crick base pairs and their associated isostericity matrices**. *Nucleic Acids Res* 2002, **30**:3497-3531.
53. Lemieux S, Major F: **RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire**. *Nucleic Acids Res* 2002, **30**:4250-4263.
54. Yang H, Jossinet F, Leontis N, Chen L, Westbrook J, Berman H, Westhof E: **Tools for the automatic identification and classification of RNA base pairs**. *Nucleic Acids Res* 2003, **31**:3450-3460.
55. Jossinet F, Westhof E: **Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure**. *Bioinformatics* 2005, **21**:3320-3321.

56. Lee JC, Gutell RR: **Diversity of base-pair conformations and their occurrence in rRNA structure and RNA structural motifs**. *J Mol Biol* 2004, **344**:1225-1249.
57. Nagaswamy U, Larios-Sanz M, Hury J, Collins S, Zhang Z, Zhao Q, Fox GE: **NCIR: a database of non-canonical interactions in known RNA structures**. *Nucleic Acids Res* 2002, **30**:395-397.
58. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences**. *Nucleic Acids Res* 1990, **18**:6097-6100.
59. Gorodkin J, Heyer LJ, Brunak S, Stormo GD: **Displaying the information contents of structural RNA alignments: the structure logos**. *Comput Appl Biosci* 1997, **13**:583-586.
60. Leontis NB, Altman R, Berman HM, Brenner SE, Brown J, Engelke D, Harvey SC, Holbrook SR, Jossinet F, Lewis SE, et al.: **The RNA Ontology Consortium: An Open Invitation to the RNA Community**. *RNA* In Press.

Figure Legends

Figure 1: Five examples of the new Ω-turn motif [45] proposed by Wadley and Pyle (2004) in their structural contexts. **Upper panel**: Three-dimensional representations highlighting the conservation of the backbone conformation of the five nucleotides comprising each Ω-turn (shown in red). **Lower panel**: Schematic representations of each Ω-turn in its structural context, annotated with symbols for basepairing [31] and base-stacking [46] interactions. The nucleotides comprising each Ω-turn are shown in red and nucleotides in the *syn* conformation are indicated with **boldface**.

Figure 2: K-turn and reverse K-turn. **Left panel**: Schematic structures of the Helix 7 K-turn of 23S rRNA of *Haloarcula marismortui* (*H.m.*) (shown in red) and the reverse K-turn in the P9/P9.0 junction of the *Azoarcus* intron (shown in blue) annotated for basepairing interactions with the geometric nomenclature of Leontis and Westhof (2001) [31]. **Right panel**: Superposition of the canonical helices (C-helix) from the 3D structures the K-turn (red) and the reverse K-turn (blue); in the K-turn structure, the non-canonical helix (NC helix) is oriented in the minor/shallow groove while for the reverse K-turn structure, NC-helix is in the major/deep groove.

Figure 3: Flow chart illustrating the use of Isostericity Matrices to integrate 3D structural and sequence information so as to produce accurate alignments and model 3D structures based on sequence. Isostericity Matrices (IM) for non-Watson-Crick basepairs organized in geometric families were proposed based on analysis of high-resolution atomic structures [52], as indicated in path **1**. Sequence signatures of RNA motifs identified in 3D structures are deduced by analyzing homologous RNA molecules having the same motif (path **2**). Isostericity Matrices are employed to productively iterate between sequence alignment and sequence signature to arrive at accurate, structure-based alignments (path **2**). Sequence signatures for recurrent motifs identified in different crystal structures are defined with reference to Isostericity Matrices (path **3**). For families of homologous RNA molecules for which no 3D structure exists (path **4**), Watson-Crick covariations and energy minimization (path **5**) can be used to determine common 2D structures, which in turn define hairpin, internal, and junction loops where 3D motifs may occur. Sequence signatures of known motifs are used to propose motifs for loops and used to refine alignments of loop regions in an iterative manner (paths **4** and **5**). Motif substitutions at corresponding positions in the alignments can also be identified (path **4**).

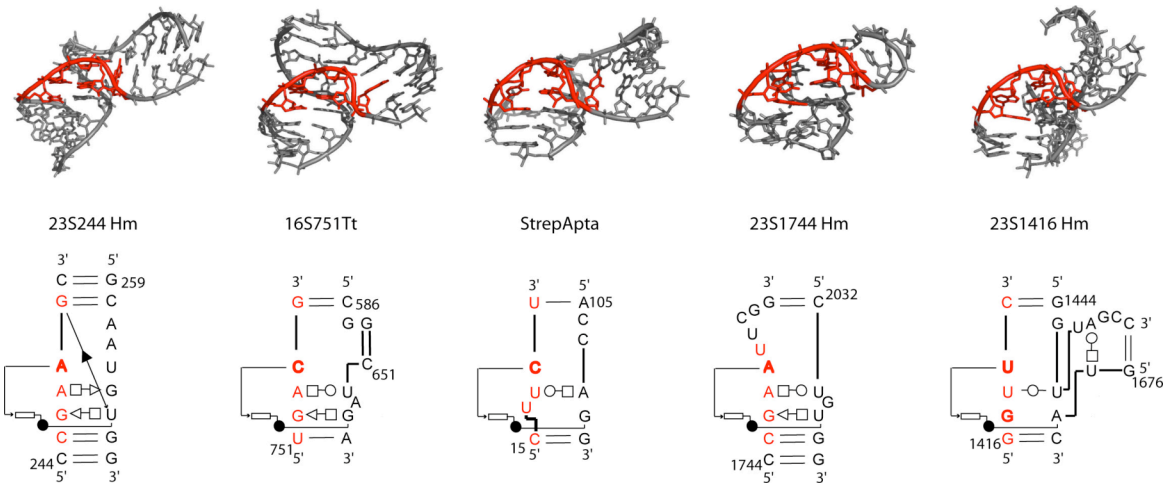


Figure 1

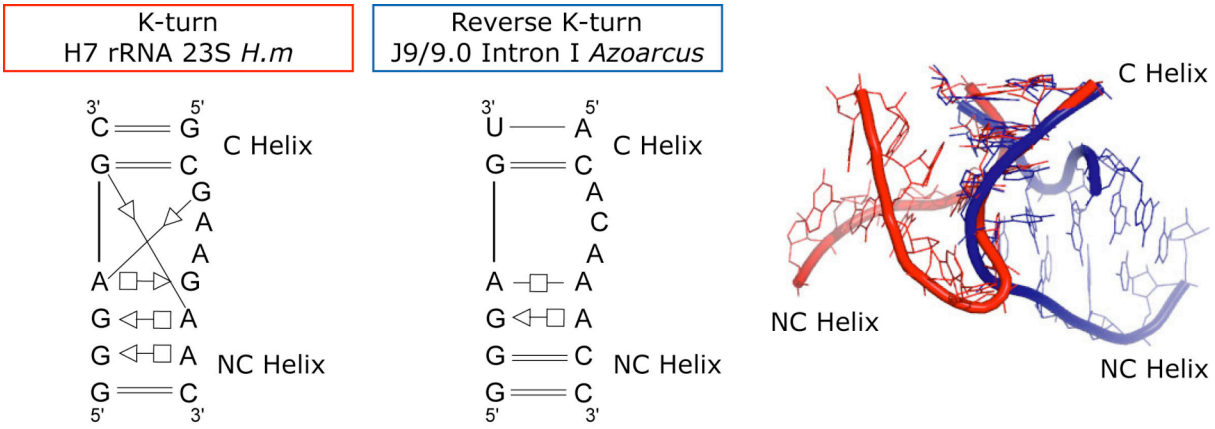

K-turn
H7 rRNA 23S *H.m*
Reverse K-turn
J9/9.0 Intron I *Azoarcus*
3'
5'
C
G
C Helix
G
C
G
A
A
A
G
G
A
G
A
NC Helix
G
C
5'
3'
3'
5'
U
A
C Helix
G
C
A
C
A
A
A
G
A
G
C
NC Helix
G
C
5'
3'
C Helix
NC Helix
NC Helix

Figure 2

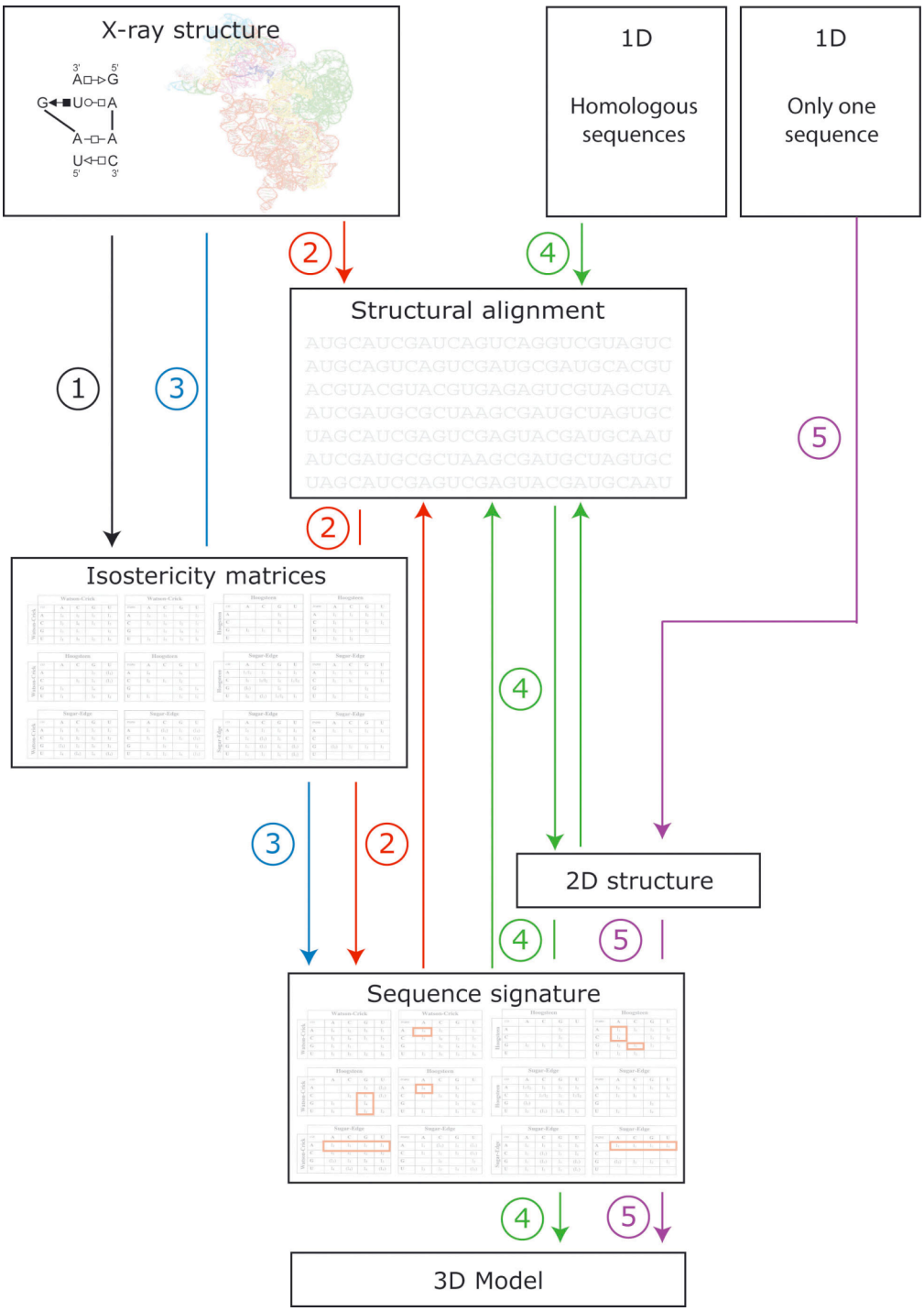

X-ray structure
1D
Homologous sequences
1D
Only one sequence
Structural alignment
Isostericity matrices
2D structure
Sequence signature
3D Model
1
2
3
4
5

Figure 3

2.3. Modélisation des éléments périphériques du ribozyme à tête de marteau

2.3.1. Introduction

REACTION CATALYTIQUE

Le ribozyme à tête de marteau est un petit ARN autocatalytique qui, en présence d'ions divalents, clive en un site spécifique le squelette sucre-phosphate du brin substrat par une réaction de transestérification libérant une extrémité 2'-3' phosphate cyclique et une extrémité 5'-hydroxyle (Figure 34).

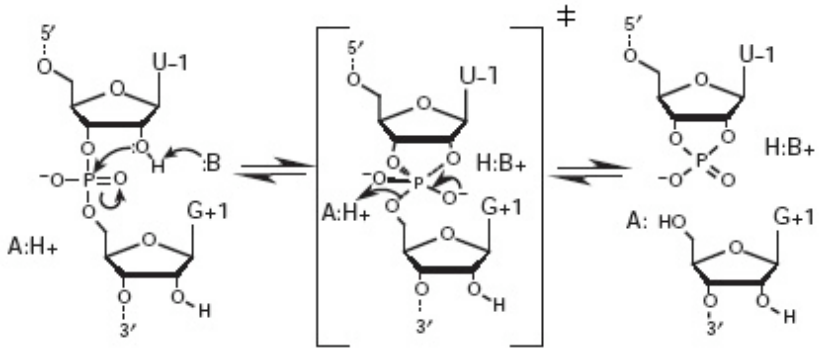


Figure 34: Réaction d'auto-clivage du ribozyme à tête de marteau. Le 2' hydroxyle adjacent au phosphate cyclique est activé par attaque nucléophile sur son proton. De manière concomitante, un proton est donné pour stabiliser l'oxygène partant. Figure extraite de la référence (Doudna & Cech, 2002)

FONCTION BIOLOGIQUE

Cet ARN autocatalytique a été d'abord découvert chez certains viroïdes et virusoïdes responsables de plusieurs maladies infectieuses chez des plantes économiquement importantes. Ces pathogènes sont de petits ARN simple brin circulaires, non encapsidés, et ne possèdent aucun cadre de lecture ce qui les classe dans les ARN non-codants. Leur réplication a lieu (i) dans le cytoplasme à l'aide d'un virus assistant dans le cas des virusoïdes, (ii) de manière autonome (sans virus assistant) dans les chloroplastes ou le noyau dans le cas des viroïdes (Figure 35). La réplication de ces génomes circulaires se fait selon le mécanisme de "rolling circle" ou "cercle roulant", qui implique une copie du génome circulaire

matrice de polarité positive (brin +) par une RNA polymérase de l'hôte ou du virus assistant pour donner un brin d'ARN de polarité négative (brin -) long de plus d'une unité. Dans le cas des viroïdes, le devenir des concatémères synthétisés diffère selon qu'ils appartiennent à la famille des Pospiviroidae ou des Avsunviroidae (Figure 36). Dans le premier cas, le brin (–) multicopie linéaire sert de matrice pour la synthèse d'un brin (+) multicopie dont la circularisation d'une unité semble demander l'intervention d'une ARN ligase comme l'indique la présence d'un 2' phosphomonoester et d'une liaison phosphodiester 3',5'-phosphodiester observés dans certains cas (Kiberstis et al., 1985).

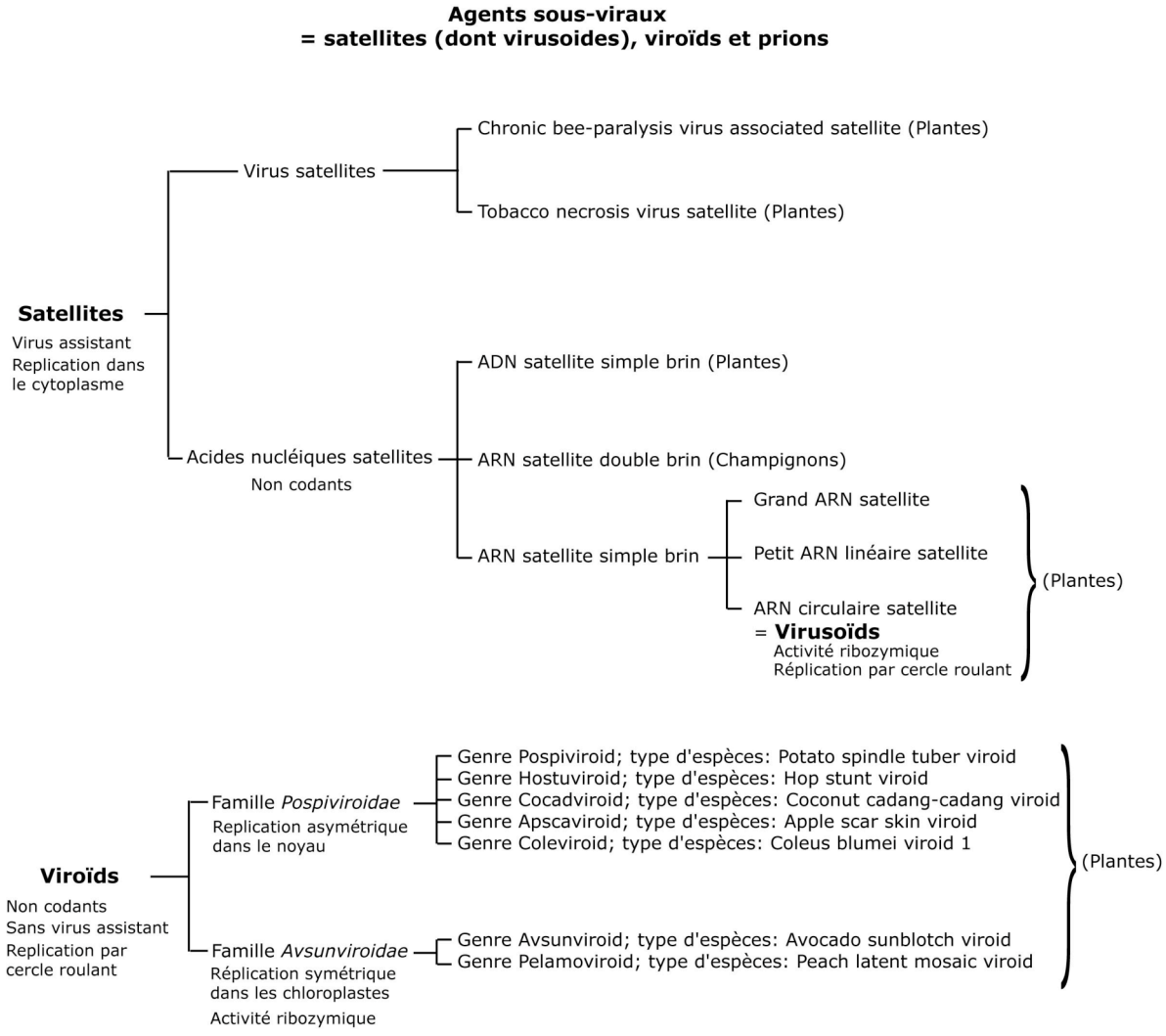


Figure 35 : Classification des satellites et viroïdes. Les viroïdes prolifèrent et se copient de manière autonome (sans virus assistant). Leur génome ARN circulaire ne possède pas de cadre ouvert de lecture. Les virusoïdes sont des ARN satellites à génome simple brin circulaire. Ils ne codent pas de protéine contrairement aux génomes de virus satellites qui codent pour des protéines de la capside. Les hôtes infectés par ces pathogènes sont indiqués entre parenthèses.

Pour les viroïdes qui appartiennent à la famille des Avsunviroidae, le brin (-) multicopie linéaire se clive lui-même grâce au ribozyme à tête de marteau présent. Le mécanisme de circularisation de ces brins (-) unicopie pour servir de matrice reste polémique. En effet deux hypothèses sont avancées : l'une proposant l'intervention d'une ARN ligase et l'autre que le ribozyme en catalysant la réaction inverse de la coupure est capable de lier l'ARN. Ceci sera développé plus longuement dans la partie discussion de ce chapitre.

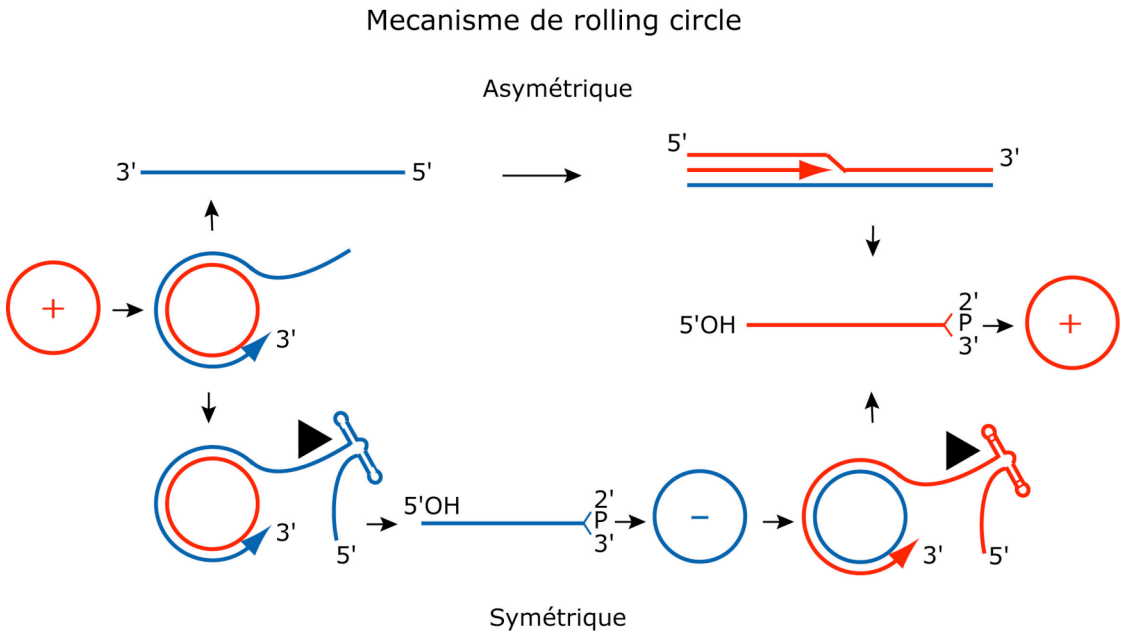


Figure 36 : Mécanisme de réplication par cercle roulant ("rolling circle"). Il existe deux voies de réplication : une voie asymétrique (en haut) qui nécessite un seul "rolling circle" empruntée par la famille des *Pospiviroidae* et une voie symétrique qui nécessite deux "rolling circle" spécifique à la famille des *Avsunviroidae*. Les brins de polarité + sont dessinés en rouge et les brins de polarité négative en bleu. Seule la voie symétrique fait intervenir les ribozymes à tête de marteau dans le clivage des concatémères.

Finalement, le brin (-) sert à son tour de matrice pour la synthèse de concatémères (+) synthétisés par le même mécanisme de "rolling circle" que précédemment. Après clivage des copies par le ribozyme à tête de marteau, les brins (+) sont circularisés à leur tour. Ainsi, la réplication n'implique que des intermédiaires ARN puisque le génome du viroïde ne code pour aucune protéine. Donc les différentes propriétés biologiques des viroïdes, comme l'identification de l'hôte, dépendent exclusivement de la séquence et de la structure de leur ARN.

Le ribozyme à tête de marteau joue un rôle fondamental dans la réplication du génome de ces pathogènes de plantes. Son rôle, pourtant, ne semble pas se limiter à cette fonction puisqu'il a été identifié par la suite dans le génome du triton (*Notophtalamus viridescens*) (Pabon-Pena et al., 1991; Zhang & Epstein, 1996), des schistosomes (Ferbeyre et al., 1998), et du cricket des cavernes (espèces Dolichopoda) (Rojas et al., 2000). Plus récemment, deux ribozymes à tête de marteau ont également été trouvés dans le génome d'*Arabidopsis thaliana* ; c'est la première fois que ce ribozyme est trouvé dans un génome de plante (voir article 2).

STRUCTURE ET REPLIEMENT

La structure du cœur catalytique du ribozyme à tête de marteau a été déterminée par cristallographie sous la forme d'un hybride ARN/ADN et sous la forme "tout ARN" avec un méthyle en 2' du site de clivage empêchant ainsi la réaction (Pley et al., 1994b; Scott et al., 1995). Les deux structures pratiquement identiques montrent un cœur catalytique composé de onze nucléotides essentiels à la catalyse, organisés en point de jonction duquel partent trois hélices I, II et III (Figure 37). L'hélice II est empilée sur l'hélice III tandis que l'hélice I est parallèle à l'hélice II. L'hélice II, composée de paires de bases Watson-Crick, repose sur trois paires non Watson-Crick : deux paires *trans* Hoogsteen/Sucre (A9G12 et A13G8) empilées sur une paire *cis* Watson-Crick/Sucre (A15.1U16.1).

De nombreuses études structurales du ribozyme en solution montrent l'importance des ions Mg^{2+} pour son repliement. En effet, en absence d'ions divalent le cœur du ribozyme n'est pas structuré et les hélices sont à équidistance les unes des autres révélant une désorganisation de l'empilement, alors qu'en présence d'ions divalent le ribozyme se replie en deux étapes en une structure active. A faible concentration, les hélices II et III sont empilées mais l'angle entre les hélices I et II est très grand et le cœur catalytique n'est pas structuré ; le ribozyme est inactif. A plus forte concentration, l'hélice I se réoriente pour se positionner à proximité de l'hélice II ; le ribozyme est alors dans une conformation active (Bassi et al., 1995; Bassi et al., 1997; Hammann et al., 2001)

La plupart des études de repliement du ribozyme à tête de marteau ont été réalisées sur des ribozymes minimalistes présentant les éléments conservés et supposés essentiels à l'activité catalytique : les treize nucléotides du cœur catalytique dont onze ne peuvent pas être mutés sans causer de perte d'activité et les trois hélices I, II et III de longueur et de terminaison variables. Ces ribozymes minimalistes sont actifs de manière optimale à forte concentration d'ions Mg^{2+} (>10mM) mais sont quasiment inactifs à concentration physiologique de Mg^{2+} (0,1-0,3 mM). Au contraire, les ribozymes dont la séquence naturelle a été conservée, sont fonctionnels aux concentrations physiologiques de Mg^{2+}. L'objectif de l'étude menée en collaboration avec A. Khvorova et S. Jayasena (Article 2) a été de déterminer quelles étaient les caractéristiques des séquences et/ou les structures responsables de l'efficacité de coupure du ribozyme naturel à une concentration de Mg^{2+} cent fois inférieure à la concentration nécessaire à l'activité d'un ribozyme minimaliste.

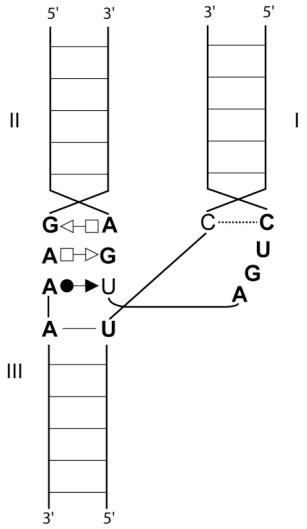


Figure 37 : Le cœur catalytique du ribozyme à tête de marteau. Les hélices I, II et III sont indiquées. Le cœur est composé de treize nucléotides dont onze sont strictement conservés (en gras). Les appariements sont représentés selon la nomenclature de Leontis et Westhof (2001).

CLASSIFICATION DES RIBOZYMES A TETE DE MARTEAU

Les ribozymes naturels peuvent être classés en trois types selon l'hélice qui porte l'extrémité 5' d'entrée dans le ribozyme (Figure 38). Trois types de ribozyme à tête de marteau sont ainsi définis :

Type 1 : l'extrémité 5' d'entrée dans la séquence du ribozyme est portée par l'hélice I qui présente toujours une bulle interne. L'hélice II présente parfois une bulle interne et une boucle terminale. La boucle terminale de l'hélice III est très

courte, inférieure ou égale à deux nucléotides. Une boucle si courte fermée par une paire Watson-Crick canonique n'est jamais observée dans une structure d'ARN cristallographique. En fait, des études montrent que certains ribozymes à tête de marteau de type 1 sont actifs sous forme dimérique (Forster et al., 1988). Le dimère présente deux cœurs catalytiques fonctionnels présentant chacun une part égale de chacune des deux molécules de ribozyme. L'hélice III ainsi formée de deux brins appartenant à deux molécules de ribozyme différentes ne présente plus de boucle terminale mais éventuellement une bulle interne. Nous reviendrons dans la discussion sur les phénomènes de dimérisation des ribozymes à tête de marteau.

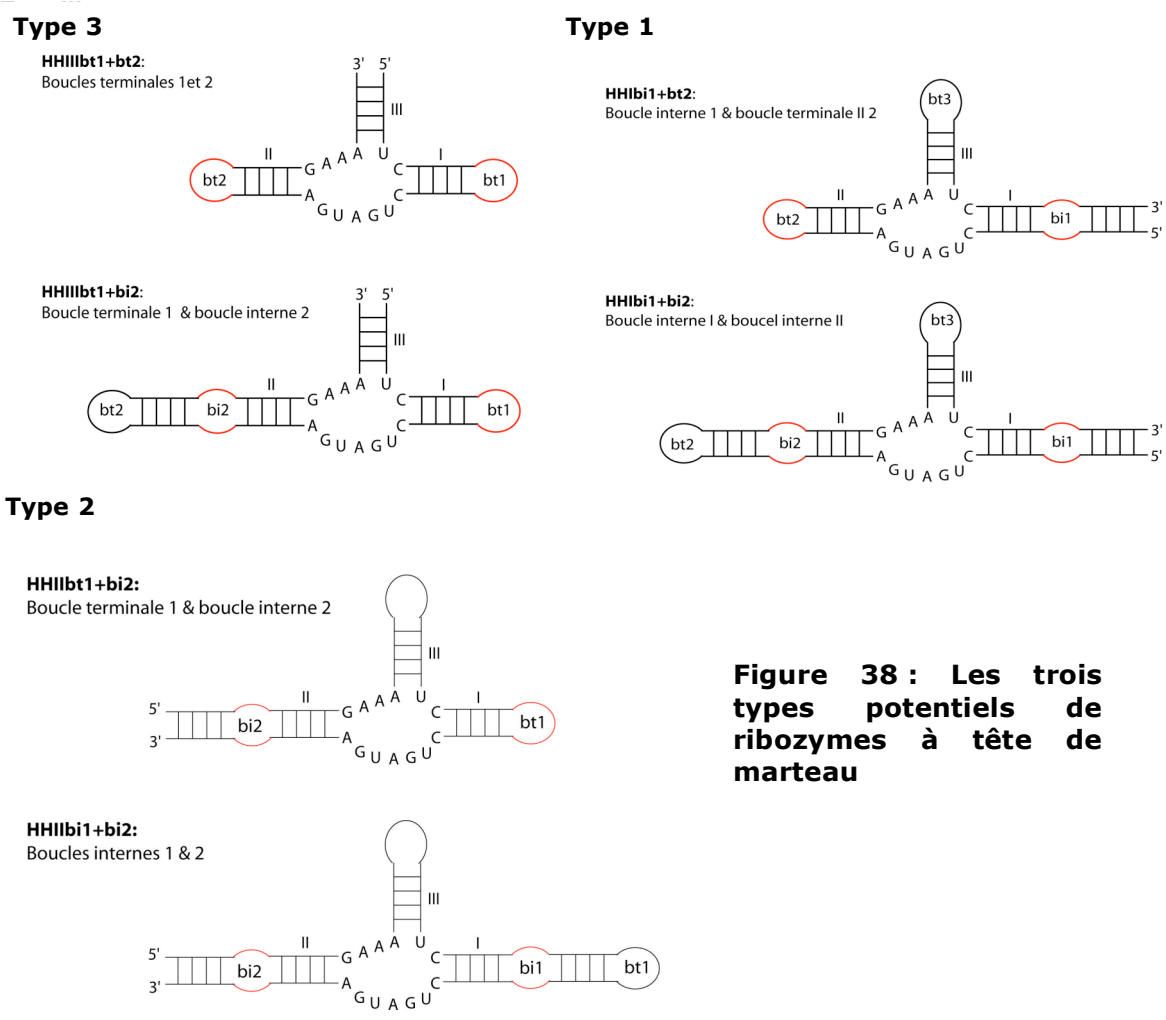


Figure 38 : Les trois types potentiels de ribozymes à tête de marteau

Type 2 : l'extrémité 5' d'entrée dans la séquence du ribozyme est portée par l'hélice II qui présente toujours une bulle interne. L'hélice II présente parfois une bulle interne et évidemment une boucle terminale. L'hélice III ne présente pas de bulle interne mais une boucle terminale. Les ribozymes de type 2 n'ont jamais été mis en évidence (une hypothèse sur les raisons de cette absence sera proposée dans la discussion).

Type 3 : l'extrémité 5' d'entrée dans la séquence du ribozyme est portée par l'hélice III. Les hélices I et II présentent parfois une bulle interne et évidemment une boucle terminale.

2.3.2. Article 2: "Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity"

Anastasia Khvorova, Aurélie Lescoute, Eric Westhof & Sumedha D Jayasena

Nature Structural Biology, 2003 Sep;10(9):708-12

Les ribozymes naturels étudiés dans l'article 1 (sTRSV, vLTSV et PLMVD) appartiennent aux ribozymes de type 3. Ils sont actifs à des concentrations très faibles de Mg^{2+} contrairement aux ribozymes minimalistes HH2 et HH15 précédemment utilisés. La différence majeure entre ces ribozymes naturels et minimalistes est l'absence de boucle terminale aux extrémités des hélices I et II de ces derniers. La mutation systématique des nucléotides des boucles L1 et L2 de sTRSV a permis de mettre en évidence l'importance de la nature des bases de ces boucles pour l'activité ribozymique. Ajoutée à ces expériences de mutations, la proximité des hélices I et II dans la structure cristallographique nous a fait suspecter des interactions entre leur boucle terminale. Les hélices I et II étant parallèles, il n'était pas possible qu'elles forment un "kissing complex" formé de paires *cis* Watson-Crick. Nous avons observé que tous les "kissing complex", dont la structure a été déterminée, montrent un angle supérieur à 90° entre les hélices impliquées. Par contre, des similarités ont été trouvées avec les boucles 6 et 8 de l'ARN de SRP (Signal Recognition Particle). Nous avons utilisé cette structure ainsi que la structure cristallographique du cœur catalytique pour modéliser les interactions tridimensionnelles entre les boucles 1 et 2. Les résultats des mesures d'activité *in vitro* et *in vivo* de ribozymes chimères basés sur sTRSV, PLMVD et vLTSV vérifient l'existence des contacts interboucles. En résumé, les résultats de l'article 1 mettent en évidence que les interactions tertiaires entre les éléments périphériques non essentiels pour la catalyse sont responsables d'une excellente activité des ribozymes naturels et des ribozymes artificiels en conditions physiologiques. Les ions Mg^{2+} à forte concentration jouent le rôle de stabilisateur conformationnel. A faible concentration d'ions divalents, les interactions tertiaires interboucles compensent la perte d'ions

divalents et réduisent l'espace conformationnel occupé par les éléments structuraux ; la structure est ainsi contrainte dans un état actif.

Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity

Anastasia Khvorova**, Aurélie Lescoute,** Eric Westhof & Sumedha D Jayasena

Nature Structural Biology, 2003, Volume 10, N° 9, Pages 708 - 712

2.3.3. Article 3 : "Functional hammerhead ribozymes naturally encoded in the genome of Arabidopsis thaliana"

Rita Przybilski, Stefan Graf, Aurélie Lescoute, Wolfgang Nellen, Eric Westhof, Gerhard Steger & Christian Hammann

Plant Cell. 2005 Jul;17(7):1877-85.

Le ribozyme trouvé dans le génome d'Arabidopsis thaliana (Article 2) appartient lui aussi aux ribozymes de type 3. Comme précédemment, nous avons proposé un modèle de la structure de ce ribozyme présentant des interactions tertiaires interboucles.

2.3.4. Discussion

En résumé, ces résultats démontrent que les contacts tertiaires entre les boucles terminales des hélices I et II, périphériques au cœur catalytique et non essentielles à la catalyse, sont indispensables à l'activité des ribozymes à tête de marteau en conditions physiologiques (faible teneur en Mg^{2+}). Les études de De la Pena et collaborateurs, réalisées en parallèle sur les ribozymes à tête de marteau naturels de PLMVd (+) et CChMVd (+), corroborent nos conclusions (De La Pena et al., 2003). Ils ont montré, en effet, que ces ribozymes, agissant en cis, sont plus efficaces que leurs dérivés dépourvus d'appendices périphériques et actifs en trans. Le rôle des structures périphériques et de leurs interactions dans l'activité catalytique des ribozymes à tête de marteau n'est pas si surprenant car il a été mis en évidence précédemment chez d'autres ribozymes de plus grande taille comme les introns de groupe I par exemple. La délétion progressive des appendices périphériques des introns de groupe I mène à des besoins croissants en sels et ultimement à l'inactivité (Doudna & Szostak, 1989b; Beaudry & Joyce, 1990). Les différentes catégories d'intron de groupe I possèdent tous le même cœur catalytique, mais les régions périphériques et les interactions tertiaires diffèrent (Costa & Michel, 1995; Jaeger, 1995; Lehnert et al., 1996). Toutes, cependant, sont responsables de la stabilité du cœur catalytique. Au cours de l'évolution, plusieurs solutions ont été trouvées pour assurer l'activité catalytique optimale du ribozyme. De la même manière, les appendices périphériques des ribozymes naturels à tête de marteau ont évolué de manière à optimiser l'activité catalytique pour les besoins biologiques aux conditions natives en Mg^{2+}. Les interactions tertiaires entre ces appendices, réduisent l'espace conformationnel accessible à la structure et facilitent le repliement du ribozyme augmentant ainsi son activité catalytique. Il est clair que le cœur catalytique du ribozyme doit subir des changements structuraux, probablement infimes, qui expliquent le gain d'efficacité. Des études cinétiques et structurales seront nécessaires pour comprendre l'influence des interactions tertiaires sur ces processus.

RÔLE DES ELEMENTS PERIPHERIQUES DANS LE MECANISME DE CLIVAGE

L'implication des éléments périphériques dans l'activité catalytique du ribozyme à tête de marteau en conditions physiologiques n'est pas réservée aux ARN des viroïdes ou aux ARN satellites. En effet, des études cinétiques montrent que l'activité de coupure et l'orientation des hélices I, II et III pendant le processus de repliement du ribozyme naturel de *Schistosoma mansoni* ont lieu à des concentrations de Mg^{2+} de l'ordre du µM et sont dépendantes de l'intégrité des interactions boucle-boucle, alors que les ribozymes minimalistes nécessitent une concentration en ions Mg^{2+} de l'ordre du millimolaire (Canny et al., 2004; Penedo et al., 2004). Une expérience de pontage UV sur un dérivé trans du ribozyme de *S.mansoni* présentant les boucles périphériques du ribozyme naturel, montre que deux uridines, l'une appartenant à la boucle 1 l'autre à la boucle 2, sont empilées (Heckman et al., 2005). Ces études cinétiques ainsi que l'interaction d'empilement détectée vont dans le sens d'interactions tertiaires entre les régions périphériques du ribozyme de *S.mansoni*. La contrainte imposée par ces interactions sur la structure globale du ribozyme réduirait les possibilités de mouvement dans l'espace de ses éléments structuraux. Les interactions tertiaires piégeraient ainsi le ribozyme dans un état proche de l'état de transition ce qui expliquerait le gain d'efficacité observé. Cette proposition soulève un certain nombre de questions : les changements conformationnels nécessaires au passage d'un état inactif à un état actif concernent-ils uniquement les éléments périphériques ou également le cœur catalytique ? S'agit-il de changements minimes ou drastiques ? Et finalement, la structure cristallographique dont nous disposons représente-t-elle le ribozyme dans son état actif ou dans l'état qui le précède ? Blount et Uhlenbeck insistent sur le nombre important d'incohérences entre les structures cristallographiques disponibles et les données biochimiques sur les groupes fonctionnels importants du ribozyme ; ils proposent que cela soit révélateur de l'isomérisation conformationnelle subie par le ribozyme antérieurement à l'état de transition (Blount & Uhlenbeck, 2005). Il est à noter que ces résultats expérimentaux ont été obtenus sur des structures de ribozymes minimalistes (avec toutes les incertitudes qu'une telle approche comporte). L'isomérisation conformationnelle implique le rapprochement significatif des hélices I et II qui pourrait être responsable d'un changement drastique au sein même du cœur catalytique et

positionner le phosphate scissile plus proche d'un arrangement en ligne. Ainsi, la structure cristallographique ne correspondrait pas à la structure active du ribozyme mais à l'état précédant l'état de transition nécessaire à la catalyse (Figure 39). Comme le montre la Figure 39, le passage de l'état inactif, qui correspondrait à la structure cristallographique, à l'état de transition nécessiterait non seulement un rapprochement très important des hélices mais également une restructuration importante du cœur catalytique même. Cette hypothèse expliquerait que les atomes impliqués dans la réaction de coupure montrent un alignement dans la structure cristallographique en désaccord avec l'alignement que nécessite une réaction SN_2. Une réorganisation structurale du cœur catalytique permettrait alors un alignement de ces atomes et le passage à l'état de transition. Il est vrai que cette hypothèse est tentante, pourtant un certain nombre de données vont à l'encontre de ces arguments. D'abord, des expériences montrent que les ribozymes contenus dans le cristal sont capables de réaliser la réaction de catalyse (Dunham et al., 2003). D'autre part, nos résultats montrent que les interactions tertiaires proposées entre les boucles 1 et 2, ont pu être modélisées sans modifier la structure du cœur catalytique du ribozyme qui a été déterminée par cristallographie.

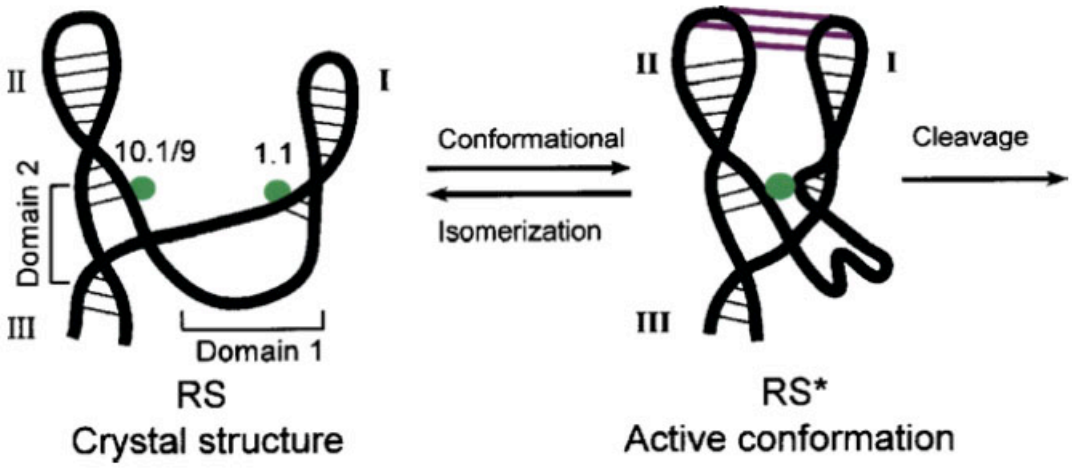


Figure 39 : Modèle du changement conformationnel nécessaire à la catalyse du ribozyme à tête de marteau Les interactions tertiaires favoriseraient le repliement du ribozyme en une conformation active. Image extraite de la référence (Blount & Uhlenbeck, 2005).

D'autre part, en accord avec les données biochimiques obtenues avec des ribozymes minimum, Lilley et collaborateurs ont proposé un modèle de repliement en deux étapes du ribozyme à tête de marteau (Figure 40) (Bassi et al., 1995; Bassi et al., 1997). La première étape qui a lieu à faible concentration de Mg^{2+} aboutit à la formation du domaine 1 c'est-à-dire des deux paires GoA consécutives et à l'empilement des hélices II et III. La deuxième étape, qui nécessite une forte concentration d'ions Mg^{2+}, mène à la structuration du domaine 2 c'est-à-dire à l'organisation de tous les nucléotides du cœur catalytique et au positionnement de l'hélice I à proximité de l'hélice II. Cette deuxième étape conduit à une structure du ribozyme équivalente à la structure cristalline qui correspond à un état très voisin de l'état actif. Au vu de nos résultats, nous proposions que les interactions tertiaires entre les boucles 1 et 2 stabilisent le positionnement de l'hélice I lors de la deuxième étape, nécessitant alors beaucoup moins de Mg^{2+} qu'en absence des éléments périphériques. Les études par FRET du repliement du ribozyme naturel de *Schistosoma mansoni* menées par Penedo et collaborateurs corroborent nos suppositions (Penedo et al., 2004). En effet, elles montrent que le repliement du ribozyme naturel en une structure active nécessite une seule étape en présence d'une concentration en Mg^{2+} de l'ordre du µM tandis que la perturbation des interactions interboucles entraîne un repliement en deux étapes nécessitant des concentrations de l'ordre du mM comme décrit précédemment. Les contacts tertiaires entre les boucles apicales propageraient dans la structure de l'hélice I un état torsionnel dont l'energie provoquerait la coupure. Ensuite, l'hélice I, tel un ressort détendu, réaliserait un mouvement de rotation en s'éloignant légèrement de l'hélice II. Hertel et Uhlenbeck ont proposé que la coupure du substrat augmente l'espace conformationnel accessible au ribozyme à tête de marteau, le rendant plus « lâche », faisant que le complexe enzyme/produit est plus favorable thermodynamiquement que le complexe enzyme/substrat (Hertel & Uhlenbeck, 1995).

De la même manière, Nahas et collaborateurs montrent que les hélices du ribozyme en épingle à cheveux sont empilées et repliées lorsqu'il est dans un état ligué mais sont non repliées après coupure (Nahas et al., 2004). Le rôle de la jonction à quatre hélices est déterminant dans ce mécanisme. Les ribozymes simplifiés, dont la jonction à quatre hélices est remplacée par une jonction à trois hélices ou un coude, ont une activité catalytique moins importante que les

ribozymes naturels (Fedor, 1999; Zhao et al., 2000). Le système du « ressort » semble cohérent avec ces observations (le ribozyme en épingle à cheveux présente également des interactions tertiaires favorisant la catalyse).

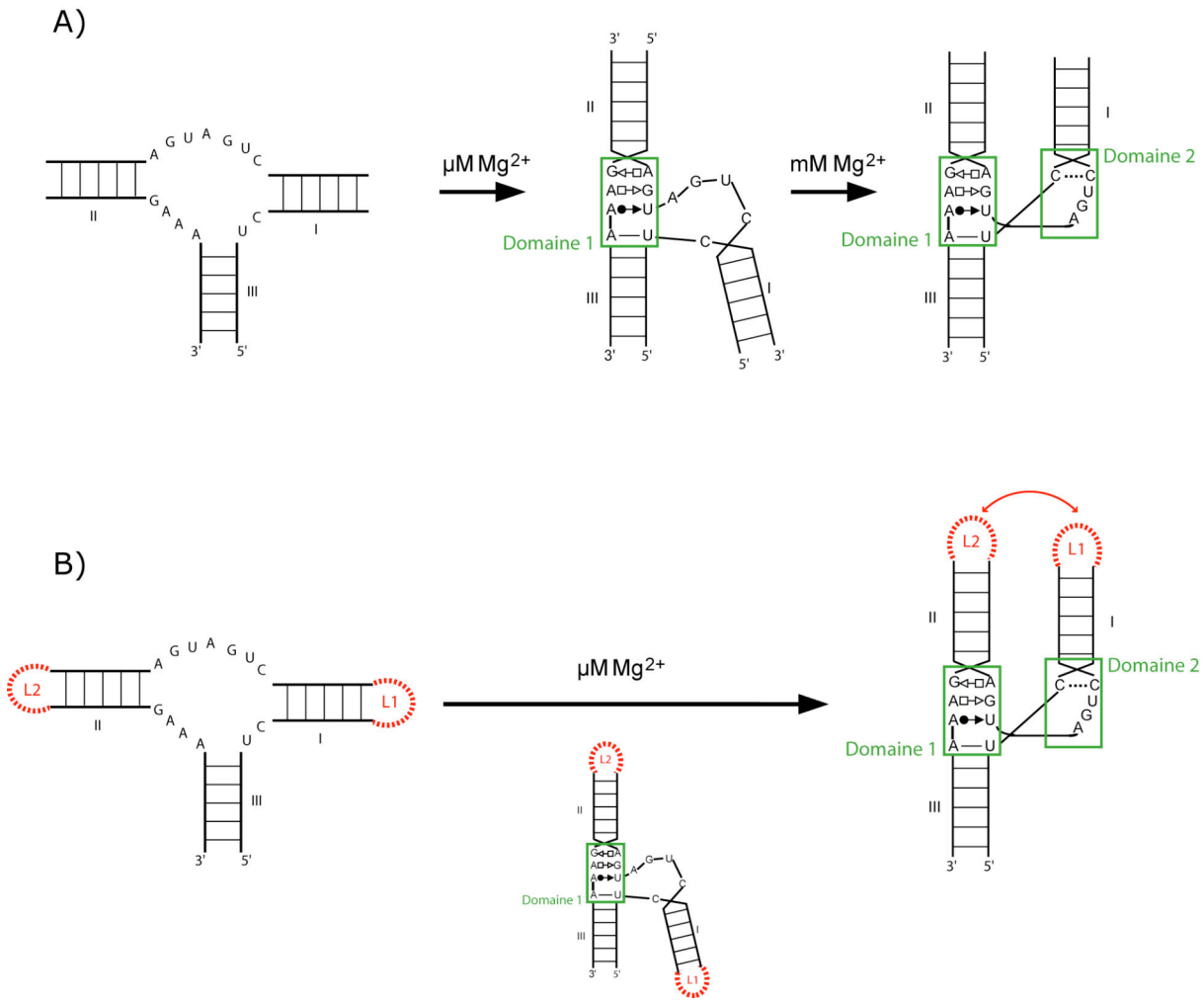


Figure 40 : Mécanisme de repliement du ribozyme à tête de marteau. (A) Le ribozyme minimum se replie en deux étapes. : d'abord formation du domaine 1 à faible concentration de Mg^{2+} puis, dans une deuxième étape à concentration forte de Mg^{2+} formation du domaine 2 et pivotement de l'hélice I à proximité de l'hélice II. (B) En présence des éléments périphériques, le repliement du ribozyme naturel se déroule en une étape, à une concentration faible de Mg^{2+}, et mène directement à l'interaction des boucles terminales des hélices I et II.

RÔLE DES ELEMENTS PERIPHERIQUES DANS LE MECANISME DE LIGATURE

Comme tout enzyme, les ribozymes à tête de marteau ou en épingle à cheveux, sont capables de réaliser la réaction inverse de la coupure c'est-à-dire la ligation. Toutefois, dans la majorité des conditions, le clivage est favorisé au moins cent fois par rapport à la ligature (Hertel & Uhlenbeck, 1995). Les

différentes étapes de la ligature sont : 1) activation du 5'-hydroxyle par une base ; 2) attaque du 2'-3' phosphate cyclique ; 3) libération et protonation de l'hydroxyle en 2'. La liaison 3'-5' phosphodiester serait ainsi restaurée.

Ribozyme en épingle à cheveux

Les études de la dynamique du ribozyme en épingle à cheveux ont permis de déterminer les propriétés cinétiques conformationnelles des réactions de ligature et de clivage. Elles montrent qu'après coupure la jonction à quatre hélices du ribozyme passe dans un état non empilé (Nahas et al., 2004). La réaction de coupure est nettement moins rapide que la réaction de ligature, mais la ligature est moins rapide que le désempilement des hélices. Le produit est donc relargué avant que la ligature n'ait lieu. Mais, dans le cas où le produit de coupure reste lié à l'enzyme, le ribozyme alterne d'un état empilé à un état non empilé avant que la ligature n'ait lieu. Ces expériences montrent aussi que la nature du produit est un facteur influant sur la vitesse de dépliement du ribozyme. En effet, la présence du phosphate 2'-3' cyclique augmente le dépliement du ribozyme comparé à un produit possédant un 3' phosphate ou hydroxyle. Donc, des altérations du produit ont des effets importants sur la dynamique structurale du ribozyme et la perte de continuité du squelette sucre/phosphate n'est pas le seul facteur augmentant le déploiement du ribozyme. Le déploiement ainsi favorisé par la présence du phosphate cyclique, entraîne un détachement du produit. Ceci est en accord avec la biologie du ribozyme dans son environnement naturel où il coupe l'ARN multicopie, produit de la réplication par cercle roulant, et relargue un produit monomérique 2'-3'phosphate cyclique. Dans le cas du ribozyme en épingle à cheveux du brin (-) de l'ARN satellite « tobacco ringspot viral satellite », l'hélice formée par le produit et l'enzyme, longue de 6 paires de bases, a une constante de dissociation assez faible. La spectroscopie montre que le dépliement-repliement peut avoir lieu plusieurs fois avant que la ligature n'ait lieu. La constante d'équilibre interne entre clivage et ligature montre un biais significatif vers la ligature. Cette favorisation de la ligature assure, lors de la réplication du virus, le maintien d'un certain nombre d'ARN circulaires qui serviront de matrice pour la synthèse du brin multicopie.

Ribozyme à tête de marteau

Pour le ribozyme à tête de marteau minimum, Hertel et Uhlenbeck ont montré que la constante cinétique de ligature (k_{-2} ≈ 0,008 min^{-1}) était 100 fois inférieure à la constante de coupure (k_2 ≈ 1 min^{-1}) ; la constante à l'équilibre Keq ≈ k_2/k_{-2} a donc une valeur de 125 (Figure 41) (Hertel et al., 1994). Ceci s'explique, comme nous l'avons vu plus haut, par l'état plus favorable thermodynamiquement du complexe enzyme/produit que du complexe enzyme/substrat qui favorise la formation de ce dernier en « tirant » la réaction dans le sens de la coupure (Hertel & Uhlenbeck, 1995). Des études identiques ont été réalisées par Osborne et collaborateurs, sur un ribozyme dérivé du ribozyme de *Schistosoma manzoni* et présentant des interactions interboucles (Osborne et al., 2005). Les valeurs obtenues, k_{-2} ≈ 0,32 min-1, k_2 ≈ 5,4 min-1 et Keq ≈ 17, montrent que la ligature est plus favorisée que pour le ribozyme minimum bien que la coupure soit encore 10 fois supérieure à la ligature. Les interactions tertiaires, en limitant l'espace conformationnel accessible c'est-à-dire en stabilisant le repliement du complexe enzyme/produit, augmentent l'activité de ligature du ribozyme. Encore une fois, les mêmes phénomènes sont observés pour le ribozyme en épingle à cheveux, puisque la présence de la jonction à quatre hélices native déplace l'équilibre interne vers la ligature et augmente le taux de repliement (Fedor, 1999; Tan et al., 2003).

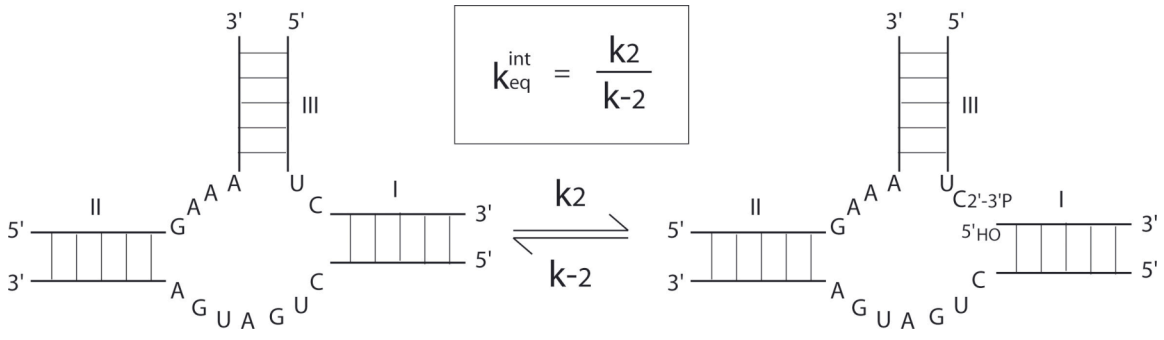


Figure 41 : Structure secondaire et réaction du cœur catalytique d'un ribozyme à tête de marteau minimum. La flèche rouge indique le site de coupure. La réaction de clivage est définie par la constante de clivage k_2 et la réaction de ligature est définie par la constante de ligation k_{-2}. La constante d'équilibre interne est définie dans l'encadré.

HYPOTHESES SUR LE MECANISME DE CIRCULARISATION DU MONOMERE APRES REPLICATION

Il apparaît donc que les éléments périphériques au cœur catalytique ne sont pas seulement impliqués dans la réaction de coupure mais également dans la réaction de ligature. Nous avons vu dans l'introduction que chaque copie monomérique du génome, produit de la réplication par cercle roulant et clivage par le ribozyme à tête de marteau, doit ensuite se circulariser pour servir à son tour de matrice. On sait peu de choses sur la circularisation *in vivo* du monomère mais elle pourrait impliquer la formation d'une structure ribozyme à tête de marteau qui catalyse une réaction de ligature 5'-3' correspondant à la réversion de la réaction de clivage. Cette réaction de ligature est identique à celle du ribozyme en épingle à cheveux. Le ribozyme à tête de marteau est capable d'auto-clivage et d'auto-ligature mais le mécanisme qui régule le passage de l'une à l'autre des activités est peu connu. Les ribozymes à tête de marteau étudiés dans les laboratoires favorisent, pour la majorité, la réaction de coupure plutôt que la ligature. Cependant, des études récentes montrent qu'une liaison covalente entre un nucléotide de l'hélice I et un nucléotide de l'hélice II, proches dans la structure cristallographique, induit un déplacement de l'équilibre de la réaction catalysée vers la ligature (Stage-Zimmermann & Uhlenbeck, 2001; Blount & Uhlenbeck, 2002). Cette liaison covalente n'a pas d'effet sur le taux de clivage, mais augmente significativement le taux de ligature d'environ 25 fois. Il a été proposé que les ribozymes ne comportant pas de liaison covalente subissaient des mouvements importants réduisant l'efficacité de ligature. La liaison covalente, en réduisant ces mouvements, augmente le taux de ligature. Ces résultats ont été confirmés par les études structurales de Dunham et collaborateurs (Dunham et al., 2003). Elles montrent qu'un lien entre les hélices I et II restreint la rotation de l'hélice I autour de son axe et prévient ainsi la réaction de coupure. Les auteurs suggèrent que cette restriction de mouvement soit responsable du passage d'un état permettant la coupure à un état permettant la ligature. Mais des expériences de circularisation du substrat qui contraint les mouvements des hélices I et II dans le complexe enzyme-substrat, montrent qu'il n'y a pas d'augmentation du taux de ligature (Stage-Zimmermann & Uhlenbeck, 1998). Ils restent ainsi un certain nombre d'incertitudes sur les

contraintes structurales et les mouvements subis par le ribozyme pour passer d'un état favorisant le clivage à un état favorisant la ligature.

Les ARN des viroïdes appartenant à la famille des Pospiviroides sont probablement ligués par une ARN ligase comme l'indique la présence d'un 2' phosphomonoester et d'une liaison 3'-5' phosphodiester au site de ligature (Kiberstis et al., 1985). Par contre, les informations sur la ligature des ARN viroïdes de la famille des Avsunviroidae sont assez limitées. Toutefois, dans le cas de la circularisation du génome de PLMVd, il a été observé la présence d'une liaison 2'-5' phosphodiester *in vitro* (Cote & Perreault, 1997) et *in vivo* (Cote et al., 2001). Cette liaison préviendrait la coupure du ribozyme et assurerait ainsi le maintien d'un ARN matriciel circulaire indispensable à la réplication par mécanisme de cercle roulant ("rolling-circle"). Peut-on imaginer que le ribozyme à tête de marteau catalyse une réaction de ligature menant à la formation d'une liaison 2'-5' phosphodiester ? Semlow et Silverman ont montré que les désoxyribozymes sélectionnés par leur capacité à lier un groupement 5'-hydroxyle avec un phosphate cyclique 2'-3', lient préférentiellement les positions 2' et 5' au lieu des positions 3' et 5' indépendamment de leur environnement (Semlow & Silverman, 2005). La réaction de ligature impliquant un phosphate 2'-3' cyclique mènerait à une liaison non native de l'ARN plutôt qu'à la liaison native 3'-5' phosphodiester. Si le ribozyme à tête de marteau était capable de catalyser la formation d'une liaison 2'-5' phosphodiester il ne pourrait pas réaliser la réaction inverse. La ligature serait alors irréversible ce qui assurerait la présence d'ARN génomique circularisé pouvant servir de matrice lors de la réplication. L'hypothèse est séduisante mais très peu probable. D'abord, la ligature 2'-5' phosphomonoester catalysée par le désoxyribozyme se forme dans un contexte structural particulier: les deux nucléotides entre lesquels se fait la liaison forment des appariements Watson-Crick et appartiennent à une hélice. D'autre part, PLMVD, après clivage du concatémère en monomère, se replierait en une structure alternative à la structure ribozymique. Dans cette structure alternative, les deux nucléotides à liguer forment des appariements Watson-Crick et appartiennent à une hélice tout comme dans le désoxyribozyme.

Ainsi, la ligature pourrait nécessiter une structuration alternative de l'ARN viroïde monomère permettant la formation d'une liaison 2'-5' phosphomonoester. L'autre hypothèse est la régulation du passage du ribozyme à tête de marteau d'un état favorisant le clivage à un état favorisant la ligature (formation d'une

liaison native 3'-5' phosphodiester) probablement par les mouvements des hélices I et II.

REGULATION DE L'ACTIVITE RYBOZYME A TETE DE MARTEAU

La réplication par mécanisme de cercle roulant symétrique demande que l'ARN linéaire obtenu par autocoupure du ribozyme à tête de marteau se circularise pour servir de matrice. La manière dont la ligature se réalise, comme nous l'avons vu précédemment, n'est pas encore claire pour tous les ARN concernés. Mais une fois que l'ARN est circularisé, dans le cas de la formation de la liaison native 3'-5' phosphodiester, il doit rester sous cette forme suffisamment longtemps pour que l'ARN polymérase en réalise plusieurs copies ; ceci implique que le ribozyme ne catalyse pas de réaction de coupure. Un mécanisme de régulation de l'activité du ribozyme à tête de marteau est donc envisageable. À ce jour, au vu des résultats obtenus, deux voies de régulation sont envisageables : l'une propose la dimérisation de ribozyme, l'autre le repliement de l'ARN en une structure alternative à la structure catalytique.

Régulation par dimérisation du ribozyme :

L'influence des régions périphériques sur l'activité du ribozyme a été mise en évidence chez les ribozymes de type 1 dont l'hélice III est fermée par une boucle très courte (voir introduction) allant de zéro à deux nucléotides. Cette boucle très courte est responsable de l'instabilité de l'hélice III dans le cas du ribozyme des brins (+) et (-) de ASBV. Dans le cas du ribozyme du triton elle est dans l'impossibilité même de se replier (Hutchins et al., 1986; Forster & Symons, 1987; Forster et al., 1988). Parmi les structures d'ARN à haute résolution disponibles, les boucles à deux nucléotides sont très rares et les boucles inférieures à deux nucléotides ne sont jamais observées. Or la formation de l'hélice III comme des deux autres hélices est indispensable à l'activité du ribozyme. Comment ces ribozymes, dont l'hélice III ne se forme probablement pas, peuvent-ils adopter une structure active ? En fait, Forster et collaborateurs ont montré en 1988 par analyse des séquences de ces ribozymes que des structures contenant deux ribozymes à tête de marteau, donc deux sites de coupure, pouvaient être construites par association de deux molécules de même

polarité. Des expériences *in vitro* ont révélé que le ribozyme du brin (+) de ASBV et le ribozyme du triton, dont l'hélice III est instable, n'étaient actifs que sous forme dimérique alors que le ribozyme du virusoïde de la mosaïque de la luzerne (vLTSV), qui possède une hélice III stable, était actif sous forme monomérique. En effet, dans le cas des ribozymes de ASBV et du triton, le pourcentage de clivage était dépendant de la concentration de la solution en ARN ce qui est révélateur d'une interaction intermoléculaire alors que le pourcentage de clivage était indépendant de la concentration dans le cas de vLTSV ce qui est caractéristique d'une réaction intramoléculaire. Ainsi, les premiers fonctionneraient sous forme de dimères dont l'hélice III alors suffisamment longue permettrait la coupure, tandis que l'hélice III stable du second lui permet d'être actif sous forme monomère (Forster et al., 1988). Lors de la réplication du génome circulaire des viroïdes ASBV, comme vu dans l'introduction, il y a coupure de l'ARN synthétisé par catalyse ribozymique et obtention d'un simple brin. Le modèle « double ribozyme » implique qu'il n'y ait pas coupure après synthèse d'une copie de l'ARN matrice puisque le ribozyme ne peut fonctionner en monomère mais qu'au minimum deux copies du génome soient synthétisées pour que la coupure ait lieu. Ce modèle doit toutefois être pris avec précaution. Les ribozymes des salamandres *Amphiuma tridactulum* et *Ambystoma talpoideum* appartiennent à la même sous classe (type 1) que le ribozyme du triton ; ils ne possèdent aucun nucléotide dans la boucle terminale de l'hélice III. Pourtant Garret et collaborateurs ont montré que le taux clivage n'était pas dépendant de la concentration en ARN ce qui implique une réaction intramoléculaire ou monomérique (Garrett et al., 1996). Ils proposent que l'instabilité de l'hélice III soit compensée par les interactions entre les hélices I et II.

Cette capacité de certains ribozymes à dimériser, nous a interpellé alors que nous tentions de cristalliser un ribozyme pourvu de régions périphériques et synthétisé chimiquement. En effet, après modélisation des interactions tertiaires interboucle, l'étape suivante était d'obtenir une structure cristallographique d'un ribozyme naturel. Nous avons testé différentes constructions basées sur VLTSV et STRSV, longues d'environ 80 nucléotides, déshydroxylées en 2' du site actif et toutes synthétisées par Dharmacon. Après déprotection et purification, les ARN ont été renaturés dans différentes conditions. Les protocoles utilisés sont décrits

dans la revue « Preparation and handling of RNA crystals » en annexe du manuscrit. Le dépôt sur gel de polyacrylamide non dénaturant des solutions d'ARN destinées à la cristallogenèse, permet de vérifier que la population d'ARN repliés qu'elle contient est homogène.

Malgré les différentes températures de renaturation, les différentes concentrations de Mg^{2+} ou les différents protocoles de renaturation testés, les solutions de renaturation contenaient systématiquement deux populations d'ARN repliés. Comme nous l'avons vu précédemment, les ribozymes possédant une hélice III instable seraient capables de dimériser. Les constructions d'ARN que nous avons tentées de cristalliser ont été conçues avec une hélice III stable. Elles n'étaient donc pas supposées s'ouvrir et interagir avec d'autres molécules. Toutefois, afin d'éviter la formation de multimères, nous avons renaturé les ARN à une concentration très diluée pour favoriser le repliement intramoléculaire et empêcher les interactions type duplex étendu avant de concentrer la solution pour obtenir des concentrations exploitables en cristallogenèse. Comme le montre la Figure 42, c'est dans ces conditions de renaturation diluée puis reconcentrée, que le conformère qui migre le plus loin est favorisé (piste 3). Mais nous n'avons pas réussi à trouver les conditions de renaturation permettant d'obtenir une seule conformation de l'ARN.

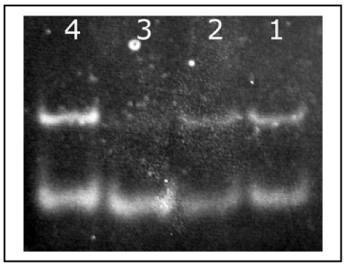


Figure 42 : Gel natif polyacrylamide 10% 5mM Mg^{2+} d'une construction de STRSV. 1 : ARN renaturé à 80°C concentration 0,5 mM, 2 : ARN renaturé à 80°C dilué 1000x puis concentré 0,25mM, 3 : ARN non renaturé concentré 0,5mM, 4 : ARN renaturé à 50°C concentration 0,5 mM

Osborne et collaborateurs, dans leurs études cinétiques de dérivés de ribozyme de *S. mansoni*, ont montré par séparation sur gel non dénaturant que le ribozyme se repliait en deux conformères différents (Osborne et al., 2005). Nous avons également testé de manière préliminaire la renaturation des différentes constructions en présence d'antibiotique. En effet, il a été montré que la néomycine, antibiotique de la classe des aminoglycosides, inhibait la réaction de

clivage du ribozyme en stabilisant le complexe E-S (Stage et al., 1995). Une stabilisation accrue de la structure ribozyme pourrait ainsi faciliter la cristallisation. Malheureusement, nous avons obtenu un nombre de conformères différents encore plus important que dans le cas des solutions de renaturation sans antibiotiques. Malgré la présence de différentes conformations dans nos solutions renaturées d'ARN, nous avons réalisé des essais de cristallisation des différentes constructions.

Différentes conditions de cristallisation utilisées pour cristalliser des ARN de longueur supérieure à 27 nucléotides ont été extraites de la littérature (voir tableau en annexe). Des conditions présentant différentes concentrations d'ARN, de $MgCl_2$, de glycérol, de spermine et de NaCl ont été testées ainsi que la nature et la concentration des agents précipitants tels que le MPD, PEG400 et le PEG4000. Le logiciel INFAC (Incomplete Factorial Plan) a été utilisé pour créer le plan factoriel incomplet présenté figure 43. Toutes les conditions contiennent 50mM de Na cacodylate pH 6.5. 1 µL de solution de cristallisation a été ajouté à 1 µL de solution d'ARN renaturé. Les gouttes ont été testées à 20°C et 37°C. Nous n'avons obtenu aucun cristal dans ces conditions. Les kits commerciaux Natrix et Mini screen ont également été testés sans plus de résultats.

Les essais de cristallisation sont à ce jour restés infructueux. Il apparaît clairement qu'il faudra à l'avenir tenter d'obtenir des solutions de cristallisation ne contenant qu'une seule espèce conformationnelle de ribozyme. Le problème de la synthèse se pose également. La synthèse chimique de ces ARN longs de 80 nucléotides environ représentait à l'époque de nos essais un véritable tour de force. La quantité produite était alors assez faible et le nombre de test de cristallisation réalisé pour chaque commande de produit était assez réduit. L'utilisation du nouveau robot de cristallogenèse « Mosquito » qui réalise des nanogouttes permettra à l'avenir de tester de nombreuses conditions avec une quantité réduite d'ARN. D'autre part, cette capacité des ribozymes testés à former des multimères, peut-être des dimères, devra être étudiée. La dimérisation pourrait être un moyen de régulation de la catalyse des ribozymes naturels dont l'hélice III est stable.

A1		A2		A3		A4		A5		A6	
MPD	5%	MPD	5%	MPD	5%	MPD	15%	MPD	15%	MPD	15%
Glycerol	5%	Glycerol	1%	Glycerol	5%	Glycerol	5%	Glycerol	1%	Glycerol	1%
MgCl2	1 mM	MgCl2	5mM	MgCl2	10 mM	MgCl2	1 mM	MgCl2	5mM	MgCl2	20 mM
spermine	5 mM	spermine	10 mM	spermine	5 mM	spermine	10 mM	spermine	10 mM	spermine	1 mM
NaCl (M)	50 mM	NaCl (M)	100 mM	NaCl (M)	100 mM	NaCl (M)	50 mM	NaCl (M)	100 mM	NaCl (M)	50 mM
RNA	0,3 mM	RNA	0,1 mM	RNA	0,1 mM	RNA	0,1 mM	RNA	0,1 mM	RNA	0,5 mM
B1		**B2**		**B3**		**B4**		**B5**		**B6**	
MPD	25%	MPD	25%	PEG400	5%	PEG400	5%	PEG400	5%	PEG400	10%
Glycerol	1%	Glycerol	5%	Glycerol	5%	Glycerol	5%	Glycerol	5%	Glycerol	1%
MgCl2	5mM	MgCl2	1 mM	MgCl2	1 mM	MgCl2	10 mM	MgCl2	20 mM	MgCl2	1 mM
spermine	1 mM	spermine	10 mM	spermine	10 mM	spermine	1 mM	spermine	5 mM	spermine	5 mM
NaCl (M)	100 mM	NaCl (M)	100 mM	NaCl (M)	100 mM	NaCl (M)	100 mM	NaCl (M)	50 mM	NaCl (M)	100 mM
RNA	0,1 mM	RNA	0,3 mM	RNA	0,5 mM	RNA	0,3 mM	RNA	0,1 mM	RNA	0,5 mM
C1		**C2**		**C3**		**C4**		**C5**		**C6**	
PEG400	10%	PEG400	10%	PEG400	20%	PEG400	20%	PEG400	20%	PEG4000	5%
Glycerol	1%	Glycerol	5%	Glycerol	1%	Glycerol	1%	Glycerol	1%	Glycerol	5%
MgCl2	5mM	MgCl2	20 mM	MgCl2	1 mM	MgCl2	20 mM	MgCl2	10 mM	MgCl2	1 mM
spermine	5 mM	spermine	10 mM	spermine	10 mM	spermine	5 mM	spermine	10 mM	spermine	1 mM
NaCl (M)	100 mM	NaCl (M)	100 mM	NaCl (M)	50 mM	NaCl (M)	50 mM	NaCl (M)	50 mM	NaCl (M)	100 mM
RNA	0,5 mM	RNA	0,3 mM	RNA	0,3 mM	RNA	0,5 mM	RNA	0,3 mM	RNA	0,5 mM
D1		**D2**		**D3**		**D4**		**D5**		**D6**	
PEG4000	5%	PEG4000	5%	PEG4000	10%	PEG4000	10%	PEG4000	20%	PEG4000	20%
Glycerol	5%	Glycerol	1%	Glycerol	1%	Glycerol	1%	Glycerol	5%	Glycerol	5%
MgCl2	5mM	MgCl2	10 mM	MgCl2	5mM	MgCl2	10 mM	MgCl2	10 mM	MgCl2	20 mM
spermine	5 mM	spermine	1 mM	spermine	1 mM	spermine	10 mM	spermine	5 mM	spermine	10 mM
NaCl (M)	50 mM	NaCl (M)	100 mM	NaCl (M)	50 mM	NaCl (M)	50 mM	NaCl (M)	50 mM	NaCl (M)	50 mM
RNA	0,3 mM	RNA	0,3 mM	RNA	0,1 mM	RNA	0,5 mM	RNA	0,1 mM	RNA	0,5 mM

Figure 43 : Conditions de cristallisation générées par le logiciel INFAC.

Régulation par formation de structures alternatives :

La formation de structures alternatives à la structure en ribozyme à tête de marteau, représente également un moyen de régulation envisageable de la réaction de la catalyse ribozymique. Miller et Silver ont montré que la structure secondaire de l'ARN satellite sBYDV comportait une structure similaire à la celle du ribozyme à tête de marteau dont la boucle terminale de l'hélice I forme un pseudonoeud avec une boucle interne de l'hélice II (Miller & Silver, 1991). En empêchant la formation du pseudonoeud par la substitution des nucléotides, l'activité d'autocoupure du ribozyme est multipliée par 400. Comme nous l'avons montré par la modélisation du ribozyme naturel, du fait de la position parallèle des hélices, la formation d'appariements Watson-Crick entre les boucles est impossible. Un autre positionnement des hélices implique des changements drastiques dans la structuration du cœur catalytique, peu favorables à la conservation d'une activité catalytique. La formation du pseudonoeud de sBYDV sous entend donc une restructuration de l'ensemble de la structure avec perte du ribozyme à tête de marteau. La formation du pseudonoeud ferait passer l'ARN en position « off » empêchant la formation du ribozyme et l'autocatalyse alors que

le dépliement du pseudonoeud permettrait la formation du cœur catalytique et donc la catalyse ; le pseudonoeud pourrait donc être considéré comme un « interrupteur ».

HYPOTHESE SUR L'ABSENCE DE RIBOZYME A TETE DE MARTEAU DE TYPE 2

Nous avons vu dans l'introduction que seuls sont observés dans la nature des ribozymes à tête de marteau de type 1 et de type 3. Les ribozymes de type 2 n'ont pas été identifiés. Nous nous sommes interrogés durant ce travail sur l'absence de ribozyme de type 2 des séquences de viroïde ou virusoïdes de plantes. Il semble en effet que la séquence d'un tel ribozyme n'ait jamais été trouvée jusqu'à présent dans un génome séquencé. Nous avons envisagé deux structurations possibles d'un ribozyme de type 2 (Figure 44). Dans la première, l'hélice III et l'hélice I sont fermées par une boucle terminale tandis que l'hélice II, sans boucle interne comporte l'entrée 5' dans la structure ribozymique. Cette structure qui ne comporte pas les éléments structuraux nécessaires à la formation des interactions tertiaires, est difficilement envisageable. Une deuxième structure possible présente une boucle interne sur l'hélice II ; les interaction tertiaires seraient alors conservées. Mais pour les deux structures envisagées, en cas de coupure, le cœur catalytique divisé ne comporterait plus aucune des paires non canoniques qui forme le cœur catalytique. Pour les types 3 et 2, le départ du produit ne conduit pas à la perte de la moitié des nucléotides du cœur catalytique et la structure catalytique est conservée. Dans l'hypothèse où c'est le ribozyme à tête de marteau qui réaliserait la réaction de ligature du monomère, la conservation de la structure du coeur catalytique après coupure faciliterait la reformation du ribozyme divisé en deux parties à chaque extrémité du monomère (Figure 45). Le type 2, dans lequel la structure du coeur catalytique serait perdue après coupure, aurait du mal à se reformer lors de la circularisation, ce qui expliquerait qu'il ne soit jamais observé dans les séquences.

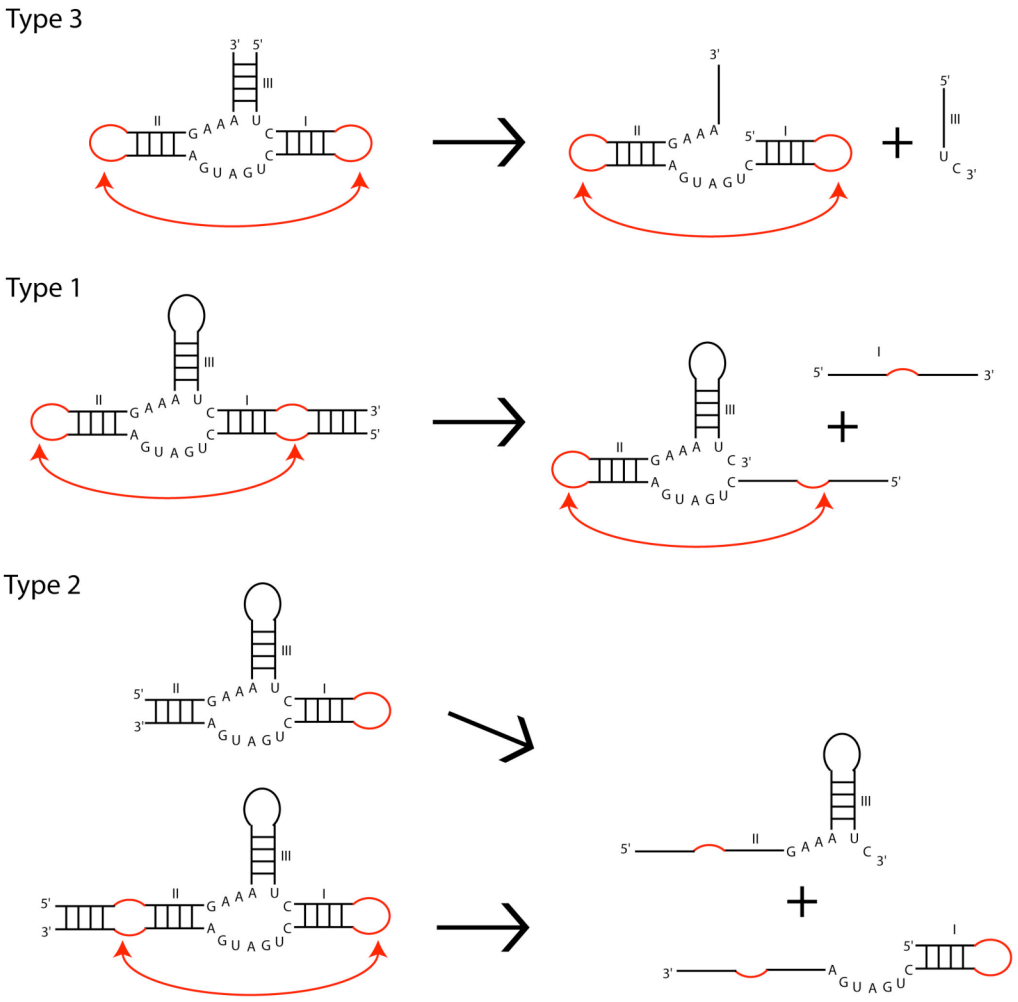


Figure 44 : Hypothèse expliquant l'absence de ribozyme à tête de marteau de type II

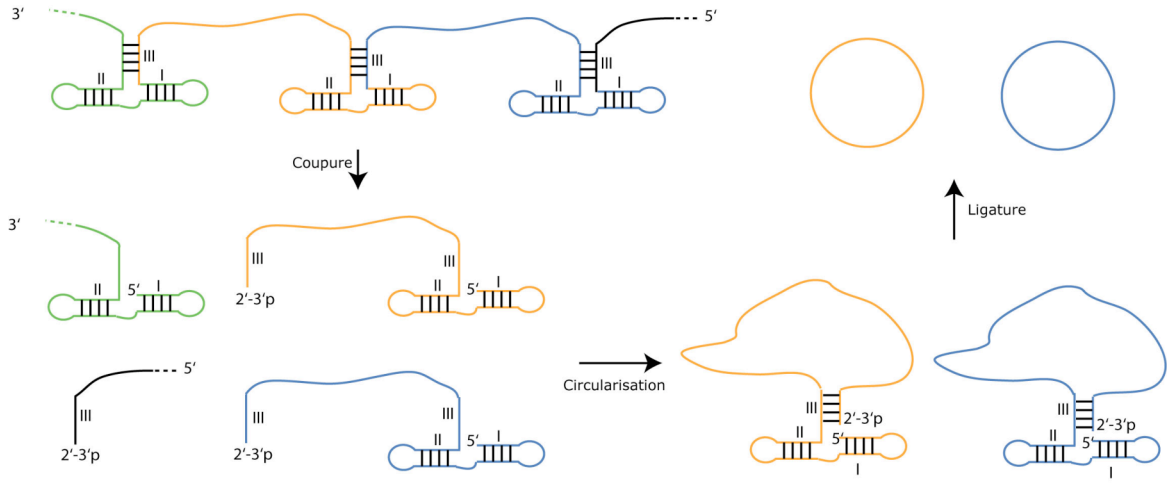


Figure 45 : Hypothèse de mécanisme de passage du multimère aux nouvelles matrices ARN circulaires. Le multimère est composé d'une succession de copies du génome séparées par un ribozyme à tête de marteau. Par clivage du ribozyme chaque copie du génome forme un monomère présentant à chaque extrémité une partie du ribozyme. Les deux extrémités doivent se rapprocher, et donc le monomère se circulariser. Après ligature réalisée par le ribozyme à tête de marteau les ARN génomiques circulaires peuvent servir à leur tour de matrices.

CONSERVATION DES INTERACTIONS ARN-ARN DANS LES RIBOZYMES OUTILS

Les ribozymes minimaux ont été souvent présentés comme des outils potentiels utilisés pour cliver en trans certains ARN pathogènes ou hôtes. Des ribozymes coupant en trans mais conservant les éléments périphériques essentiels à l'activité catalytique en conditions dénaturantes, pourraient être utilisés dans des applications intracellulaires. Une étude récente a montré qu'il était possible de construire des dérivés du ribozyme sTRSV qui clivent en trans tout en conservant les boucles terminales 1 et 2 (Burke & Greathouse, 2005). Ce ribozyme artificiel, dont les extrémités 5' du ribozyme et 3' du substrat appartiennent à l'hélice I, est actif à des concentrations physiologiques de Mg^{2+}. En effet, malgré la discontinuité de l'hélice I, les interactions tertiaires sont maintenues. Ce ribozyme artificiel pourrait être utilisé dans différentes applications intracellulaires dont éventuellement la thérapie génique.

POUR CONCLURE

Comme nous l'avons vu, les ribozymes à tête de marteau tout comme les ribozymes en épingle à cheveux ont été pendant longtemps étudiés sous leur forme simplifiée et affaiblis par la perte des appendices périphériques indispensables à leur activité optimale en conditions natives. Ces régions périphériques ont pour rôle de faciliter le repliement en une structure active et sont nécessaires en conditions physiologiques bien qu'elles ne soient pas directement impliquées dans la catalyse. Ces résultats soulignent les liens fondamentaux qui lient le repliement, les interactions tertiaires et la catalyse des ribozymes. D'un point de vue plus général, cela pose la question de la pertinence des conclusions tirées des résultats obtenus sur des systèmes simplifiés à l'extrême.

2.4. Classification des jonctions triples

2.4.1. Introduction

Nous avons vu dans les chapitres précédents que différents motifs participaient de manière indirecte à la formation des interactions tertiaires. Ainsi, dans la structure de P4-P6 de *Tetrahymena*, le coude J5/J5a entraîne un positionnement parallèle des hélices P5 et P5a. Ainsi orientée, la boucle terminale GAAA qui ferme P5b peut interagir avec le récepteur de GAAA à 11 nucléotides de l'hélice P6b. Points de branchement de plusieurs hélices, les jonctions sont courantes dans les structures d'ARN. Les jonctions à trois ou quatre hélices sont fréquemment observées dans les structures secondaires prédites ou les structures tridimensionnelles connues d'ARN fonctionnels où elles jouent un rôle majeur dans l'organisation de l'architecturale globale. Au même titre que les motifs, elles participent à l'orientation dans l'espace d'éléments qui formeront des interactions tertiaires. Les structures du ribozyme à tête de marteau et du ribozyme en épingle à cheveux sont organisées, comme nous l'avons vu dans le chapitre 2.3, autour d'une jonction qui joue un rôle clé dans l'architecture globale de ces ribozymes. Dans le cas du ribozyme en épingle à cheveux, une corrélation directe entre la structure de la jonction à quatre hélices et l'activité du ribozyme a été démontrée. Le ribozyme en épingle à cheveux catalyse la même réaction de transestérification que le ribozyme à tête de marteau et est également capable de catalyser la réaction de ligature inverse. Le ribozyme en épingle à cheveux comporte deux éléments structuraux majeurs, (i) la jonction à quatre hélices, et (ii) les deux boucles internes portées par les hélices adjacentes A et B (figure 46A). Les études de FRET montrent que les hélices sont empilées par paires et que le brin continu est orienté de manière anti-parallèle (Murchie et al., 1998). Cette architecture permet aux boucles des hélices A et B de former des interactions tertiaires comme le montrent les structures cristallographiques (Figure 46B) (Rupert & Ferre-D'Amare, 2001; Rupert et al., 2002). Les interactions entre les boucles internes A et B sont essentielles à l'activité du ribozyme ; elles sont responsables d'une augmentation de l'activité du ribozyme de 10^5 fois. Le remplacement de la jonction par un coude entre les deux hélices A et B induit la perte des hélices C et D mais

conserve les interactions entre les boucles. L'activité de cette forme minimale est présente à des concentration élevées de Mg^{2+} mais elle est abolie à des concentrations physiologiques de Mg^{2+} (Berzal-Herranz et al., 1993; Zhao et al., 2000). Les études de FRET de molécules uniques montrent que le repliement du ribozyme en épingle à cheveux est accéléré en présence de la jonction à quatre hélices et qu'il s'effectue en deux étapes : (i) positionnement parallèle des hélices A et B, et (ii) formation des interactions tertiaires entre les boucles (Tan et al., 2003). La dernière étape mène à la forme active du ribozyme impliquant des interactions tertiaires entre les boucles. Le taux de formation de ces interactions en présence de la jonction à quatre hélices est 500 fois supérieur à ce qui est mesuré dans le cas du ribozyme minimum. Le parallèle avec le rôle des éléments périphériques dans l'activité du ribozyme à tête de marteau est frappant : la jonction à quatre hélices du ribozyme en épingle à cheveux n'est pas essentielle à l'activité chimique proprement dite, mais assure une activité optimale en conditions cellulaires. Dans les deux cas, les éléments auxiliaires, la jonction pour le ribozyme en épingle à cheveux et les boucles pour les ribozymes à tête de marteau, facilitent le repliement en sélectionnant ceux où les interactions tertiaires sont les plus favorables à la catalyse.

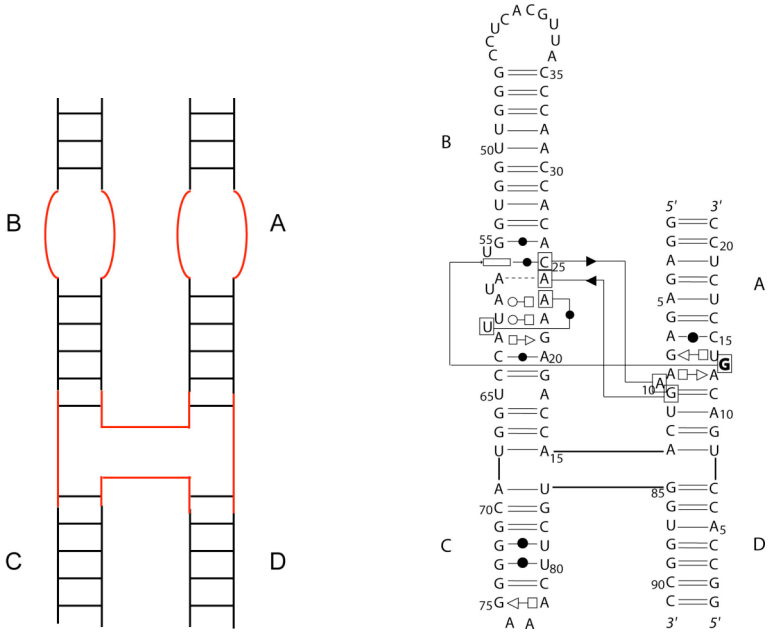


Figure 46 : Structure du ribozyme en épingle à cheveux. A) Deux éléments structuraux fondamentaux : la jonction à quatre hélices et les boucles internes (en rouge). B) Diagramme de la structure cristallographique. Code PDB : 1HP6 (Rupert & Ferre-D'Amare, 2001).

Nous venons de voir que la jonction à quatre hélices du ribozyme en épingle à cheveux, indispensable à l'activité cellulaire du ribozyme, favorise le rapprochement des boucles internes des hélices A et B et donc la formation des interactions tertiaires. Dans le cas du ribozyme à tête de marteau, les interactions tertiaires entre les boucles des hélices I et II sont indispensables à son activité cellulaire. Ces interactions entre les boucles sont possibles grâce à l'orientation parallèle des hélices imposée par la jonction à trois hélices qui constitue le cœur catalytique du ribozyme. Dans les deux cas, c'est une jonction qui assure le positionnement d'éléments structuraux leur permettant d'établir les interactions tertiaires assurant la fonctionnalité de l'ARN dans son environnement naturel. Les jonctions constituent dans ces deux ribozymes mais également dans de nombreux ARN, des modules structurés critiques pour la promotion des interactions à longue distance assurant une architecture globale fonctionnelle (Figure 47).

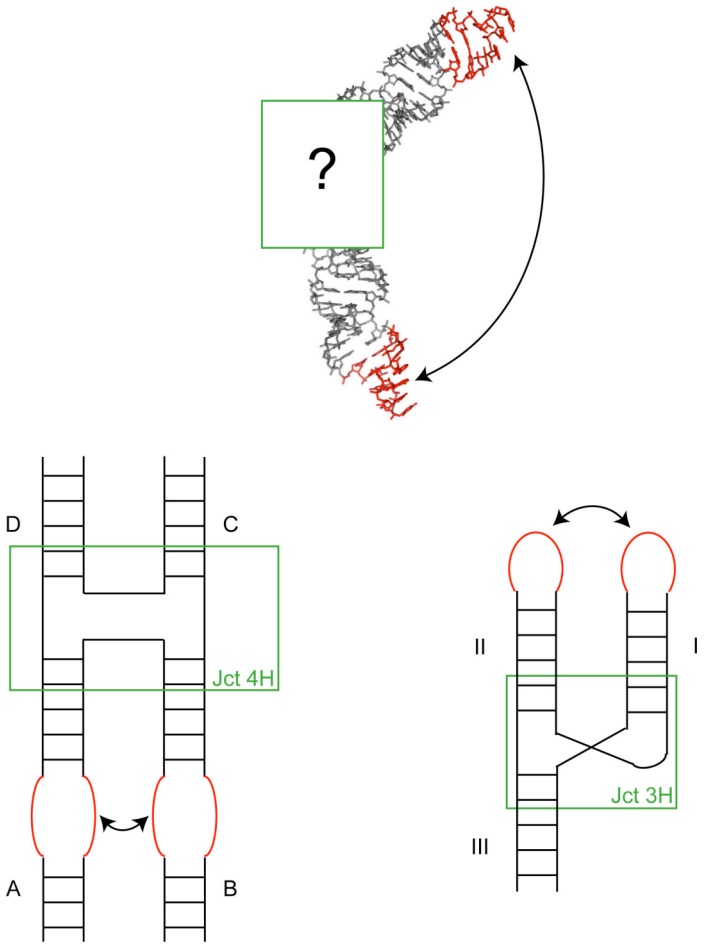


Figure 47 : Différentes topologies permettent d'assurer le rapprochement dans l'espace d'éléments qui doivent interagir. La jonction à quatre hélices du ribozyme en épingle à cheveux et la jonction à trois hélices du ribozyme à tête de marteau assurent le positionnement des boucles qui interagissent.

La structure et le repliement des jonctions à trois ou quatre hélices d'ADN ont été beaucoup plus étudiées que les jonctions d'ARN. Certaines caractéristiques des jonctions ADN ont pu être étendues aux jonctions ARN comme, par exemple, l'empilement des hélices. Mais globalement, malgré leur importance dans de nombreux ARN fonctionnels, la structuration des jonctions à trois hélices d'ARN est peu connue. Cela nous a posé de nombreux problèmes lors de la modélisation d'ARN structuré de grande taille présentant des jonctions à trois hélices car leur rôle sur le positionnement des éléments périphériques, peut être, comme nous l'avons vu plus haut, déterminant pour la structure globale de l'ARN. La difficulté réside dans le choix de la topologie de la jonction. En effet, à partir de la structure secondaire d'une jonction triple (ou à trois hélices), trois topologies de structures tridimensionnelles différentes, en envisageant toutes les possibilités d'empilement entre les trois hélices, peuvent être proposées (figure 48). Quelle que soit la topologie, on distingue un brin commun aux deux hélices empilées dont les nucléotides du segment entre hélices, quand il y en a, sont libres ou impliqués dans des interactions non-Watson-Crick et deux segments entre les hélices empilées et la troisième hélice. Il fallait donc déterminer quelles étaient les caractéristiques de la structure secondaire qui nous permettraient de faire un choix entre les trois topologies possibles de la structure tridimensionnelle.

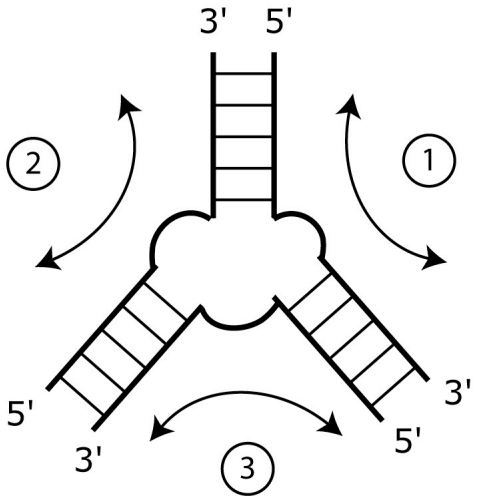


Figure 48 : Topologie de jonction triple. A partir de la structure secondaire d'une jonction à trois hélices, trois empilements différents sont envisageables entraînant trois topologies différentes et diverses possibilités d'interactions à longue distance.

Nous avons choisi de récolter et d'analyser les jonctions triples dont la structure cristallographique est connue, pour extraire les caractéristiques tridimensionnelles lisibles au niveau de la structure secondaire et qui nous permettraient d'établir des règles topologiques. Les objectifs principaux du travail étaient de déterminer si les jonctions triples pouvaient être divisées en familles en fonction (i) de la longueur des brins jonction et (ii) de la présence d'interactions non Watson-Crick caractéristiques. Au vu de ces analyses et afin de faciliter la déduction de la topologie d'une jonction triple à partir de la structure secondaire, nous proposons quelques règles basées sur la longueur des segments simple brin entre les hélices et la nature des nucléotides qu'ils comportent. Trois familles, A, B et C ont pu être isolées en fonction de la longueur des segments simple brin et du type d'interactions tertiaires entre ces segments et les hélices empilées. Pour les deux familles les mieux caractérisées (A et C) nous proposons une structure consensus révélatrice du maintien du nombre de nucléotides dans les simples brins jonctions et de la nature des bases impliquées dans des paires non Watson-Crick conservées. La famille C, comportant le plus de membres, nous semble être la plus intéressante des trois. Elle présente deux types d'interactions tertiaires : d'une part entre l'une des hélices empilées et la troisième hélice, et, d'autre part, entre les nucléotides libres du segment qui joint la troisième hélice et la deuxième hélice des hélices empilées et le petit sillon de cette hélice. Ces interactions assurent une stabilité de l'ensemble de la jonction. La famille C est la seule famille comportant des jonctions d'autres ARN que les ARN ribosomiques.

À l'aide des règles établies, nous proposons finalement une solution topologique pour différentes jonctions triples qui sont d'un grand intérêt fonctionnel pour les ARN auxquels elles appartiennent et dont la structure tridimensionnelle n'est pas encore connue. Une de nos propositions a été validée par la publication de la structure cristallographique de l'ARN concerné pendant la révision de l'article.

2.4.2. Article 4 : "Topology of three-way junctions in folded RNAs"

Aurélie Lescoute & Eric Westhof

RNA. 2006 Jan;12(1):83-93.

2.4.3. Discussion

Nous avons, dans l'article précédent, proposé des règles basées sur la structure 2D permettant de faciliter la prédiction de la structure tridimensionnelle des jonctions à trois hélices. Nous donnons quelques exemples de prédictions dont un concerne la jonction triple du complexe U4-U6 du spliceosome.

Le spliceosome, formé de plusieurs complexes (snRNP) constitués de petits ARN nucléaires (snARN) riches en uraciles et de protéines Sm, catalyse la réaction d'épissage des ARN pré-messagers (pré-ARNm) eucaryotes. L'assemblage du spliceosome sur le pré-ARNm implique une dynamique des interactions entre snRNP, et, snRNP et ARN pré-messager. Une première étape consiste en la fixation des snRNP U1 et U2 sur le site de clivage en 5' et le site de branchement respectivement, suivie de la fixation du complexe U4/U6.U5 tri-snRNP. La conversion du spliceosome en une machine active implique de nombreux réarrangements touchant en particulier l'ARN du complexe tri-snRNP. Dans le complexe U4/U6 snRNP, les ARN U4 et U6 forment une jonction triple hautement conservée phylogénétiquement et constituée de deux hélices intermoléculaires (hélices I et II) et d'une hélice, appartenant à U4, terminée par une boucle. Le passage à l'état actif du spliceosome entraine l'interruption des hélices intermoléculaires. L'ARN U6 se replie alors sur lui même pour interagir avec l'ARN U2 et former ainsi une partie du site catalytique. Finalement le complexe snRNP U4 se dissocie du spliceosome ou reste faiblement attaché. Le mécanisme du passage à l'état actif du spliceosome est encore mal compris mais il apparait clairement que le complexe des snRNP U4 et U6 joue un rôle important.

Nous travaillons en ce moment à l'obtention d'un modèle tridimensionnel de la jonction à trois hélices U4/U6 qui soit cohérent avec les données de fluorescence de l'équipe de R. Lührmann avec laquelle nous collaborons. Leur différents résultats d'empreinte en solution leur ont permis de nous fournir la structure secondaire du complexe (Figure 49). Nous avons utilisé les règles de repliement déterminées dans l'article 4 pour identifier la topologie de la jonction. Les séquences des régions en simple brin de la jonction ne montraient pas les caractéristiques particulières d'une jonction appartenant à la famille C. Nous avions donc le choix entre la famille A et la famille B. Le nombre à peu près

identique de nucléotides dans les brins jonctions suggérait une appartenance à la famille B. La jonction 16S H33H33aH33b de la famille B présente un certain nombre de similarités de séquence et nous nous en sommes inspirés pour modéliser la jonction triple du complexe U4/U6. Nous avons ensuite mesuré les distances entre les positions sur lesquelles nos collaborateurs avaient fixé les fluorophores et nous les avons comparées avec les distances qu'ils avaient obtenues. Les positions des donneurs et accepteurs sont indiquées à la fois sur la représentation 2D et le modèle 3D ci-dessous (Figures 49 et 50). Les distances mesurées sur le modèle et par fluorescence sont indiquées dans le diagramme (Figure 50). La comparaison avec les données de fluorescence laisse penser que le modèle s'approche de la structure en solution. Pourtant, nous avons besoin de données supplémentaires pour pouvoir affiner cette hypothèse. Nous avons ainsi suggéré à nos collaborateur de positionner les fluorophores à des distances plus lointaine du "coeur" de la jonction. Ceci pour éviter d'une part les potentiels problèmes de repliement du à la présence du fluorophore dans le brin jonction, comme ca pourrait être le cas pour U11, mais également pour augmenter l'amplitude des différences entre les données. En effet, plus les fluorophores seront loin du coeur de la jonction plus les distances seront interprétables. Si les fluorophores sont aux extrémités de ces hélices, on pourra mieux déterminer si l'hélice I est très proche de l'hélice III, et donc très loin de l'hélice II. Il pourrait même être envisageable d'allonger ces hélices pour augmenter l'amplitude des données.

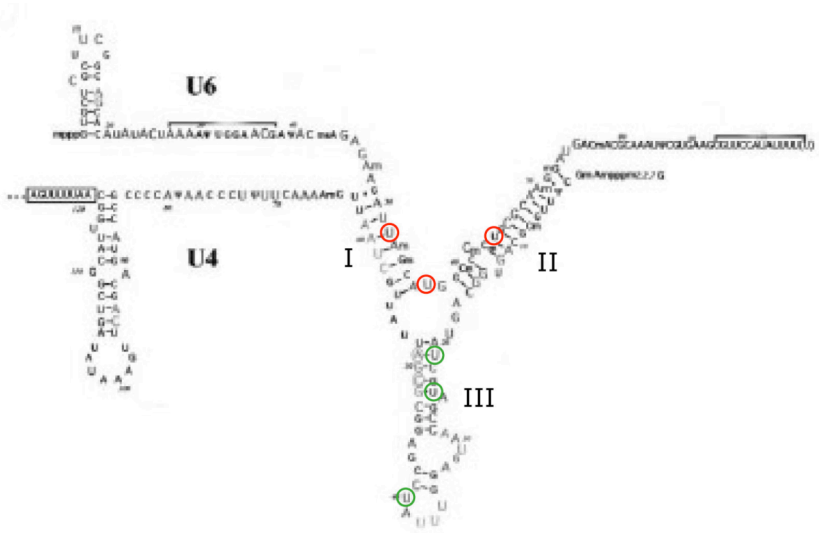
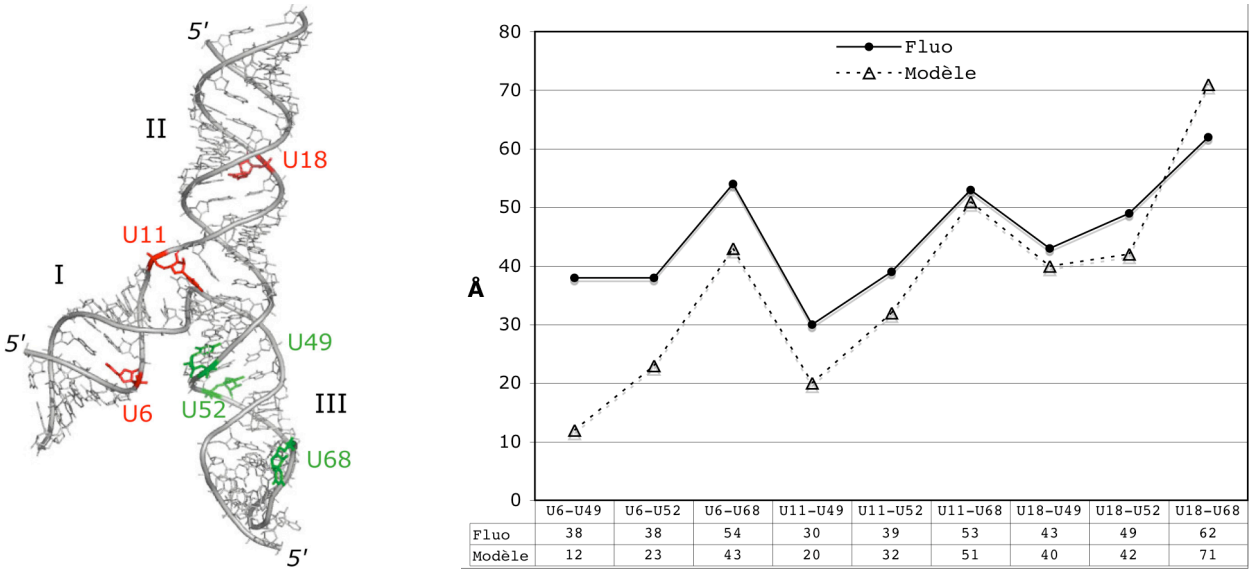


Figure 49: Structure 2D du complexe U4U6. Les fluorophores donneurs et accepteurs sont indiqués respectivement en vert et rouge.



	U6–U49	U6–U52	U6–U68	U11–U49	U11–U52	U11–U68	U18–U49	U18–U52	U18–U68
Fluo	38	38	54	30	39	53	43	49	62
Modèle	12	23	43	20	32	51	40	42	71

Figure 50 : Modèle 3D de la jonction triple du complexe U4U6 (à gauche) et graphique des distances entre fluorophores mesurées expérimentalement (Fluo) et sur le modèle (Modèle).

3. CONCLUSIONS ET PERSPECTIVES

Un des objectifs de ce travail de thèse a été d'extraire, à partir de comparaison de séquences et d'analyse de structures tridimensionnelles, des contraintes structurales qui régissent l'architecture d'un ARN. Dans un premier temps, nous avons cerné et proposé les définitions des notions qui seraient utlisées tout au long de ce travail de thèse, c'est-à-dire précisé un vocabulaire spécifique à la structure d'ARN. Pour commencer, nous nous sommes intéressés aux différentes définitions de la notion de "motif ARN" trouvées dans la littérature et nous avons rappelé la définition du motif ARN utilisée au laboratoire (Leontis & Westhof, 2003) : une succession ordonnée de paires de bases non Watson-Crick sous contraintes. Les motifs ARN, tout comme les jonctions, sont responsables de l'organisation dans l'espace des hélices formées de paires Watson-Crick. L'hélice est l'élément majeur de structure secondaire. Les hélices sont régulières et ce sont les motifs tridimensionnels et les jonctions qui apportent l'irrégularité et la compaction nécessaires à l'obtention de structures originales et fonctionnelles. Pour faciliter l'extraction de l'information de structures tridimensionnelles connues d'ARN, nous avons proposé une représentation des réseaux d'interactions. Nous avons illustré dans ce travail l'utilité de ces représentations en montrant l'exploitation des diagrammes de quelques structures cristallographiques d'ARN. Une rapide lecture de ces diagrammes de réseaux d'interactions nous a permis de mettre en évidence des similitudes et des différences qui pourraient être importantes pour les mécanismes moléculaires dans lesquels ces ARN sont impliqués. Nous pensons que les structures résolues d'ARN n'ont pas encore révélé tous leurs secrets et qu'une analyse approfondie est indispensable pour la détermination des règles générales de repliement de l'ARN.

Des motifs d'interactions très spécifiques sont présents dans les ARN de taille moyenne comme les introns de groupe I, mais assez peu nombreux dans les ARN ribosomiques. Par exemple, il y a deux récepteurs à 11 nucléotides dans l'intron de *Azoarcus* mais aucun dans l'ARN ribosomique 16S de *Thermus thermophilus*. Les ARN ribosomiques comportent par contre de très nombreuses interactions en A mineur et en particulier des motifs typeI/typeII. L'identification

de toutes les interactions en A mineur des ARN ribosomiques dont on connait la structure, mais également des interactions qui assurent un lien entre la grande et la petite sous-unité ribosomique, est un travail important qui reste à réaliser. En effet, ces interactions pourraient jouer un rôle très important dans la dynamique du ribosome lors des différentes étapes du processus de traduction. H. Noller, dans un article récent (2005), propose le recensement des interactions en A mineur de l'ARN ribosomique 16S. Il souligne, en particulier, le nombre important d'adénines dans une géométrie favorable à la formation d'interactions en A mineur mais qui ne réalisent pas complètement ces interactions dans la structure cristallographique. Ces adénines, par un léger rapprochement des éléments qui les portent, sont susceptibles de former des interactions à longue distance et de participer ainsi au changement conformationnel subi par les sous-unités ribosomiques au cours de la traduction.

Nous avons vu que les motifs en A mineur sont peu spécifiques et que leur formation est assurée par le positionnement dans l'espace des éléments structuraux qui les portent. Ce sont les motifs tridimensionnels et les jonctions à plusieurs hélices qui permettent ce positionnement et donc les interactions à longue distance. Au vu du rôle important que jouent les motifs sur l'architecture globale des ARN, il est important de pouvoir les identifier dans des séquences ARN de structure inconnue. Le caractère récurrent des motifs et la succession caractéristique de paires non Watson-Crick propre à chacun permettent, au sein d'un alignement de séquences homologues et à l'aide des matrices d'isostérie (Figure 51 voie 1), de déterminer leur séquence signature (Figure 51 voie 2). Au cours de ce travail de thèse, nous avons déterminé, comme cela avait été fait précédemment pour le motif boucle E (Leontis et al., 2002a), les séquences signatures du motif tournant K et du motif C. L'exemple de la détermination théorique de la séquence signature du motif en A mineur montre que lorsque le motif a suffisamment de contraintes, il est possible de proposer une séquence signature sans passer par les calculs de covariations au sein d'un alignement (Figure 51 voie 3). La détermination, de l'une ou l'autre manière, des séquences signatures de motifs permettra d'identifier dans un alignement de séquences homologues d'ARN de structure tridimensionnelle inconnue les motifs correspondants (Figure 51 voie 4). Cette étape nécessitera la détermination initiale de la structure secondaire de l'ARN par la méthode de choix que constitue l'analyse des covariations dans l'alignement de séquences. Dans le cas où une

seule séquence est disponible, la détermination de la structure secondaire pourra se faire par minimisation de l'énergie libre accompagnée d'analyses de type sondes enzymatiques et chimiques, et les régions en simple brin pourront être examinées afin d'identifier, à l'aide des séquences signatures, la présence d'un motif ou d'un autre (Figure 51 voie 5).

Nous avons vu que certains motifs sont responsables d'interactions à longue distance. Dans de nombreuses structures cristallographiques de ribozymes, les interactions ARN-ARN entre éléments à la périphérie du coeur de la structure jouent un rôle important pour l'activité ribozymique. La modélisation moléculaire des interactions entre les boucles 1 et 2 du ribozyme à tête de marteau a permis de confirmer les données en solution montrant une augmentation de 100 fois de l'activité du ribozyme en présence des éléments périphériques au coeur catalytique. Différentes études montrent que le mouvement des hélices I et II, dont les interactions à longue distance pourraient être responsables, aurait une influence sur l'activité du ribozyme. Ainsi, pour confirmer la nature de ces interactions et comprendre mieux l'activité ribozymique, la structure cristallographique du ribozyme à tête de marteau natif devra être déterminée. C'est dans ce but que les essais de cristallisation du ribozyme à tête de marteau natif devraient être repris au laboratoire.
L'histoire montre qu'après la découverte du ribozyme à tête de marteau, le premier objectif a été de définir la structure minimale qui conservait une activité d'autocoupure. Ce n'est que tout récemment que l'on s'est interrogé sur le rôle des régions périphériques et que leur importance dans l'activité physiologique d'autocoupure a été démontrée. Cet exemple illustre la simplification réductionniste des systèmes peut être source d'incertitude et d'erreur, et devrait servir de leçon sur la manière d'appréhender les systèmes mécanistiques moléculaires.

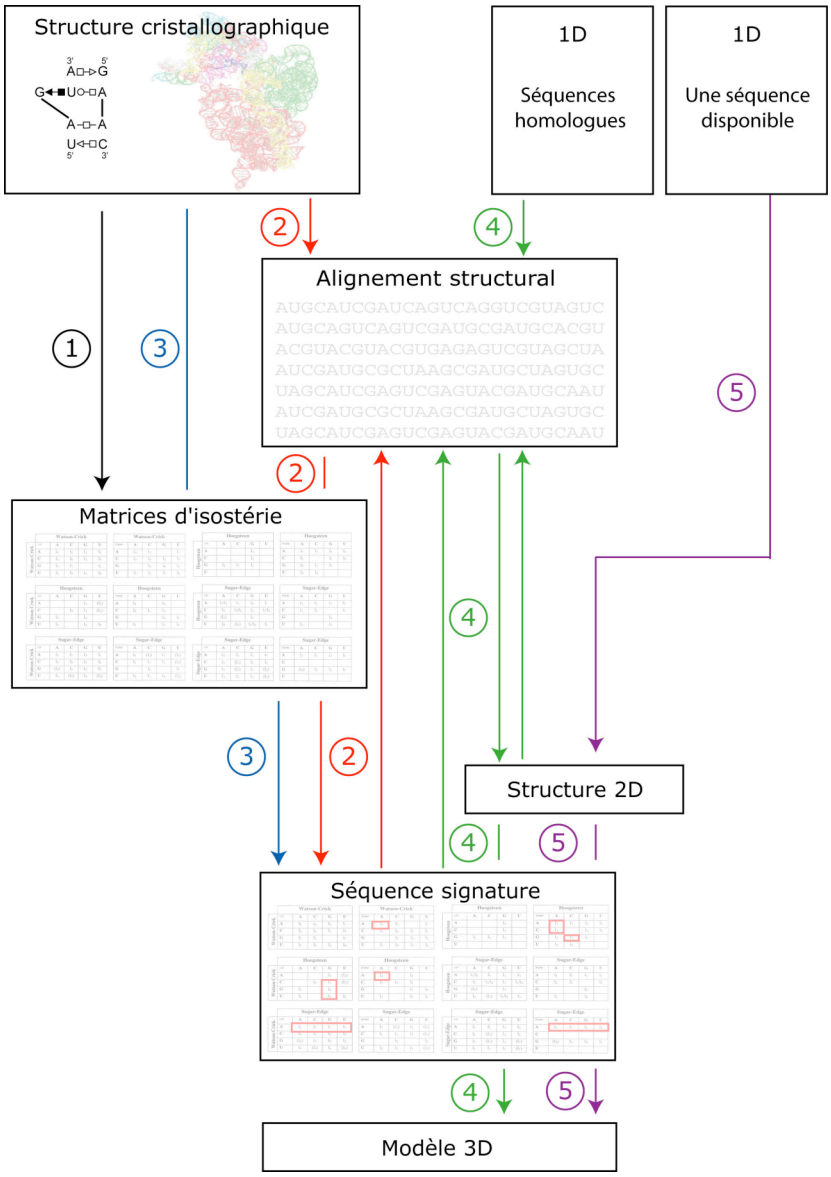


Figure 51 : Illustration de l'utilisation des matrices d'isotérie pour l'intégration des informations de séquence et de structure 3D dans le but de produire des alignements précis et des modèles 3D basés sur la séquence. En premier lieu, les matrices d'isostérie des paires de bases non Watson-Crick classées en 12 familles géométriques ont été proposées sur la base de l'analyse des structures atomiques à haute résolution comme indiqué dans la voie 1. L'analyse des molécules d'ARN homologues présentant le même motif a permis, à l'aide des matrices d'isostérie, de proposer des séquences signatures de plusieurs motifs (voie 2). Les séquences signature de motifs de structures cristallographiques, présentant des contraintes structurales fortes, peuvent être déterminées directement par l'intermédiaire des matrices d'isostérie comme nous l'avons fait pour le motif en A mineur (voie 3). Pour des familles des molécules homologues d'ARN (voie 4) pour lesquelles aucune structure 3D n'existe, les covariations Watson-Crick et la minimisation d'énergie (voie 5) peuvent être utilisées pour déterminer les structures 2D communes qui définissent les régions en simple brin porteuses de motif tridimensionnel. Les signatures des motifs connus sont employées pour identifier des motifs et pour affiner l'alignements des régions présentant des motifs selon un processus itératif (voies 4 et 5). Les substitutions de motifs peuvent être identifiées dans les alignements.

Tout comme les motifs intrahélicaux, les jonctions à plusieurs hélices peuvent intervenir dans le positionnement dans l'espace d'éléments qui doivent interagir. Le rôle structural de la jonction à trois hélices qui constitue le coeur catalytique du ribozyme à tête de marteau est apparu déterminant lors de la modélisation. En effet, l'orientation des hélices I et II était cruciale pour la formation des interactions ARN-ARN entre les boucles 1 et 2. Comme nous l'avons vu, de nombreuses études ont montré qu'une modification de cette orientation pourrait favoriser la coupure ou la ligature du ribozyme. Dans la structure du ribozyme, seule la jonction à trois hélices peut assurer l'orientation des hélices I et II.

Les jonctions à trois hélices sont nombreuses dans les ARN structurés mais contrairement aux jonctions à quatre hélices, peu de choses sont connues à leur sujet. Ce manque d'information nous a posé problème au moment du choix de topologie pour une jonction à trois hélices que nous devions modéliser. Afin de mieux comprendre leur structuration et ainsi faciliter leur modélisation, nous avons choisi d'étudier leur structure. La récolte et l'analyse des jonctions triples des structures cristallographiques d'ARN nous ont permis de les classer en trois familles de jonctions (A, B et C) en fonction de la taille de leurs simples brins jonctions. Nous avons pu déterminer pour les familles A et C une structure consensus qui révèle les contraintes imposées par des appariements non Watson-Crick spécifiques au coeur de la jonction. Finalement, nous proposons des règles structurales qui permettront, à l'avenir, de modéliser une jonction triple en fonction d'éléments de séquence nucléotidique et de structure secondaire.

L'établissement des interactions ARN-ARN dépend donc des conformations imposées par les motifs tridimensionnels et les jonctions. Nous avons vu que les modifications architecturales des sous-unités ribosomiques lors de la traduction pouvaient être dues à l'alternance de formation et rupture d'interactions en A mineur. Les éléments structuraux portant les motifs d'interactions ARN-ARN doivent donc montrer une certaine dynamique mais contrairement aux interrupteurs moléculaires qui demandent un changement de structure secondaire, les modifications des régions qui créent un coude ou des points de jonction de plusieurs hélices demanderaient un changement de la seule structure

tridimensionnelle. Un changement conformationnel d'un motif ou d'une jonction pour favoriser une nouvelle interaction est donc envisageable.

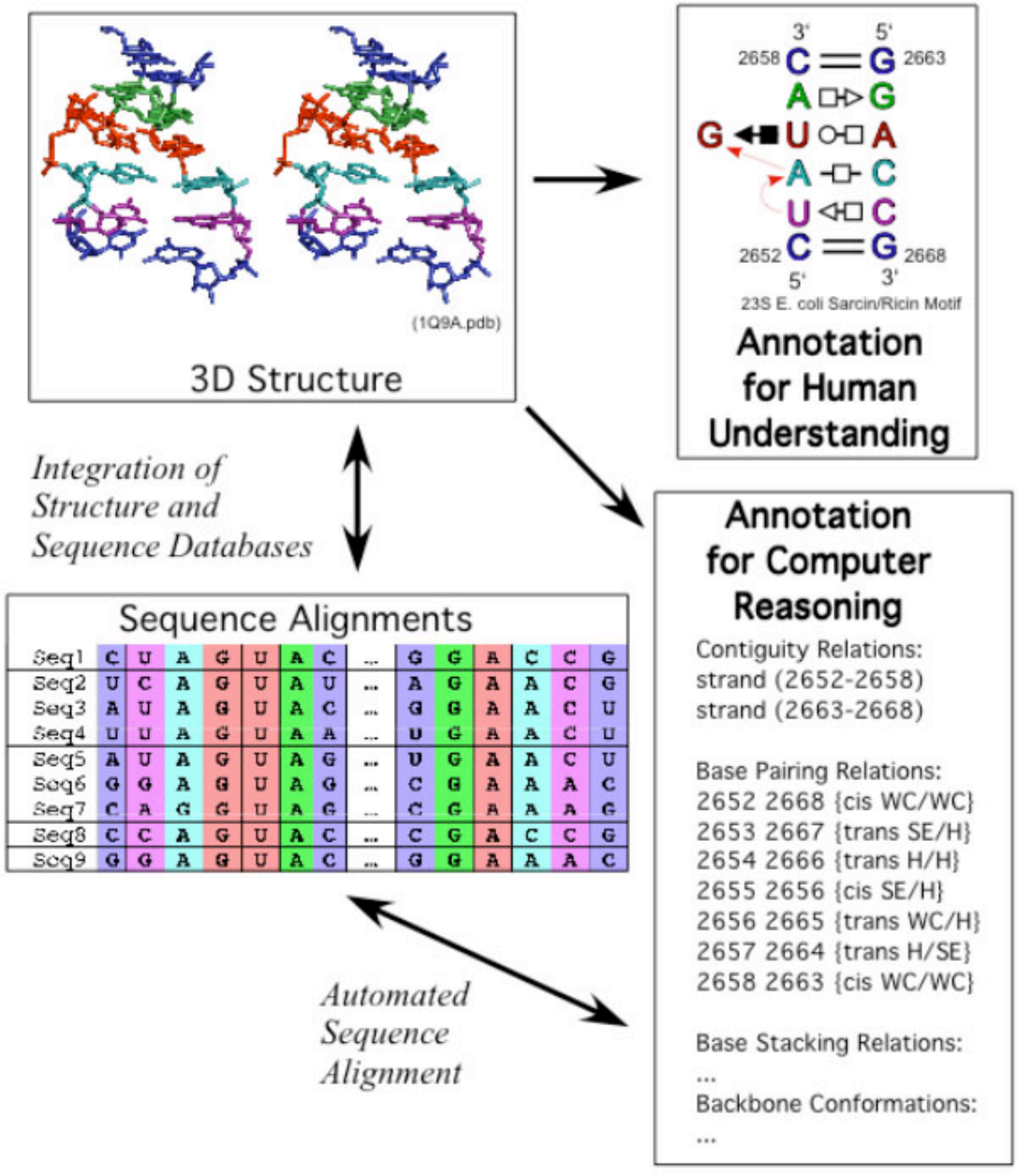


Figure 52: L'ontologie facilitera l'intégration des données hétérogènes d'ARN comprenant les structures 3D expérimentales (en haut à gauche) et les séquences nucléotidiques (à gauche en bas). Les informations sur l'ARN seront compréhensibles par des humains (en haut à droite) et des machines (en bas à droite). L'exemple présenté est celui du motif boucle sarcine. Schéma extrait de l'article sous presse de Leontis et collaborateurs (Leontis et al., 2006).

Au cours de ce travail, les difficultés de communication entre les différents scientifiques travaillant sur l'ARN sont apparues nombreuses et entraînent confusion et perte de temps. La raison principale est que les notions de base de la structure de l'ARN sur lesquelles reposent les études réalisées ne sont pas appréhendées de la même manière par tous (voir chapitre 2.2). Afin de pallier à ces manques, il a été créé en 2005 un consortium d'ontologie de l'ARN qui a pour objectif la création d'un cadre conceptuel de travail sur l'ARN avec un vocabulaire commun, dynamique, contrôlé et structuré pour décrire et caractériser les séquences nucléotidiques, les structures secondaires et tridimensionnelles et les dynamiques permettant aux ARN d'être fonctionnels (Leontis et al., 2006). Pour cela, des outils adaptables à l'informatique et permettant la description de l'architecture, de la fonction et de l'évolution de l'ARN devront être créés et capables d'inter-opérer. L'objectif le plus urgent est de définir, d'identifier et de classer, dans un langage expoitable en informatique, les motifs structuraux décrits dans la littérature ou apparaissant dans les bases de données. Pour atteindre cet objectif, des réunions de travail sont organisées entre les membres de la "communauté ARN" pour discuter, débattre et enfin trouver des solutions aux difficultés conceptuelles rencontrées. Il est intéressant de noter que pour atteindre ce premier objectif, deux équipes travaillant sur des visions différentes du motif ARN (succession d'appariements et conformations du squelette sucre-phosphate) en utilisant le même langage, pourront joindre leurs efforts et exploiter les deux notions pour décrire de la manière la plus complète possible les structures tridimensionnelles d'ARN et les relier aux séquences nucléotidiques (Figure 52).

La connaissance des règles structurales reliant les séquences d'ARN à leurs architectures devrait permettre dans l'avenir de mieux comprendre l'évolution moléculaire de ces molécules ubiquitaires en biologie et au delà, l'évolution biologique au sens plus large.

4. ANNEXES

4.1. Matériel et méthodes

4.1.1. Analyse comparative de séquence

L'approche comparative est la méthode de choix de détermination de structure secondaire d'un ARN quand des molécules homologues, de fonction identique et contenant des variations suffisantes, sont disponibles. Elle est basée sur une caractéristique fondamentale des ARN structurés : la robustesse. En effet, au cours de l'évolution, les éléments importants des structures secondaires et tertiaires d'un ARN évoluent moins vite que la séquence nucléotidique. Tout changement dans la séquence qui peut perturber la structure est compensé par un autre changement ; la structure est ainsi restaurée. Des séquences homologues présentent ainsi des changements compensatoires ou covariations qui permettent la conservation de la structure 2D et ensuite tridimensionnelle active. La structure d'un ARN peut être donc déterminée par identification des covariations au sein de l'alignement structural des séquences homologues. Un alignement structural implique la détermination des régions appariées le long de chacune des séquences de l'alignement puis leur arrangement au niveau horizontal de manière que les régions formant des paires Watson-Crick se juxtaposent verticalement. Ensuite, les bases impliquées dans des appariements Watson-Circk sont alignées verticalement ce qui nécessite parfois l'insertion de blancs. La solidité de l'approche augmente avec la diversité des séquences et la distance évolutive qui les sépare tandis que l'exactitude des prédictions dépend du nombre d'évènements de covariations observées (Michel et al., 2000). Les paires de bases Watson-Crick conservées doivent être considérée avec précaution car l'absence de covariation révèle souvent un autre type d'appariement.

Les analyses phylogénétiques peuvent être utilisées également pour la prédiction d'interactions tertiaires. Levitt a ainsi pu prédire dès 1969, c'est à dire cinq ans avant la structure cristallographique, une des paires non canoniques et une triple de la structure de l'ARNt (Levitt, 1969). De la même manière, l'analyse comparative de séquence a mis en évidence des interactions tertiaires qui ont

été incorporées dans les modèles tridimensionnels de l'ARN de la RNase P ou des introns de groupe I (Michel & Westhof, 1990; Jaeger, 1995; Massire, 1998; Massire et al., 1998) et validées par la suite par les structures cristallographiques correspondantes. Le principe est le même que pour les paires Watson-crick mais les appariements non Watson-Crick obeissent à d'autres règles de covariation (voir introduction). L'analyse comparative comme nous l'avons vu dans l'article 1 est un processus itératif ; l'alignement de séquences homologues donne des informations structurales qui elles-mêmes permettent d'affiner l'alignement. Nous avons utilisé le logiciel Bioedit pour les alignements de séquences .

Une autre manière d'obtenir des informations structurales est l'analyse des mutations dont l'effet est compensé par mutation sur un second site. Cette méthode est intéressante lorsqu'on cherche à confirmer la présence d'une caractéristique particulière de la structure. Elle a été utilisée pour mettre en évidence les interactions tridimensionnelles entre éléments périphériques au coeur catalytique des ribozymes à tête de marteau de STRSV et *Arabidodpsis Thaliana* (voir articles 2 et 3).

L'analyse comparative de séquence est, lorsqu'on dispose de séquences homologues, une étape nécessaire avant la modélisation.

4.1.2. Modélisation moléculaire

La modélisation moléculaire d'ARN consiste à englober les données expérimentales accessibles dans un modèle tridimensionnel. Le modèle atomique tridimensionnel est construit en tirant parti de la stéréochimie des nucléotides et autres contraintes intrinsèques aux ARN comme la structure des hélices (Westhof 1992, 1993).

La modélisation repose sur le caractère modulaire de l'ARN qui permet de penser la construction de molécules structurées complexes comme l'assemblage d'éléments structuraux récurrents. Ainsi, les molécules d'ARN qui présentent des structures 2D différentes, partagent des modules qui adoptent des structures tridimensionnelles analogues.

La construction d'un ARN débute par l'analyse et la découpe de sa structure secondaire en éléments de structure secondaire comme les hélices Watson-Crick, les bulles internes ou les boucles. Les coordonnées des hélices

sont générées à l'aide du logiciel NAHELIX. Les éléments de structure secondaire non-hélicoïdaux sont modélisés avec le logiciel FRAGMENT à partir des structures cristallographiques d'ARN. Les modules ainsi formés sont ensuite assemblés à l'écran graphique avec le programme MANIP (Massire & Westhof, 1998) en tenant compte des données expérimentales comme les résultats de cartographies chimiques et enzymatiques (Ehresmann et al., 1987). Le modèle de structure est alors affiné par moindres-carrés (konnert & hendrickson, 1980) avec le programme NUCLIN/NUCLSQ (Westhof et al., 1985) (Figure 51)

La modélisation moléculaire permet donc de construire des modèles *ab initio* sur la base de données en solution et d'analyse comparative de séquences. De nombreux modèles d'ARN structurés ont été publiés et l'analyse des structures cristallographiques publiées postérieurement valide l'approche comme outil de prédiction de l'architecture globale des molécules d'ARN.

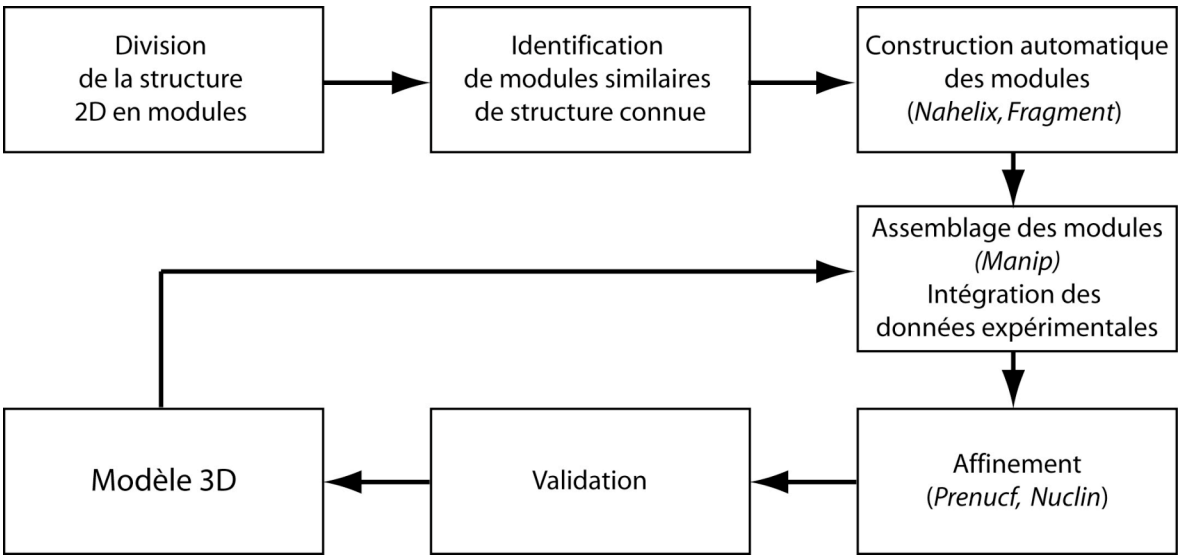


Figure 53: Représentation schématique de la stratégie de modélisation d'ARN. Les modules de structure secondaire identifiés par analys de séquences sont construits automatiquement en 3D. Ces modules sont ensuite assemblés en tenant compte des données expérimentales. Après affinement, le modèle est comparé aux données expérimentales. Il est validé s'il est cohérent ou modifié dans le cas contraire.

4.1.3. Revue 2: "Preparation and handling of RNA Crystals"

En 2004, dans le cadre d'un livre de méthodologie dédié à la biochimie des acides nucléiques, notre laboratoire a rédigé un chapitre traitant de la cristallisation des ARN, étape par étape, de la conception de fragments d'ARN et leur purification jusqu'à la mise en oeuvre de stratégies de cristallisation. J'ai utilisé les techniques décrites dans ce chapitre lors des essais de cristallisation des constructions de ribozyme naturel.

Dans l'objectif des études de cristallogénèse des ribozymes à tête de marteau présentant des éléments périphériques, j'ai identifié les conditions expérimentales utilisées dans la litérature pour la cristallogenèse d'ARN d'une taille à peu près similaire à celle du ribozyme. Ces conditions sont rassemblées dans la table 2 qui suit le chapitre du Handbook.

Handbook of RNA Biochemistry

Preparation and Handling of RNA Crystals

Boris François, **Aurélie Lescoute-Phillips**, Andreas Werner and Benoît Masquida

Pages 438–452 :

Structure	*Domaine P4-P6 intron gp I*	*tRNAi euca*	*tRNA Asp Yeast*	*tRNAi E.Coli*	*HH all-RNA*	*HH DNA-RNA*
Pdb	1GID	1YFG			1MME	1HMH
Nbr nts	160 nts	74 nts	76 nts	76 nts	16nt + 25nt	13dnt + 34nts
Résolution	2.8 Å	4.5 Å	3.5 Å	3.5 Å	3.1 Å	2.6 Å
Synhèse	transcription in vitro		levure		Synthèse chimique	RNA: trans. In vitro DNA: solid-phase synth.
Purification	PAGE 6%		Chromato.			PAGE filtre, concentre, dessale
Renaturation	Supposés : 5 mM MgCl2 10 mM NaCl 5mM HEPES pH7 50°C pdt 5 min.	Élution dans : 1,6 M AS 10 mM NaOAc 10 mM MgCl2 pH 4.5			10 mM NH4Caco pH 6.5	
Méthode de cristallisation	Diffusion vap. (g.susp.)			Diffusion vap.	Diffusion vap. (G.ass..)	
Goutte						
Conc. RNA	Supposés : 3,5 mg/mL				0,5 mM	0,3 mM
Tampon	60 mM KCaco pH 6.0		40 mM NaCaco		50 mM NH4Caco pH 6.5	10 mM NaCaco pH 6.5
Agent précipitant		2 M AS	62% AS	Isopropanol^{-1}-2%		1,9-2,2 AS ou 1,5-2,2 M LiSO4
Sels et additifs	30 mM MgCl2 0,3 mM spermine 0,2^{-1},0 mM CoHex.	5 mM MgCl2 2 mM spermine	2 mM MgCL2	10 mM MgCl2 8 mM spermine 2 mM BaCl2 50 mM AS	2^{-1}0 uL réservoir v:v	0^{-1}00 mM spermine
Réservoir						
Tampon					50 mM NH4Caco pH 6.5	
Agent précipitant	MPD			isopropanol ?	23% PEG 6000	
Sels et additifs					10 mM Mg(OAc)2 100 mM NH4OAc 1 mM spermine 5 % glycérol	
T°C			20		20	4
Référence	1312 ; 790	1977 ; 1978			391	532

Rmq	! Renat conc. RNA d'après 1312 !	Purif et renat d'après 1977		BaCl2 + MgCl2 pas dans matrices	Natrix: cond. 27 et 20	
AS: (NH4)2SO4						
Co.Hex.: Co(NH3)6Cl3						
MOPS: 3N morpholino propane sulfonique acid						
PIPES: piperazine^{-1},4-bis (2-ethane sulfonique acid)						
Structure	*HH DNA-RNA*	*10-23 DNA enzyme*	*Complexe ARN-ADN dérivé de DNA-enz.*	*Malachite G aptamer*	*RNA triplex (1)*	*RNA triplex + atomes lourds (2)*
Pdb	1HMH	1br3	1EGK	1F1T	437D	
Nbr nts	13dnt + 34nts	82 nts	108 nts	38 nts	28 nts	
Résolution	2.6 Å			2.8 Å	1.6 Å	
Synhèse	RNA: trans. In vitro DNA: solid-phase synth.	Synthèse chimique	Synthèse chimique	Transcription in vitro	Transcription in vitro	
Purification	PAGE filtre, concentre, dessale	PAGE dénat. Dialyse contre eau	PAGE dénat. Dialyse contre eau	desal. Sephadex G-50 PAGE 20%	PAGE, extraction, conc. Amicon, dialyse contre eau mQ	
Renaturation		vDNA:vRNA 70°C et refroidissement lent jsq room t°	vDNA:vRNA 70°C et refroidissement lent jsq room t°		300 mM NaCaco (pH6.5) 15 mM Mg 60°C 10min. Refroidissement j. 25°C	
Méthode de cristallisation		Diffusion vap. (g.ass.)	Diffusion vap. (g.ass.)	Diffusion vap. (G.susp..)	Diffusion vap. (g.ass.)	
Goutte						
Conc. RNA	0,3 mM	0,4 mM		0,5 mM +1 mM TMR	0,43 mM	ARN brome
Tampon	10 mM NaCaco pH 6.5	50 mM NaCaco pH 6.5	50 mM NaCaco pH 6.0	13,2 mM NaCaco pH 4.5	100 mM K MOPS pH 7.0	100 mM K PIPES pH 6.5
Agent précipitant	1,9-2,2 AS ou 1,5-2,2 M LiSO4	0,9 M Li2SO4	25% MPD	3,33% (v/v) MPD	5% sec-butanol	5% MPD
Sels et additifs	0^{-1}00 mM spermine	30 mM MgCl2 1 mM spermine	20 mM MgCl2 1 mM spermine	6,66 mM MgCl2 0,165 mM spermine 23,16 mM KCl 39,6 mM SrCl2	5 mM MgCl2 2 mM spermine	5 mM MgCl2 1 mM spermine
Réservoir						

Tampon						
Agent précipitant				35% MPD	18% sec-butanol	
Sels et additifs						
T°C	4	24,5	24,5	Room t°	25°C	
Référence	532	1325		2130	1331	1331
Rmq		Jonctions à 3 hélices cristaux dans mêmes conditions + co.hex. Natrix : cond. 18 et 19	Pas d'info sur conc. d'ARN ; doit etre meme que dna enzyme Natrix cond. 15	Pas ce pH dans Natrix pas trouve cette combinaison	Natrix cond. 31	Seule condition où cristaux apparaissent pas trouve cette combinaison

Structure	*Viral RNA pseudoknot cristal cubique (3)* *Pour cristal trigonal cf 1 et 2*	*Jonction à 4 hélices HCV*	*Rnase P*	*Preliminary X-ray diff. Studies STRSV*
Pdb	Trigo. 1L2X Cubic 1L3D	1KH6	1NBS	
Nbr nts	28 nts	53 nts	153 nts	69 nts
Résolution	2.85 Å	2.8 Å		2.4 Å
Synhèse	Transcription in vitro	Transcription in vitro HDV en 3'	Transcription in vitro	Transcription in vitro
Purification	PAGE, extraction, conc. Amicon, dialyse contre eau mQ	PAGE denat., elution passive, conc.		Filtration Microcon colonne echange d'ions
Renaturation		30 mM HEPES-KOH pH7.5 65°C 1min. 5min. T° ambiante 2,5 mM MgCl2 1 mM spermidine		
Méthode de cristallisation	diffusion vap. (g.ass.)	diffusion vap. (g.ass.)	diffusion vap. (g.susp.)	diffusion vap.
Goutte				
conc. RNA	0,3 mM	0,3 mM 1v pour 2v de solution de xtalli.	0,1 mM	0,5 mM RNA

tampon	0,33 M citrate de sodium pH 5.0	0,1M citrate de sodium pH5.6	v:v reservoir	35 mM NaCaco pH 6.5
agent precipitant	1M AS	30% MPD	"	6% PEG 400
sels et additifs		0,2 M NH4OAC	"	28 mM MgCl2 18,9 mM spermine
Reservoir				
tampon			50 mM MES pH 6.0	0,1 M NaCaco pH 6.5
agent precipitant	2M AS		10,5% isopropanol	18% PEG 400
sels et additifs			5 mM MgCl2 1,5 mM spermine 80 mM SrCl2	80 mM MgCl2 18 mM spermine
T°C	4°C	30°C	30°C	room t°
Référence	2001	1778		
Rmq	Natrix : pas de citrate	Natrix : pas de citrate	Natrix : cond. 10	problèmes de repro. des cristaux pas de cristaux avec atomes lourds Natrix cond. 22

Structure	*etudes preli. vit. B12 binding RNA pseudoknot cristx A*	*cristx B*	*cristx B*	*etudes preliminaires biotin-binding RNA pseudonoeud*	
Pdb					
Nbr nts	35 nts			36 nts	
Résolution (Å)	2.9 Å	2.9 Å		2.8 Å	
Synhèse	Transcription in vitro			Transcription in vitro	
Purification	PAGE 15%			PAGE 20%	

Renaturation	B12-binding buffer:1 M LiCl ; 5 mM Na-HEPES pH 7.4			binding buffer : 0,1 M KCl ; 10 mM Na-HEPES pH 7.4 ; 5 mM MgCl2	
Méthode de cristallisation	diffusion vap. (g.susp.)			diffusion vap. (g.susp.)	
Goutte					
conc. RNA	0,1 mM			0,4-0,7 mM	
tampon	20 mM K Caco pH 6.0	20 mM K-PIPES pH 6.5	20 mM K-MOPS pH 7.0	rapport arn:reservoir 1:2 ; 2:2 ; 2:1	
agent precipitant	7% 2-propanol	7% 2-propanol	20% dioxane	10	
sels et additifs	5 mM MgCl2	15 mM MgCl2 0,5 mM BaCl2	5 mM MgCl2 ; 1 mM spermine 1 M LiCl	"	
Reservoir					
tampon				20 mM K-HEPES pH 7.5	20 mM K Caco pH 6.5
agent precipitant				4% PEG 8000	5% PEG 4000
sels et additifs	10 M LiCl			20 mM MgCl2 1 mM spermine	15 mM MgCl2 ; 1,5 mM spermine ; 2 mM CoHex.
T°C				4	
Référence					
Rmq	pas d'atomes lourds Natrix cond. 10	pas d'atomes lourds	pas d'atomes lourds	3 mois pour cristaux mauvaise reproduc. Natrix cond. 40 ; 42	

4.2. Revue 3 : "Riboswitch structures: purine ligands replace tertiary contacts"

Aurélie Lescoute & Eric Westhof

Chem Biol. 2005 Jan;12(1):10-3.

Riboswitch structures: purine ligands replace tertiary contacts

Aurélie Lescoute & Eric Westhof

Chemistry & Biology, 2005, Vol.12, N°1, Pages 10-13

5. BIBLIOGRAPHIE

Adams PL, Stahley MR, Gill ML, Kosek AB, Wang J, Strobel SA. 2004a. Crystal structure of a group I intron splicing intermediate. *Rna 10*:1867-1887.

Adams PL, Stahley MR, Kosek AB, Wang J, Strobel SA. 2004b. Crystal structure of a self-splicing group I intron with both exons. *Nature 430*:45-50.

Auffinger P, Louise-May S, Westhof E. 1999. Molecular dynamics simulations of solvated yeast tRNA(Asp). *Biophys J 76*:50-64.

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.

Bassi GS, Mollegaard NE, Murchie AI, von Kitzing E, Lilley DM. 1995. Ionic interactions and the global conformations of the hammerhead ribozyme. *Nat Struct Biol 2*:45-55.

Bassi GS, Murchie AI, Walter F, Clegg RM, Lilley DM. 1997. Ion-induced folding of the hammerhead ribozyme: a fluorescence resonance energy transfer study. *Embo J 16*:7481-7489.

Batey RT, Gilbert SD, Montange RK. 2004. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature 432*:411-415.

Battle DJ, Doudna JA. 2002. Specificity of RNA-RNA helix recognition. *Proc Natl Acad Sci U S A 99*:11676-11681.

Beaudry AA, Joyce GF. 1990. Minimum secondary structure requirements for catalytic activity of a self-splicing group I intron. *Biochemistry 29*:6534-6539.

Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T, Hsieh SH, Srinivasan AR, Schneider B. 1992. The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys J 63*:751-759.

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

Berzal-Herranz A, Joseph S, Chowrira BM, Butcher SE, Burke JM. 1993. Essential nucleotide sequences and secondary structure elements of the hairpin ribozyme. *Embo J 12*:2567-2573.

Blount KF, Uhlenbeck OC. 2002. Internal equilibrium of the hammerhead ribozyme is altered by the length of certain covalent cross-links. *Biochemistry 41*:6834-6841.

Blount KF, Uhlenbeck OC. 2005. The structure-function dilemma of the hammerhead ribozyme. *Annu Rev Biophys Biomol Struct 34*:415-440.

Brion P, Westhof E. 1997. Hierarchy and dynamics of RNA folding. *Annu Rev Biophys Biomol Struct 26*:113-137.

Brodersen DE, Clemons WM, Jr., Carter AP, Wimberly BT, Ramakrishnan V. 2002. Crystal structure of the 30 S ribosomal subunit from Thermus thermophilus: structure of the proteins and their interactions with 16 S RNA. *J Mol Biol 316*:725-768.

Brown JW, Haas ES, Pace NR. 1993. Characterization of ribonuclease P RNAs from thermophilic bacteria. *Nucleic Acids Res 21*:671-679.

Burke DH, Greathouse ST. 2005. Low-magnesium, trans-cleavage activity by type III, tertiary stabilized hammerhead ribozymes with stem 1 discontinuities. *BMC Biochem 6*:14.
Burke JM, Belfort M, Cech TR, Davies RW, Schweyen RJ, Shub DA, Szostak JW, Tabak HF. 1987. Structural conventions for group I introns. *Nucleic Acids Res 15*:7217-7221.
Canny MD, Jucker FM, Kellogg E, Khvorova A, Jayasena SD, Pardi A. 2004. Fast cleavage kinetics of a natural hammerhead ribozyme. *J Am Chem Soc 126*:10848-10849.
Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Kundrot CE, Cech TR, Doudna JA. 1996a. Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science 273*:1678-1685.
Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Szewczak AA, Kundrot CE, Cech TR, Doudna JA. 1996b. RNA tertiary structure mediation by adenosine platforms. *Science 273*:1696-1699.
Cech TR, Damberger SH, Gutell RR. 1994. Representation of the secondary and tertiary structure of group I introns. *Nat Struct Biol 1*:273-280.
Conn GL, Draper DE, Lattman EE, Gittis AG. 1999. Crystal structure of a conserved ribosomal protein-RNA complex. *Science 284*:1171-1174.
Costa M, Michel F. 1995. Frequent use of the same tertiary motif by self-folding RNAs. *Embo J 14*:1276-1285.
Costa M, Michel F. 1997. Rules for RNA recognition of GNRA tetraloops deduced by in vitro selection: comparison with in vivo evolution. *Embo J 16*:3289-3302.
Cote F, Levesque D, Perreault JP. 2001. Natural 2',5'-phosphodiester bonds found at the ligation sites of peach latent mosaic viroid. *J Virol 75*:19-25.
Cote F, Perreault JP. 1997. Peach latent mosaic viroid is locked by a 2',5'-phosphodiester bond produced by in vitro self-ligation. *J Mol Biol 273*:533-543.
De La Pena M, Gago S, Flores R. 2003. Peripheral regions of natural hammerhead ribozymes greatly increase their self-cleavage activity. *Embo J 22*:5561-5570.
Doudna JA, Cech TR. 2002. The chemical repertoire of natural ribozymes. *Nature 418*:222-228.
Doudna JA, Szostak JW. 1989a. Miniribozymes, small derivatives of the sunY intron, are catalytically active. *Mol Cell Biol 9*:5480-5483.
Doudna JA, Szostak JW. 1989b. RNA-catalysed synthesis of complementary-strand RNA. *Nature 339*:519-522.
Duarte CM, Pyle AM. 1998. Stepping through an RNA structure: A novel approach to conformational analysis. *J Mol Biol 284*:1465-1478.
Dunham CM, Murray JB, Scott WG. 2003. A helical twist-induced conformational switch activates cleavage in the hammerhead ribozyme. *J Mol Biol 332*:327-336.
Ehresmann C, Baudin F, Mougel M, Romby P, Ebel JP, Ehresmann B. 1987. Probing the structure of RNAs in solution. *Nucleic Acids Res 15*:9109-9128.
Engelhardt MA, Doherty EA, Knitt DS, Doudna JA, Herschlag D. 2000. The P5abc peripheral element facilitates preorganization of the tetrahymena group I ribozyme for catalysis. *Biochemistry 39*:2639-2651.
Fedor MJ. 1999. Tertiary structure stabilization promotes hairpin ribozyme ligation. *Biochemistry 38*:11040-11050.
Ferbeyre G, Smith JM, Cedergren R. 1998. Schistosome satellite DNA encodes active hammerhead ribozymes. *Mol Cell Biol 18*:3880-3888.

Ferre-D'Amare AR, Zhou K, Doudna JA. 1998. Crystal structure of a hepatitis delta virus ribozyme. *Nature 395*:567-574.
Forster AC, Davies C, Sheldon CC, Jeffries AC, Symons RH. 1988. Self-cleaving viroid and newt RNAs may only be active as dimers. *Nature 334*:265-267.
Forster AC, Symons RH. 1987. Self-cleavage of virusoid RNA is performed by the proposed 55-nucleotide active site. *Cell 50*:9-16.
Francois B, Russell RJ, Murray JB, Aboul-ela F, Masquida B, Vicens Q, Westhof E. 2005. Crystal structures of complexes between aminoglycosides and decoding A site oligonucleotides: role of the number of rings and positive charges in the specific binding leading to miscoding. *Nucleic Acids Res 33*:5677-5690.
Garrett TA, Pabon-Pena LM, Gokaldas N, Epstein LM. 1996. Novel requirements in peripheral structures of the extended satellite 2 hammerhead. *Rna 2*:699-706.
Golden BL, Kim H, Chase E. 2005. Crystal structure of a phage Twort group I ribozyme-product complex. *Nat Struct Mol Biol 12*:82-89.
Guo F, Gooding AR, Cech TR. 2004. Structure of the Tetrahymena ribozyme: base triple sandwich and metal ion at the active site. *Mol Cell 16*:351-362.
Haas ES, Armbruster DW, Vucson BM, Daniels CJ, Brown JW. 1996a. Comparative analysis of ribonuclease P RNA structure in Archaea. *Nucleic Acids Res 24*:1252-1259.
Haas ES, Banta AB, Harris JK, Pace NR, Brown JW. 1996b. Structure and evolution of ribonuclease P RNA in Gram-positive bacteria. *Nucleic Acids Res 24*:4775-4782.
Haas ES, Brown JW, Pitulle C, Pace NR. 1994. Further perspective on the catalytic core and secondary structure of ribonuclease P RNA. *Proc Natl Acad Sci U S A 91*:2527-2531.
Haas ES, Morse DP, Brown JW, Schmidt FJ, Pace NR. 1991. Long-range structure in ribonuclease P RNA. *Science 254*:853-856.
Hammann C, Norman DG, Lilley DM. 2001. Dissection of the ion-induced folding of the hammerhead ribozyme using 19F NMR. *Proc Natl Acad Sci U S A 98*:5503-5508.
Heckman JE, Lambert D, Burke JM. 2005. Photocrosslinking detects a compact, active structure of the hammerhead ribozyme. *Biochemistry 44*:4148-4156.
Hershkovitz E, Tannenbaum E, Howerton SB, Sheth A, Tannenbaum A, Williams LD. 2003. Automated identification of RNA conformational motifs: theory and application to the HM LSU 23S rRNA. *Nucleic Acids Res 31*:6249-6257.
Hertel KJ, Herschlag D, Uhlenbeck OC. 1994. A kinetic and thermodynamic framework for the hammerhead ribozyme reaction. *Biochemistry 33*:3374-3385.
Hertel KJ, Uhlenbeck OC. 1995. The internal equilibrium of the hammerhead ribozyme reaction. *Biochemistry 34*:1744-1749.
Holbrook SR. 2005. RNA structure: the long and the short of it. *Curr Opin Struct Biol 15*:302-308.
Huang HC, Nagaswamy U, Fox GE. 2005. The application of cluster analysis in the intercomparison of loop structures in RNA. *Rna 11*:412-423.
Hutchins CJ, Rathjen PD, Forster AC, Symons RH. 1986. Self-cleavage of plus and minus RNA transcripts of avocado sunblotch viroid. *Nucleic Acids Res 14*:3627-3640.

Jaeger L. 1995. Les introns auto-catalytiques de groupe I comme modèle d'étude du repliement des acides ribonucléiques. Strasbourg: Université Louis Pasteur. pp 245.
Jaeger L. 1997. The New World of ribozymes. *Curr Opin Struct Biol 7*:324-335.
Jaeger L, Michel F, Westhof E. 1994. Involvement of a GNRA tetraloop in long-range RNA tertiary interactions. *J Mol Biol 236*:1271-1276.
Jaeger L, Westhof E, Michel F. 1991. Function of P11, a tertiary base pairing in self-splicing introns of subgroup IA. *J Mol Biol 221*:1153-1164.
Johnson TH, Tijerina P, Chadee AB, Herschlag D, Russell R. 2005. Structural specificity conferred by a group I RNA peripheral element. *Proc Natl Acad Sci U S A 102*:10176-10181.
Jossinet F, Westhof E. 2005. Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. *Bioinformatics 21*:3320-3321.
Joyce GF, van der Horst G, Inoue T. 1989. Catalytic activity is retained in the Tetrahymena group I intron despite removal of the large extension of element P5. *Nucleic Acids Res 17*:7879-7889.
Jucker FM, Pardi A. 1995. GNRA tetraloops make a U-turn. *Rna 1*:219-222.
Kiberstis PA, Haseloff J, Zimmern D. 1985. 2' phosphomonoester, 3'-5' phosphodiester bond at a unique site in a circular viral RNA. *Embo J 4*:817-822.
Klein DJ, Moore PB, Steitz TA. 2004. The contribution of metal ions to the structural stability of the large ribosomal subunit. *Rna 10*:1366-1379.
Klein DJ, Schmeing TM, Moore PB, Steitz TA. 2001. The kink-turn: a new RNA secondary structure motif. *Embo J 20*:4214-4221.
konnert j, hendrickson w. 1980. Restrained parametersthermal factors refinement procedures. *Acta Crystallographica A36*:344-349.
Krasilnikov AS, Mondragon A. 2003. On the occurrence of the T-loop RNA folding motif in large RNA molecules. *Rna 9*:640-643.
Krasilnikov AS, Xiao Y, Pan T, Mondragon A. 2004. Basis for structural diversity in homologous RNAs. *Science 306*:104-107.
Krasilnikov AS, Yang X, Pan T, Mondragon A. 2003. Crystal structure of the specificity domain of ribonuclease P. *Nature 421*:760-764.
Laserson U, Gan HH, Schlick T. 2005. Predicting candidate genomic sequences that correspond to synthetic functional RNA motifs. *Nucleic Acids Res 33*:6057-6069.
Lehnert V, Jaeger L, Michel F, Westhof E. 1996. New loop-loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3D model of the Tetrahymena thermophila ribozyme. *Chem Biol 3*:993-1009.
Leontis N, Altman R, Berman H, Brenner SE, Brow J, Engelke D, Harvey S, Holbrook SR, Jossinet F, Lewis S, Major F, Mathews D, Richardson J, Williamson JR, Westhof E. 2006. The RNA Ontologie Consortium: An Open invitation to the RNA Community. *RNA*.
Leontis NB, Stombaugh J, Westhof E. 2002a. Motif prediction in ribosomal RNAs Lessons and prospects for automated motif prediction in homologous RNA molecules. *Biochimie 84*:961-973.
Leontis NB, Stombaugh J, Westhof E. 2002b. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res 30*:3497-3531.
Leontis NB, Westhof E. 1998. The 5S rRNA loop E: chemical probing and phylogenetic data versus crystal structure. *Rna 4*:1134-1153.

Leontis NB, Westhof E. 2001. Geometric nomenclature and classification of RNA base pairs. *Rna 7*:499-512.
Leontis NB, Westhof E. 2003. Analysis of RNA motifs. *Curr Opin Struct Biol 13*:300-308.
Levitt M. 1969. Detailed molecular model for transfer ribonucleic acid. *Nature 224*:759-763.
Mandal M, Breaker RR. 2004. Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nat Struct Mol Biol 11*:29-35.
Mandal M, Lee M, Barrick JE, Weinberg Z, Emilsson GM, Ruzzo WL, Breaker RR. 2004. A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science 306*:275-279.
Massire C. 1998. Développement de logiciels d'aide à la modélisation d'ARN. Application à la composante ribonucléique de la ribonucléase P bactérienne. Strasbourg: Université Louis Pasteur.
Massire C, Jaeger L, Westhof E. 1998. Derivation of the three-dimensional architecture of bacterial ribonuclease P RNAs from comparative sequence analysis. *J Mol Biol 279*:773-793.
Massire C, Westhof E. 1998. MANIP: an interactive tool for modelling RNA. *J Mol Graph Model 16*:197-205, 255-197.
Michel F, Costa M, Massire C, Westhof E. 2000. Modeling RNA tertiary structure from patterns of sequence variation. *Methods Enzymol 317*:491-510.
Michel F, Jacquier A, Dujon B. 1982. Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. *Biochimie 64*:867-881.
Michel F, Westhof E. 1990. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J Mol Biol 216*:585-610.
Miller WA, Silver SL. 1991. Alternative tertiary structure attenuates self-cleavage of the ribozyme in the satellite RNA of barley yellow dwarf virus. *Nucleic Acids Res 19*:5313-5320.
Moore PB. 1999. Structural motifs in RNA. *Annu Rev Biochem 68*:287-300.
Moras D, Comarmond MB, Fischer J, Weiss R, Thierry JC, Ebel JP, Giege R. 1980. Crystal structure of yeast tRNAAsp. *Nature 288*:669-674.
Murchie AI, Thomson JB, Walter F, Lilley DM. 1998. Folding of the hairpin ribozyme in its natural conformation achieves close physical proximity of the loops. *Mol Cell 1*:873-881.
Murphy FL, Cech TR. 1993. An independently folding domain of RNA tertiary structure within the Tetrahymena ribozyme. *Biochemistry 32*:5291-5300.
Murphy FL, Cech TR. 1994. GAAA tetraloop and conserved bulge stabilize tertiary structure of a group I intron domain. *J Mol Biol 236*:49-63.
Nagaswamy U, Fox GE. 2002. Frequent occurrence of the T-loop RNA folding motif in ribosomal RNAs. *Rna 8*:1112-1119.
Nahas MK, Wilson TJ, Hohng S, Jarvie K, Lilley DM, Ha T. 2004. Observation of internal cleavage and ligation reactions of a ribozyme. *Nat Struct Mol Biol 11*:1107-1113.
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*:4899-4903.
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.
Noller HF. 2005. RNA structure: reading the ribosome. *Science 309*:1508-1514.

Noller HF, Hoang L, Fredrick K. 2005. The 30S ribosomal P site: a function of 16S rRNA. *FEBS Lett 579*:855-858.
Nottrott S, Hartmuth K, Fabrizio P, Urlaub H, Vidovic I, Ficner R, Luhrmann R. 1999. Functional interaction of a novel 15.5kD [U4/U6.U5] tri-snRNP protein with the 5' stem-loop of U4 snRNA. *Embo J 18*:6119-6133.
Ogle JM, Brodersen DE, Clemons WM, Jr., Tarry MJ, Carter AP, Ramakrishnan V. 2001. Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science 292*:897-902.
Olson W. 1980. Configuration statistics of polynucleotide chains. *Macromolecules 13*:721-728.
Osborne EM, Schaak JE, Derose VJ. 2005. Characterization of a native hammerhead ribozyme derived from schistosomes. *Rna 11*:187-196.
Oubridge C, Kuglstatter A, Jovine L, Nagai K. 2002. Crystal structure of SRP19 in complex with the S domain of SRP RNA and its implication for the assembly of the signal recognition particle. *Mol Cell 9*:1251-1261.
Pabon-Pena LM, Zhang Y, Epstein LM. 1991. Newt satellite 2 transcripts self-cleave by using an extended hammerhead structure. *Mol Cell Biol 11*:6109-6115.
Penedo JC, Wilson TJ, Jayasena SD, Khvorova A, Lilley DM. 2004. Folding of the natural hammerhead ribozyme is enhanced by interaction of auxiliary elements. *Rna 10*:880-888.
Pley HW, Flaherty KM, McKay DB. 1994a. Model for an RNA tertiary interaction from the structure of an intermolecular complex between a GAAA tetraloop and an RNA helix. *Nature 372*:111-113.
Pley HW, Flaherty KM, McKay DB. 1994b. Three-dimensional structure of a hammerhead ribozyme. *Nature 372*:68-74.
Quigley GJ, Rich A. 1976. Structural domains of transfer RNA molecules. *Science 194*:796-806.
Reijmers TH, Wehrens R, Buydens LM. 2001. The influence of different structure representations on the clustering of an RNA nucleotides data set. *J Chem Inf Comput Sci 41*:1388-1394.
Reuter K, Nottrott S, Fabrizio P, Luhrmann R, Ficner R. 1999. Identification, characterization and crystal structure analysis of the human spliceosomal U5 snRNP-specific 15 kD protein. *J Mol Biol 294*:515-525.
Rich A, Kim SH. 1978. The three-dimensional structure of transfer RNA. *Sci Am 238*:52-62.
Robertus JD, Ladner JE, Finch JT, Rhodes D, Brown RS, Clark BF, Klug A. 1974. Structure of yeast phenylalanine tRNA at 3 A resolution. *Nature 250*:546-551.
Rojas AA, Vazquez-Tello A, Ferbeyre G, Venanzetti F, Bachmann L, Paquin B, Sbordoni V, Cedergren R. 2000. Hammerhead-mediated processing of satellite pDo500 family transcripts from Dolichopoda cave crickets. *Nucleic Acids Res 28*:4037-4043.
Rupert PB, Ferre-D'Amare AR. 2001. Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. *Nature 410*:780-786.
Rupert PB, Massey AP, Sigurdsson ST, Ferre-D'Amare AR. 2002. Transition state stabilization by a catalytic RNA. *Science 298*:1421-1424.
Salvo JL, Belfort M. 1992. The P2 element of the td intron is dispensable despite its normal role in splicing. *J Biol Chem 267*:2845-2848.
Schultes EA, Spasic A, Mohanty U, Bartel DP. 2005. Compact and ordered collapse of randomly generated RNA sequences. *Nat Struct Mol Biol 12*:1130-1136.

Scott WG, Finch JT, Klug A. 1995. The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage. *Cell 81*:991-1002.
Semlow DR, Silverman SK. 2005. Parallel selections in vitro reveal a preference for 2'-5' RNA ligation upon deoxyribozyme-mediated opening of a 2',3'-cyclic phosphate. *J Mol Evol 61*:207-215.
Serganov A, Keiper S, Malinina L, Tereshko V, Skripkin E, Hobartner C, Polonskaia A, Phan AT, Wombacher R, Micura R, Dauter Z, Jaschke A, Patel DJ. 2005. Structural basis for Diels-Alder ribozyme-catalyzed carbon-carbon bond formation. *Nat Struct Mol Biol 12*:218-224.
Stage TK, Hertel KJ, Uhlenbeck OC. 1995. Inhibition of the hammerhead ribozyme by neomycin. *Rna 1*:95-101.
Stage-Zimmermann TK, Uhlenbeck OC. 1998. Circular substrates of the hammerhead ribozyme shift the internal equilibrium further toward cleavage. *Biochemistry 37*:9386-9393.
Stage-Zimmermann TK, Uhlenbeck OC. 2001. A covalent crosslink converts the hammerhead ribozyme from a ribonuclease to an RNA ligase. *Nat Struct Biol 8*:863-867.
Sundaralingam M. 1969. Stereochemistry of nucleic acids and their constituents. *Biopolymers 7*:821-838.
Szewczak AA, Moore PB, Chang YL, Wool IG. 1993. The conformation of the sarcin/ricin loop from 28S ribosomal RNA. *Proc Natl Acad Sci U S A 90*:9581-9585.
Tamura M, Holbrook SR. 2002. Sequence and structural conservation in RNA ribose zippers. *J Mol Biol 320*:455-474.
Tan E, Wilson TJ, Nahas MK, Clegg RM, Lilley DM, Ha T. 2003. A four-way junction accelerates hairpin ribozyme folding via a discrete intermediate. *Proc Natl Acad Sci U S A 100*:9308-9313.
Tinoco I, Jr., Bustamante C. 1999. How RNA folds. *J Mol Biol 293*:271-281.
Torres-Larios A, Swinger KK, Krasilnikov AS, Pan T, Mondragon A. 2005. Crystal structure of the RNA component of bacterial ribonuclease P. *Nature 437*:584-587.
Turner DH. 1996. Thermodynamics of base pairing. *Curr Opin Struct Biol 6*:299-304.
Varani G, Wimberly B, Tinoco I, Jr. 1989. Conformation and dynamics of an RNA internal loop. *Biochemistry 28*:7760-7772.
Vicens Q, Westhof E. 2001. Crystal structure of paromomycin docked into the eubacterial ribosomal decoding A site. *Structure 9*:647-658.
Wadley LM, Pyle AM. 2004. The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. *Nucleic Acids Res 32*:6650-6659.
Weichenrieder O, Wild K, Strub K, Cusack S. 2000. Structure and assembly of the Alu domain of the mammalian signal recognition particle. *Nature 408*:167-173.
Westhof E, Dumas P, Moras D. 1985. Crystallographic refinement of yeast aspartic acid transfer RNA. *J Mol Biol 184*:119-145.
Westhof E, Fritsch V. 2000. RNA folding: beyond Watson-Crick pairs. *Structure Fold Des 8*:R55-65.
Westhof E, Michel F. 1994a. *Prediction and experimental investigation of RNA secondary and tertiary foldings*: Oxford:IRL Press, Oxford Univ. Press.
Westhof E, Michel F. 1994b. *Prediction and experimental investigation of RNA secondary and tertiary foldings.*

Wilson DS, Szostak JW. 1999. In vitro selection of functional nucleic acids. *Annu Rev Biochem 68*:611-647.
Wimberly B, Varani G, Tinoco I, Jr. 1993. The conformation of loop E of eukaryotic 5S ribosomal RNA. *Biochemistry 32*:1078-1087.
Wimberly BT, Brodersen DE, Clemons WM, Jr., Morgan-Warren RJ, Carter AP, Vonrhein C, Hartsch T, Ramakrishnan V. 2000. Structure of the 30S ribosomal subunit. *Nature 407*:327-339.
Woese CR, Winker S, Gutell RR. 1990. Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc Natl Acad Sci U S A 87*:8467-8471.
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 Res 31*:3450-3460.
Youngman EM, Brunelle JL, Kochaniak AB, Green R. 2004. The active site of the ribosome is composed of two layers of conserved nucleotides with distinct roles in peptide bond formation and peptide release. *Cell 117*:589-599.
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.
Zhang Y, Epstein LM. 1996. Cloning and characterization of extended hammerheads from a diverse set of caudate amphibians. *Gene 172*:183-190.
Zhao ZY, Wilson TJ, Maxwell K, Lilley DM. 2000. The folding of the hairpin ribozyme: dependence on the loops and the junction. *Rna 6*:1833-1846.
Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res 31*:3406-3415.