2013-07-15

Tetraloop-Receptor Interactions in RNA Crystal Structures

Wu, Li

UNIVERSITY OF CALGARY

Tetraloop-Receptor Interactions in RNA Crystal Structures

by

Li Wu

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF BIOLOGICAL SCIENCES

CALGARY, ALBERTA

JULY, 2013

Abstract

This dissertation addresses two aspects of RNA tertiary structure. The first section is a computational analysis of tetraloop-receptor interactions in RNA crystal structures. A total of 78 loop-receptor interactions were collected from the Protein Data Bank and grouped into four structural classes. The majority have the standard conformation discovered by earlier studies, indicating that these are the most favoured conformations. However, much structural diversity was found in the rest of this set, and several potential motifs were identified that can be studied in the future. In the second section, a FRET-based analysis was designed to obtain distance information between two selected sites in a group II intron. This project is at an initial stage, and will be extended in future. Together, the results advance our understanding of a specific type of RNA structural motif, and lay groundwork for further modelling a group II intron RNA in the future.

Acknowledgements

I would like to show my appreciation to my supervisor, Dr. S. Zimmerly, for directing me into the field of RNA biology, as well as many academic advices during these years. I always have a hard time while doing formal writing, and once again I would like to thank him for his careful reviewing of all of my writings, including this thesis. I would also like to thank my committee members, Dr. S-L. Wong and Dr. V. Zaremberg, as well as my internal examiner, Dr. D. C. Schriemer. Their advices during the defence helped me a lot to revise thesis and give it a much better organisation. I would also like to thank Dr. M. Fraser for helping with a part of this project, Dina R. for going through my thesis, and finally, the members in Zimmerly lab for the past years.

Table of Contents

List of Figures

List of Tables

List of Symbols, Abbreviations and Nomenclature

bp	basepair
CF	correction factor
CL	chloroplast-like
cpm	count per minute
cpRNA	circularly permutated RNA
CRISPR	clustered regularly interspaced short palindromic repeats
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
DTT	dithiothreitol
EBS	exon binding sequence
EDTA	ethylenediaminetetraacetic acid
FRET	Förster (Fluorescence) resonance energy transfer
IBS	intron binding sequence
IEP	intron-encoded protein
kb	kilo base
LSU	large subunit
LTR	long terminal repeat
miRNA	micro RNA
ML	mitochondrial-like
mRNA	messenger RNA
nm	nanometer

ncRNA	non-coding RNA
NMR	nuclear magnetic resonance
nt	nucleotide
NTL	non-tetraloop
NTP	nucleoside triphosphate
-OH	hydroxyl group
ORF	open reading frame
PCR	polymerase chain reaction
PDB	protein data bank
pKS+	pBluescript KS II +
RMSD	root-mean-square deviation
RNA	ribonucleic acid
rRNA	ribosomal RNA
RT	room temperature; reverse transcriptase
siRNA	small interfering RNAs
smFRET	single molecule Förster resonance energy transfer
SSU	small subunit
tmRNA	transfer-messenger RNA
TPP	thiamine pyrophosphate
tRNA	transfer RNA
UTP	uridine 5'-triphosphate
WC	Watson-Crick

CHAPTER ONE: INTRODUCTION

The discovery and study of ribonucleic acid (RNA) has dramatically shaped our understanding of the flow of genetic interaction. Inside a biological system, genetic information is controlled by a precise framework known as the central dogma of molecular biology. Deoxyribonucleic acid (DNA) is the basic information storage molecule. Information stored in DNA can be copied into messenger RNA (mRNA) through transcription. The mRNA then serves as the template for synthesis of protein. The process of synthesising protein is termed translation, and this requires two additional types of RNA: ribosomal RNA (rRNA) and transfer RNA (tRNA). Ribosomal RNA is one of the basic components of the ribosome, which scans along mRNA in the 5'-3' direction and recognises the codons. Then, the correct protein sequence is formed by joining the amino acids that are carried by tRNA to the ribosome. This basic concept of the central dogma of molecular biology was further refined with the discovery of reverse transcriptase, an enzyme that enables the conversion of RNA back to DNA, confirming that the flow of information between DNA and RNA may occur in either direction [1]. All of these processes are facilitated and controlled by various biological catalysts that are termed enzymes. Enzymes were once thought to be proteins exclusively. However, catalytic RNAs (ribozymes) were experimentally confirmed in 1981, demonstrating that RNA is capable to act as biological catalysts [2-4]. In addition, many regulatory RNA elements have been identified, such as micro RNAs (miRNAs), small interfering RNAs (siRNAs) or clustered regularly interspaced short palindromic repeats (CRISPR) RNAs. In our modern view, RNA is not merely an intermediate during gene expression, but carries out a multitude of tasks within the cell, some of which are complicated tasks usually carried by proteins.

After its discovery, catalytic RNA was rationalised by the fact that RNA can fold into compact, ordered structures. The conservation in both secondary and tertiary structures of different ribozymes proved the biological significance of their specific structures [5]. RNA tertiary structure is the main focus of this thesis. For the rest of this introductory chapter, various topics will be introduced, including examples of resolved RNA molecules, common methods used in RNA structural study, RNA structural building blocks, and finally, the aims of this thesis.

1.1 Examples of RNAs Associated with Specific Tertiary Structure

To reveal the biological functions of RNA at molecular level, an essential approach is to understand its tertiary structure, and this is facilitated by resolved RNA crystal structures. To date, many RNA crystal structures have been solved, and these are responsible for our understanding of RNA structure. Several representative examples of them will be introduced in this section.

1.1.1 Ribosomal RNA

Ribosomal RNA (rRNA) is one of the most important biomacromolecules and is involved in the process of translation. It normally comprises the majority of RNA from living cells. It makes up the structured catalytic portion of ribosomes but requires the assistance of ribosomal proteins to function properly *in vivo* [6]. A complete ribosome consists of a small subunit (SSU) and a large subunit (LSU). Prokaryotes and eukaryotes have different types of subunits, but they are similar in terms of secondary and tertiary structures. The first crystal structure of the ribosome in high-resolution was resolved in 2000, giving detailed access to the structural composition at atomic

level [7-9].

1.1.2 Transfer RNA

Transfer RNA (tRNA) is also involved in the process of translation. It serves as an adaptor between mRNA and the ribosome. During translation, it pairs with the mRNA codons and delivers amino acids to the ribosome. A tRNA molecule is generally small (~100 nucleotides [nt]) and has conservation in sequence [10]; its secondary structure has been characterised as the cloverleaf structure, while its tertiary structure has been resolved and described as L-shaped since the 1970s [11-14].

1.1.3 RNase P

RNase P was the first characterised ribozyme along with the group I intron. It is a type of ribonuclease and functions to cleave the tRNA precursor and lead to the maturation of tRNA molecules [4]. RNase P exists in all organisms, while its sequence and structural composition vary. RNase P has been one of the most-studied ribozymes, and its highly resolved crystal structure has been reported by several groups [15-18].

1.1.4 Riboswitches

Riboswitches are mainly found in bacteria. They are a segment within a mRNA that can regulate the expression of the mRNA by interacting with a specific effector molecule [19-21]. Riboswitches can be divided into multiple types based on the effector molecule. Some of their

tertiary structures have been determined, such as the TPP (thiamine pyrophosphate) riboswitch and the lysine riboswitch [22-24]. A riboswitch usually consists of two parts, an aptamer and an expression platform. When the concentration of effector increases, it will be bound to the aptamer, causing the expression platform to undergo a structural change [25]. The altered structure may then prevent either transcription or translation or trigger alternative splicing [26]. As a result, the level of gene expression is down-regulated. Because of its biological function, engineered riboswitches are sometimes introduced to study of control gene expression [27-29].

1.1.5 Group I and II Introns

Group I and II Introns are considered large and complex ribozymes. They can self-splice themselves out of the pre-mRNA transcript. Group I introns are mainly found in organelles and ribosomal RNA of fungi, protists and plants, while group II introns are mainly found in bacterial and archaeal genomes, as well as the organelles of fungi, protists and plants. Both of group I and II introns are mobile elements and have the ability to insert into intronless genes, which is also known as homing [30-33]. These introns usually encode a protein cofactor (intron-encoded protein, IEP) and this protein is required for splicing and retrohoming *in vivo*. However, they may undergo self-splicing under non-physiological conditions. Group I and II introns both have little similarity in sequence, but can form into conserved secondary and tertiary structures [34]. On the other hand, the mechanisms of splicing of group I and II introns are different, making the mechanisms a remarkable distinguishing feature. Splicing of group I introns is initiated by nucleophilic attack of guanosine at the 5’ splice site, and after splicing the intron is usually linear topologically. The splicing of group II introns is initiated by the attack of a single nucleotide, usually an A, near the 3' end of the intron. This causes the formation of an intron lariat, and the

intron after splicing remains as a lariat shape [34]. Group II introns are of special interest because they have been hypothesised as the ancestor of eukaryotic spliceosomes because of their comparable splicing mechanism [35]. To generate highly resolved crystal structures of such large RNAs takes much effort due to their rather large size. Crystal structures of group I and II introns both came out within the past decade, which happened much later than the construction of their tertiary models [36, 37]. Since group II introns are one of the subjects of this thesis, more details about them will be introduced later.

Besides the RNAs discussed above, other types of RNA exist, including the hairpin ribozyme, the hammerhead ribozyme [38-41], and the universal signal recognition particle RNA (SRP RNA), which is involved in directing protein translocation or transport at translational or post-translational level [42-46]. These RNA do not have highly conserved sequences, but in general, they are all conserved in secondary and tertiary structures.

1.2 Methodologies of RNA Structural Study

Structure-related studies are essential for understanding how RNA functions. There are many techniques that have been developed to either resolve the entire structure or reveal local structural arrangement. Structural information from different sources can be combined and used to refine an existing structural model or as a reference for other similar molecules. Common methodologies include both physical and chemical methods, such as X-ray crystallography, nuclear magnetic resonance (NMR), Förster resonance energy transfer (FRET), and RNA inline probing. In addition, for the modelling of a tertiary structure, computational assistance is required.

1.2.1 X-Ray Crystallography and Nuclear Magnetic Resonance

In terms of resolving the entire structure of a macromolecule, X-ray crystallography and NMR are the most favoured as they both can result in atomic resolution [47-49]. An important advantage of NMR compared to X-ray crystallography is that NMR can be used to trace dynamic changes in structure [50, 51]. To date, there are over one million of X-ray/NMR -resolved structures available [http://www.rcsb.org/, 52]. Since NMR is only capable of resolving relatively small molecules, X-ray crystallography is the primary method to generate tertiary structures of large RNAs (>75 nt). In this study, all structures analysed were resolved by X-ray crystallography. However, X-ray crystallography requires the growth of crystal, making it difficult to solve structures for many large RNAs.

1.2.2 Förster Resonance Energy Transfer

Förster resonance energy transfer (FRET) is a mechanism of energy transfer between two chromophores. By selecting proper pairs of chromophores, FRET can be used to measure the distance between two chosen chemical groups. A variant application derived from FRET is called single molecule Förster resonance energy transfer (smFRET), in which single molecules are monitored. FRET-based experiments usually measure distances ranging from 1 to 10 nanometres (nm) and are preferred in studies of dynamic inter-molecule interactions as well as tracing the formation of intermediates [53]. By labelling two chromophores at interacting sites within a RNA molecule, conformational changes can be monitored dynamically by measuring changes in FRET against time. Therefore, FRET-based measurements are especially informative for refining specific local regions inside a RNA structure.

1.2.3 RNA Inline Probing

When a RNA molecule undergoes a conformational change, usually caused by binding to a ligand, it may form two or more phases corresponding to the presence or absence of ligand, as well as the ligand concentration. The reacting core is always compacted when it binds to the ligand, whereas in a ligand-free environment, this region tends to be less organised. Because RNA can be hydrolysed relatively easily in a basic environment, less structured regions will be hydrolysed first, whereas compacted regions have higher resistance to hydrolysis. Combining other techniques such as primer extension, it is possible to determine the regions involved during the conformational change by comparing the cleavage patterns with and without ligand. This method is especially useful for the study of riboswitches [54]. A recent report also demonstrated that taking advantage of either internal- or end- labelled fluorescent dyes, inline probing can be used to achieve a more complicated structural analysis by screening the cleavage pattern under different wavelengths [55].

1.2.4 Computational Assistance

In addition to the methods discussed above, computational approaches are usually involved in tertiary structure modelling and structural prediction. They provide a different way to carry out structural studies of macromolecules. Instead of experimentally resolving a RNA structure, computational approaches construct a tertiary model solely based on thermodynamic calculations and structural homology. Experimental data is usually relied upon in order to narrow down the predicted candidate structures, and the final tertiary model reflects the integration between experimental data and computational analysis. Currently, many databases, online servers and

programs are available for either data collection or structure prediction, including those that will be discussed in this thesis.

1.3 Common Structural Elements of RNA

Like DNA or protein molecules, RNA molecules have their own specific structural elements or motifs. Understanding the common motifs that can be found in RNA will assist analysis of RNA tertiary structure in a larger context. In general, a RNA tertiary structure can be broken down into several substructures. Each substructure is the combination of several RNA tertiary motifs. A tertiary motif is made by a set of interacting secondary motifs, while a secondary motif is formed by the primary RNA sequence. In this section, some well-known secondary and tertiary elements of RNA will be discussed.

1.3.1 Secondary Structural Elements

Base pairs are the most basic secondary elements in both DNA and RNA structures. A RNA base has three edges: the Waston-Crick (WC) edge, the Hoogsteen edge and the sugar edge. A RNA base-base interaction may be formed between any two of these three edges, resulting in different geometric families of basepairing. The canonical WC basepair is formed between the two WC edges of each base, while one common example of non-WC basepairs is the G-U wobble basepair. RNA base-base interactions can be notated by the Leontis-Westhof annotation, which uses open/solid circle, triangle or square to present the interacting edges of base as well as the glycosidic bond orientation (Figure 1) [56, 57]. When an RNA strand contains portions that have complementary sequences, a stem-loop structure will be formed through basepairing. A short,

single-stranded terminal loop usually appears at the end of the stem. Nucleotides that do not pair with others inside the helix will become bulges or internal loops, depending on the length of unpaired region. Both the terminal loop and the internal loop may interact with other regions within a RNA structure, including tetraloop-receptor interaction, which is one of the topics of this thesis and will be discussed below.

When more than two stem-loops come together, pseudoknots may be formed between them. A pseudoknot is a stem-loop complex, in which one loop makes up part of the stem region of another stem-loop structure [58]. It has been reported that pseudoknots formed in viral RNA are required in order to interact with bacterial RNase P, demonstrating the significant biological effect of the pseudoknot and ensuring it to be an important RNA secondary element [59].

Symbol	Interacting edges of the bases	Glycosidic bond orientation
	Watson-Crick/Watson-Crick	*Cis*
	Watson-Crick/Watson-Crick	*Trans*
	Watson-Crick/Hoogsteen	*Cis*
	Watson-Crick/Hoogsteen	*Trans*
	Watson-Crick/Sugar	*Cis*
	Watson-Crick/Sugar	*Trans*
	Hoogsteen/Hoogsteen	*Cis*
	Hoogsteen/Hoogsteen	*Trans*
	Hoogsteen/Sugar	*Cis*
	Hoogsteen/Sugar	*Trans*
	Sugar/Sugar	*Cis*
	Sugar/Sugar	*Trans*

Figure 1: Leontis-Westhof notation for base-base interaction.

1.3.2 Tertiary Structural Motifs

An RNA double helix is similar to an A-form DNA helix in three dimensional structures. Interactions between a RNA double helix and other regions are usually formed at the minor groove. One example is the A-minor motif, in which an adenosine interacts with the minor groove of the double helical RNA [60, 61]. The A-minor motif was first discovered through the analysis of the crystal structure of *Haloarcular marismortui* by its abundant and ubiquitous occurrence, and now it is known as one of the most common features in RNA structures [60]. Other single nucleotides may also interact with the minor groove to form triple basepairs [62]. This type of interaction mainly serves to stabilise the long-range interaction in RNA structures, including the loop-helix interactions [60].

Besides the A-minor motif, another common feature in RNA structures is the loop-receptor interaction. In this type of interaction, nucleotides of a terminal loop contact with a double helix and form a long-range interaction. The most common example of loop-receptor interactions is the GNRA-tetraloop-receptor interaction. The loop consists of a conserved "GNRA" base composition, and the receptor can be different structural elements such as the minor groove of a double helical structure or the well-studied 11-nt or IC3 motifs. The A-minor motif can sometimes be observed as part of the entire interaction. This type of interaction is always found in ribozymes such as group I and II introns or RNase P, and they were studied by biochemical approaches [63, 64]. As one of the main topics of this thesis, this type of interaction will be given more details in next chapter.

Interactions may also form between two RNA strands. Because a hydroxyl group (-OH) can

serve as both hydrogen bond donor and acceptor, the 2'-OH of ribose sugar in one RNA strand can form hydrogen bond with those from another strand, and this type of interaction is termed the ribose zipper [65]. In addition, two adjacent RNA helices may form into one helix and is further stabilised by noncovalent base stacking, and this type of structure is known as coaxial stacking.

1.4 Topics of This Thesis

As more RNA crystal structures became available, more studies focusing on RNA structural motifs or RNA structural modelling were carried out. Currently, several different RNA structural motifs have been identified and described, but there is no comprehensive study that analyses all these motifs together and provides relative information such as the sequence-structural correlation.

The first topic of this thesis is mainly about the tetraloop-receptor motif in RNA structures. Relying on computational approaches, a collection of known and potential loop-receptor interactions were first extracted from currently resolved crystal structures and then analysed. They were grouped into four major classes according to the tertiary structural features, and into subclasses. Features of each class are then described, followed by an analysis of the correlation between sequence and structure.

The second part of this thesis focuses on refining the current tertiary structural model of the group II intron Ll.LtrB. As another aspect of RNA structure studies, modelling provides an alternative way to learn about large RNA molecules that does not have resolved X-ray crystal

structures. The tertiary model of Ll.ltrB has been previously constructed, but there are uncertain regions because there is no crystal structure available. A FRET-based experiment has been designed in order to obtain more accurate distance information between selected sites and test the model. However, this experiment remains at an initial stage of development and will be extended in the future.

CHAPTER TWO: LOOP-RECEPTOR INTERACTIONS IN RNA CRYSTAL STRUCTURES

2.1 Introduction

2.1.1 The Tetraloop-Receptor Interaction

Tetraloops were identified many years ago as an important element in RNA architecture. They consist of four single-stranded nucleotides and generally are located at the very end of a stem-loop structure. Some loops interact with other regions in a larger structural context. Although the exact nucleotide composition of each tetraloop varies, they fall into classes based on a consensus sequence. The most common classes are UNCG and GNRA tetraloops, while other rare sequence patterns exist, such as CUYG, ANYA and AUGC [66-70].

The structural features of some of these classes have been characterised. In a UNCG tetraloop (N=A/C/G/U), U1 and G4 form a wobble basepair, and N2 and C3 extend along each side of the backbone while C3 stacks on U1 (Figure 2A) [66, 71]. Together with several other hydrogen bonds inside the structure, UNCG-tetraloops are extremely stable, even compared with DNA loops that have the same sequence [72, 73]. They act as nucleation sites during RNA folding, and stabilise or protect the local tertiary structure from the surrounding blocks. The U-G pair also serves as a recognition site under certain situations [71, 74]. Nonetheless, interactions between UNCG tetraloops and other RNA structural elements almost never occur [75, 76].

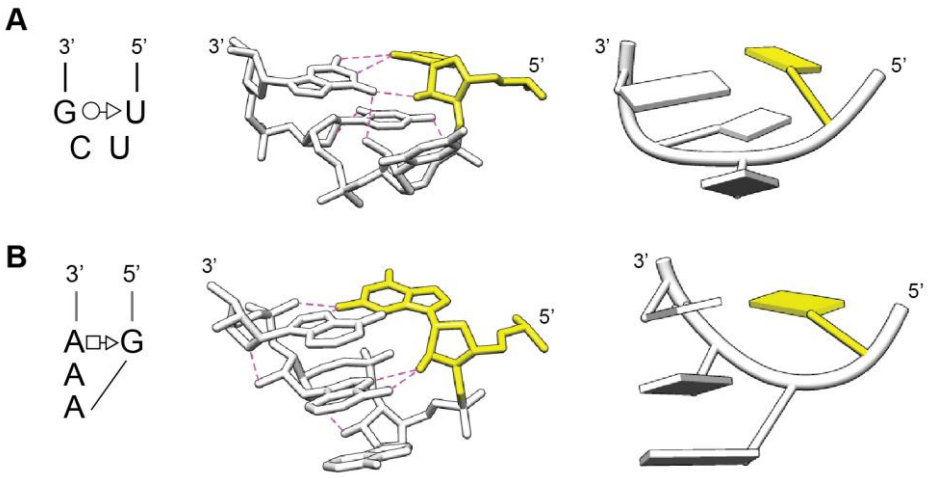


Figure 2: Examples of UNCG and GNRA tetraloops.

A) *A UNCG-tetraloop from crystal structure 1F7Y;* ***B)*** *A GNRA-tetraloop from crystal structure 2R8S. Each of the tetraloops is depicted by its secondary structure (left) and its tertiary structure represented in atom/bonds format (middle) as well as slab/tube format (right). The Leontis-Westhof notation of the structure is shown in the secondary structure. Yellow nucleotides indicate the 5' end in tertiary structures, and pink dashed lines in the atoms/bonds representative reflect hydrogen bonds.*

In case of GNRA-tetraloops (R=A/G), G1 and A4 form a non-WC G-A basepair, while N2, R3 and A4 stack on each other (Figure 2B) [77]. Compared to UNCG tetraloops, they are less stable thermodynamically and are much more likely to interact with other RNA building blocks or proteins [77, 78]. GNRA-tetraloops tend to be more variable in sequence as well. There are cases in which a loop contains more than four nucleotides but still forms a conformation capable of imitating the GNRA-tetraloop interaction; sequences of these long loops can be described as GNR[X_n]A, in which X_n stands for one or more extra nucleotides existing between R3 and A4. Moreover, non-adjacent nucleotides can come together to trans-form a functional GN/RA conformation, in which the slash indicates a sequence break [79].

GNRA-tetraloops have been detected to interact with three major types of receptors: the 11-nt

motif, the IC3 motif and the double-helical minor groove [80]. The 11-nt motif was first identified in the P4-P6 domain from the *Tetrahymena* group I intron [81]. There are 11 nucleotides involved in this motif, and a sequence of CCUAAG-UAUGG is always favoured. It includes four base pairs and a platform structure mostly consisting of two adenosines, as well as a flipped-out single nucleotide (mostly a uridine). This 11-nt motif receptor has been shown to have a specific interaction with GAAA-tetraloops, with several hydrogen-bonds formed between the loop and the receptor, and the loop partially stacking onto the adenosine platform, making it a generally stable interaction [63, 78, 80-83]. The GAAA/11-nt interaction motif is sometimes introduced into RNA constructs for crystallisation, because it facilitates ordered packing [84].

The IC3 motif was identified in group IC3 introns. This type is similar to the 11-nt motif but has a slightly different sequence of CCCUAAC-GAGGG [80]. Unlike the 11-nt motif, the IC3 motif does not require a specific loop sequence; any GNRA-tetraloop may bind to an IC3 receptor with lower affinity compared to the 11-nt motif [80].

Double-helical minor groove receptors are very simple and ubiquitous in RNA structures. Of any interaction from this category, the GNRA-tetraloop faces into the minor groove and forms multiple hydrogen-bonds with the receptor surface. For most cases, R3 and A4 from the loop form triple base pairs with the two receptor base pairs. Some variations from this category have been discovered; however, there has not been the reported of strongly supported correlations of sequence and structure, indicating the diversity of this simple yet common interaction [18, 64, 77, 78, 80, 85].

2.1.2 Identifying Loop-Receptor Interactions

As an important building block in RNA tertiary structures, tetraloop-receptor interactions have been studied both experimentally and computationally. At an earlier time, most motifs were discovered by biochemical methods including the 11-nt motif and the IC3 motif [64, 78, 80, 85]. With the development of computational approaches, additional tools have been integrated into this field to facilitate identification of structural motifs. Most computation-based procedures are based on the comparison of the backbone geometry or root mean square deviation (RMSD) to look for superposable regions from a set of X-ray crystal structures. Recurring structural elements could thus be assigned as a motif [68, 69, 79, 86-88]. Some motifs, such as the A-minor motif, the E-loop, the C-loop, the kink-turn and ribose zippers were identified by these approaches [9, 60, 89-92]. However, potential motifs that fall outside of any superposable sets would be overlooked by these approaches.

2.1.3 The Aim of Study

Although loop-receptor interactions have been identified and studied for a long time, there is no any overall survey that describes all currently available types, and that analyses the variety that exists in nature. Therefore, the first part of this thesis focuses on the loop-receptor interactions with an overall compilation and analysis. A list of 78 unique tetraloop-receptor interactions was first collected from all available crystal structures. These structures were assigned into different classes based on their structural features, followed by the description of each class. For each class, the ratio, the diversity in sequence and structure, and possible sequence-structural correlations were analysed. Some of the results were in agreement with previous studies, such as

the GAAA-11-nt motif interaction, but there were also a few new potential correlations that to be confirmed in future. Finally, one subset of this study was connected to the topic of structural evolution in group II introns.

By this approach, instead of performing a direct comparison to look for superposable regions among a set of given structures, this study carried out the analysis in a reverse way. It started by compiling all interactions between loops and helices before performing analyses onto them. This avoids the overlooking of potential motifs or patterns due to the lack of multiple superposable occurrences.

Although this study was somewhat limited by the number of available crystal structures, the information gathered here display all possible loop-receptor patterns that form in known RNA structures. A few new structural patterns were also revealed, which may be a potential reference for structural modelling.

2.2 Materials and Methods

2.2.1 Data Collection and the Nonredundant List

All crystal structure files were downloaded from Protein Data Bank (http://www.rcsb.org/pdb/) in the format of PDB [93]. The download criteria included: RNA as the macromolecule type; the chain length was longer than 50 nucleotides; the experimental method was X-ray crystallography; the resolution was less than 4 Å; and the deposit date was prior to January 1st, 2011. In addition, sequences with greater than 90% identity to other downloaded sequences were

removed. When multiple structures were available for a given molecule, only the structure with highest resolution was retained.

2.2.2 Detecting Potential Loop Interactions

The secondary structure of each crystal structure was obtained from the relevant publication. According to the secondary structures, all loops (including both tetraloops and non-tetraloops) were located. The crystal structure was loaded in SwissPDB Viewer (http://www.expasy.org/spdbv/) [94], and all atoms within a radius of 7 Å surrounding the loop were selected. Potential loop interactions were then identified visually, extracted and saved into individual PDB files. Interactions that span across unit cells were not considered as valid because this study only focused on those occurring within the same macromolecule. In the case of ribosomal RNAs, a quick superposition was done by SwissPDB Viewer for every single interaction at equivalent positions of both the eukaryotic and prokaryotic representatives, and if they were visually identical, only the one with higher resolution was retained.

2.2.3 Classification Procedure

All interactions were initially divided into several "starter" clusters based on their secondary and gross tertiary structures. Afterwards, pairwise superposition and the resulting RMSD values were used to reassign structures into classes. Superposition and RMSD calculation were performed by the UCSF Chimera package (http://www.cgl.ucsf.edu/chimera/) [95], mainly using four functions: Match, MatchMaker, Ensemble Cluster and Ensemble Match. Match can generate the RMSD value between two molecules by calculating user-defined atoms; MatchMaker takes one

point per nucleotide to calculate RMSD values between a reference and a set of structures; Ensemble Cluster and Ensemble Match are always used together. Ensemble Cluster divides the input molecules into groups based on RMSD values that are calculated by using user-defined atoms. Ensemble Match displays all the pairwise RMSD values of the input molecules. In this study, unless otherwise indicated, all RMSD values generated by Match and Ensemble Match were calculated using the default settings and the following 12 backbone atoms for each nucleotide: P, OP1, OP2, C1', C2', C3', C4', C5', O2', O3', O4', O5'.

2.2.4 The Criteria for Each Class

Because the classification mainly relied on structural superposition and pairwise RMSD values, to choose a fixed set of atoms for the calculation is important to keep the result consistent. Based on the result of this study, a maximum pairwise RMSD value was set to every RMSD-based group together with the number of nucleotides that should be measured for the calculation (Table 1). In general, all members from the same class should fulfill the criteria except some special cases such as atom/nucleotide modification or truncated/extended sequence. In addition, while assigning a new interaction to an existing class, judgement taking account of the overall structural arrangement is sometimes required.

Table 1: Overview of criteria used for RMSD calculation of each class.

Class	Subclass	Acceptable maximal pairwise RMSD value (Å)	Required minimal number of loop nucleotides (backbone atoms)	Required minimal number of receptor nucleotides (backbone atoms)
I[1]	—	1	4 (48)	11 (132)
II	1[2]	1.5	4 (48)	4 (48)
	2-4[3]	1.5	4 (48)	4 (48)
	5[4]	—	—	—
III	1	1	2 (24)[5]	4 (48)
	Individual[6]	3	4 (48)	—
IV	1	1	2 (24)[7]	4 (48)
	2-5[8]	—	—	—

[1] The 11-nt motif is another feature that can support this group separately.
[2] Value was taken from 1.1.1 and 1.1.2, the two standard subsets for GNRA/helix interaction.
[3] The RMSD value was given based on the case lf subclass 1. Subclasses 2-4 differ from subclass 1 by the relative loop/receptor positioning, and thus the RMSD value within each subclass should be similar to that of subclass 1.
[4] Subclass 5 only contains one occurrence and thus cannot be calculated.
[5] The two loop nucleotides correspond to T2 and T3 of a standard tetraloop.
[6] Class III focuses on interactions that resemble a standard tetraloop, and thus receptors are not considered in the calculation.
[7] The two loop nucleotides correspond to T3 and T4 of a standard tetraloop
[8] These subsets are not RMSD-based.

2.2.5 Verifying the Electron Density Maps

The electron density maps were downloaded from the Electron Density Server if available (http://eds.bmc.uu.se/eds/) [96]. If the server did not have the map, the electron density maps were calculated according to the structure factors associated with each structure in Protein Data Bank using either PHENIX (http://www.phenix-online.org/) [97] or programs from the CCP4 package (http://www.ccp4.ac.uk/) [98]. The program Coot was used to visualise the models and electron density maps (http://lmb.bioch.ox.ac.uk/coot/) [99]. The analysis was based on the

average electron density at the centres of each atom in the model not considering the nearby electron density. The final judgement was based on the consistency between the electron density and the structure model itself.

2.2.6 Miscellaneous

The Leontis-Westhof notation of base-base interaction for depicting secondary structure was extracted by program S2S (http://bioinformatics.org/assemble/) [100]. All molecular graphics were generated by UCSF Chimera.

2.3 Results and Discussion

2.3.1 The Nonredundant Pool of RNA Crystal Structures

The first step in the project was to assemble a collection of RNA crystal structures that had potential loop-receptor interactions. A set of RNA-containing crystal structures was first retrieved from the Protein Data Bank, resulting in 1348 entries spanning across different types of RNAs, including ribozymes, RNase P and RNA-protein complexes from various organisms. However, since the exact same molecule may be reported multiple times, the majority of 1348 structures were duplicates. To generate a nonredundant list, these files were further screened. Only one representative with highest resolution was retained per species when multiple occurrences existed. For ribosomal RNAs, the *Escherichia coli* large and small subunits (LSU and SSU) were chosen as representatives because prokaryotic and eukaryotic ribosomal subunits share some common features. Meanwhile, the *H. marismortui* LSU was chosen as a representative of archaea ribosomes. After removing the redundancy, the finalised list contained

41 unique RNA crystal structures (Table 2).

Table 2: The list of 41 unique RNA crystal structures analysed in this study.

PDB ID	Macromolecule Name	Source Organism	Length (nt)	Resolution (Å)
RNase P				
3OK7[1]	RNase P holoenzyme with tRNA (type A)	*Thermotoga maritima*	347	3.8
1U9S[2]	RNase P RNA specificity domain (type A)	*Thermus thermophilus*	161	2.9
2A64[3]	RNase P RNA (type B)	*Bacillus stearothermophilus*	417	3.3
1NBS[4]	RNase P RNA specificity domain (type B)	*Bacillus subtilis*	155	3.15
Group I introns				
2R8S[5]	P4-P6 ribozyme domain	*Tetrahymena thermophila*	159	1.95
1X8W[6]	Group I ribozyme	*Tetrahymena thermophila*	247	3.8
1U6B[7]	Group I ribozyme with both exons	*Azoarcus sp.*	197	3.1
1Y0Q[8]	Group I ribozyme	*Staphylococcus phage Twort*	229	3.6
Group II introns				
1KXK[9]	Group II intron domains 5,6 (ai5g)	*Saccharomyces cerevisiae*	70	3.0
3IGI[10]	Group IIC intron	*Oceanobacillus iheyensis*	412	3.12
Small ribozymes				
3NKB[11]	Hepatitis delta virus ribozyme	Hepatitis delta virus	64	1.92
1M5O[12]	Hairpin ribozyme	Tobacco ringspot virus	92	2.2
2QUW[13]	Hammerhead ribozyme, cleaved fragment	Tobacco ringspot virus	57	2.2
3IVK[14]	RNA polymerase ribozyme	Synthetic	128	3.1
3CUL[15]	Aminoacyl-tRNA synthetase ribozyme	Synthetic	92	2.8
2Z75[16]	GlmS ribozyme RNA	*Thermoanaerobacter tengcongensis*	125	1.7
3L3C[17]	GlmS ribozyme RNA	*Bacillus anthracis*	141	2.85
2OIU[18]	L1 Ribozyme, ligase circular adduct	Synthetic	71	2.6

Riboswitches				
3GX5[19]	SAM-I riboswitch variant bound to SAM	*Thermoanaerobacter tengcongensis*	94	2.4
2QWY[20]	SAM-II riboswitch bound to SAM	Environmental sequence	52	2.8
3E5C[21]	SMK box (SAM-III) riboswitch with SAM	*Enterococcus faecalis*	53	2.25
3NPB[22]	TL5 RNA (SAM-I), larger molecule	*Bacillus subtilis*	119	3.02
2QBZ[23]	M-Box riboswitch aptamer domain	*Bacillus subtilis*	161	2.6
3LA5[24]	Adenosine riboswitch	*Vibrio vulnificus*	71	1.7
3NPQ[25]	S-adenosylhomocysteine riboswitch	*Ralstonia solanacearum*	54	2.18
2G9C[26]	Guanine riboswitch	*Bacillus subtilis*	67	1.7
3OWW[27]	Domain II of glycine riboswitch with glycine	*Vibrio cholerae*	88	2.8
3F2Q[28]	FMN riboswitch bound to FMN	*Fusobacterium nucleatum*	112	2.95
3DIL[29]	Lysine riboswitch bound to lysine	*Thermotoga maritima*	174	1.9
3MXH[30]	c-di-GMP riboswitch	*Vibrio cholerae*	92	2.3
2GDI[31]	TPP riboswitch	*Escherichia coli*	80	2.05
3D2V[32]	TPP-specific riboswitch	*Arabidopsis thaliana*	77	2.0
Ribosomes				
3OFO[33]	30S ribosomal subunit	*Escherichia coli*	1533	3.1
3OFR[33]	50S ribosomal subunit	*Escherichia coli*	2904	3.1
1VQO[34]	50S ribosomal subunit	*Haloarcula marismortui*	2922	2.2
SRP RNAs				
1MFQ[35]	7S RNA of SRP	*Homo sapiens*	128	3.1
1LNG[36]	SRP19-7S.S SRP RNA complex	*Methanocaldococcus jannaschii*	97	2.3
3KTW[37]	SRP19/S-domain SRP RNA complex	*Sulfolobus solfataricus*	96	3.2
Other RNAs				
2IL9[38]	Ribosomal binding domain of the IRES RNA	*Plautia stali intestine virus*	142	3.1
1KH6[39]	JIIIabc (IRES)	Hepatitis C virus	53	2.9
2CZJ[40]	tRNA domain of tmRNA	*Thermus thermophilus*	63	3.01

The reference of each crystal structure is listed in Appendix A as numbered in subscript in this table.

2.3.2 The Extracted Sub-PDB Files Containing Potential Interactions

Each PDB file was examined visually to look for potential loop-receptor interactions. In this study, any observed interaction where a loop participates was extracted. Potential interactions were extracted into individual files that contained only the loop and its receptor. Every extracted file was saved following the format of "XXXX:#-#" as the filename, in which "XXXX" stands for the four-character PDB ID, and "#-#" stands for the range of loop (e.g. 2R8S:189-192 stands for the loop ranging from positions 189 to 192 of the crystal structure 2R8S has an interaction with another region within the same structure). Considering that the same ribosomal subunit is structurally similar to other ribosomes, every pair of interactions at the same secondary position of *E. coli* and *H. marismortui* was superposed using SwissPDB viewer to examine the similarity. From each pair, if they were essentially identical, only the one from *H. marismortui* was retained because it had a higher resolution.

Overall, 91 potential loop-helix interactions were initially extracted; however, after another round of screening, 21 of them were considered as a "false" loop-helix interaction, as they were either loop-loop interactions or simply nearby blocks with no actual contacts. The remainder of the 78 substructures coming from 21 unique crystal structures were confirmed as the final set used in this study (Table 3).

Table 3: The list of the 78 extracted interactions and their classes.

PDB ID	Loop positions[1]	Receptor positions[1]	ED[2]
Class I			
1NBS	205-208	145-150, 159-163	++
1U6B	24-27	146-151, 160-164	++
1U6B	189-192	60-65, 80-84	++
2R8S	150-153	222-227, 247-251	++
Class II subclass 1.1.1			
2Z75	114-117	10-11, 30-31	++
3IGI	90-93	272-273, 280-281	+
3OK7	93-96	3-4, 340-341	+
3OK7	285-288	75-76, 84-85	+
3OFO	1077-1080	16-17, 918-919	++
3OFO	1266-1269	1311-1312, 1325-1326	++
3OFR	2857-2860	1708-1709, 1749-1750	++
Class II subclass 1.1.2			
1U9S	205-208	80-81, 93-94	+
3MXH	32-35	59-60, 78-79	++
1VQO	1629-1632	1553-1554, 1567-1568	++
1VQO	1863-1866	1467-1468, 1474-1475	++
Class II subclass 1.1 (individual)			
1Y0Q	22-25	170-171, 177-178	+
1Y0Q	205-208	60-61, 78-79	+
1VQO	469-472[3]	773-774, 887-888[3]	++
1VQO	577-580	1110-1111, 1252-1253	++
1VQO	1327-1330[3]	905-906, 1299-1300[3]	++
1VQO	2630-2633[3]	2114-2115, 2470-2471[3]	++
3OFO	1013-1016	987-988, 1217-1218	+
3OFR	1807-1810	1362-1363, 1368-1369	+
Class II subclass 1 (individual)			
1X8W	323-326	118-119, 202-203	+
1VQO	734-737	2382-2383, 2405-2406	+
3OFO	898-901	769-770, 809-810	++
3OFO	1516-1519	1404-1405, 1496-1497	++
Class II subclass 1 (NTL)			
1MFQ	169-174	126-127, 223-224	+
1NBS	175-179	132, 234-235	+
1U9S	182-188	135-136, 162-163	+
2GDI	67-72	21-22, 37-38	++
3D2V	55-60	13-14, 25-26	++
1VQO	119-121	50-51, 110-111	++
1VQO	873-877[3]	1832, 1844[3]	++
1VQO	1055-1059	2491-2492, 2529-2530	++

1VQO	1077-1082[3]	2067-2068, 2077-2078[3]	++
1VQO	1499-1506[3]	1420-1421, 1443-1444[3]	++
1VQO	1991-1997[3]	2583-2584, 2594-2595[3]	++
3OFO	1166-1170	1088-1089, 1096-1097	+
3OFR	956-961	2456-2457, 2494-2495	++
Class II subclass 2			
1VQO	691-694	2439-2440, 2452-2453	++
3OFR	630-633	2401-2403, 2414-2415	+
3OFR	1364-1367	186-187, 209-210	++
1VQO	1469-1473	156-157, 179-180	++
1VQO	2390-2398	915-916, 927-928	++
Class II subclass 3			
1LNG	163-166	208-209, 212-213	+
1MFQ	147-150	197-198, 201-202	+
3KTW	164-167	209-210, 213-214	++
1VQO	2564-2569[3]	2695-2696, 2699-2670[3]	++
Class II subclass 4			
3OFO	159-162	341-342, 347-348	+
1VQO	1595-1599	1537-1538, 1647-1648	++
Class II subclass 5			
3IGI	369-372	128-129, 234-238	+
Class III subclass 1			
1VQO	2837-2843[3]	2087-2088, 2656-2657[3]	++
3OFR	642-646	2348-2349, 2368-2369	++
Class III individual			
3DIL	125-129	23-24, 68-69	++
1VQO	218-222[3]	164-165, 170-171[3]	++
1VQO	1706-1712[3]	790-791, 823-824[3]	++
Class IV subclass 1			
2A64	98-107	55-56, 392-393	+
3OFR	124-127	54-55, 115-116	++
Class IV type 2			
2QBZ	100-106	21, 167-168	+
1VQO	2300-2307[3]	952, 1014-1015[3]	++
Class IV type 3			
1VQO	2069-2076[3]	2490, 2531[3]	++
3OFR	2210-2214	1359-1360, 1371-1372	++
3OFR	1493-1497	1418-1421, 1577-1580	+
Class IV type 4			
1VQO	196-200	415-416, 424-425	++
1VQO	1770-1773[3]	1829, 1885, 2017-2018[3]	++
1VQO	1834-1842[3]	2621-2622, 2642-2643[3]	++[4]
1VQO	1917-1922	418-419, 2448-2449	+
3OFO	461-470	202-203, 214-215	+

3OFO	523-526	11-12, 22-23	++
3OFR	159-167	2206-2207, 2217-2218	++
3OFR	226-229	409-410, 417-418	++
3OFR	2552-2556	2507, 2581-2582	++
Class IV type 5			
1VQO	391-398	2441-2442, 2450-2451	++
1VQO	671-675	36, 446	++
1VQO	838-845	1369-1371, 2054-2055	++
1VQO	2784-2788[3]	1153, 1213[3]	+
3OFR	1728-1732	1516	+

[1] The position numbers correspond to the specific PDB files.
[2] "ED" as "Electron density". "++" indicates at most minor deviations between the electron density map and the extracted local substructure (~8-15 nts). "+" indicates a greater degree of unmodeled positive or negative electron density.
[3] Indicates the positions when E. coli and H. marismortui have identical conformations. The interaction from E. coli was then omitted due to a lower resolution.
[4] The electron density indicates that base 1835 should be flipped 180° around the glycosidic bond.

To evaluate the X-ray resolution of each extracted structure, the electron density map of each of them was examined. Only the specific extracted local region was analysed and either "+" or "++" was used to indicate the qualitative degree of agreement between the electron density map and the actual model (Table 3). All structures marked with "+" had a greater degree of unmodeled positive or negative density, while the models with "++" were consistent with their electron density maps with only minor discrepancies. (This part of work was done by Dr. Marie Fraser at University of Calgary.)

2.3.3 "Starter Clusters" and Initial Class Assignments

The 78 extracted interactions were first divided into four "starter" clusters, roughly based on general features of the secondary and the tertiary structures. Starter Cluster 1 consisted of the

four GNRA/11-nt interactions, since this interaction was already well characterised and easily identified. Starter Cluster 2 contained 34 interactions that involved a tetraloop interacting with a double-helix. Starter Cluster 3 contained 26 substructures in which a non-tetraloop interacts with a double-helix. Starter Cluster 4 included 14 interactions, and had loops of different sizes interacting with non-helical regions (Table 4).

Table 4: Overview of the four starter clusters.

Starter Cluster	**Number of loop nucleotides**	**Receptor Type**	**Number of structures**
1	4	11-nt motif	4
2	4	Double-helix	34
3	More/less than 4	Double-helix	26
4	any	Non-helix	14

After more detailed examinations described below and in Materials and Methods, it was decided that the following four classes are the most appropriate divisions of structural types: Class I would include all the GNRA/11-nt interactions, Class II would include all of the standard tetraloop/minor groove interactions plus their variations, Class III would be the collection of loop/helix interactions that resemble those in Class II but have other types of contacts between the loop and helix, and Class IV would consist of all the remaining interactions, which are irregular and do not belong to any of the previous classes.

2.3.4 The Process of Classification

The process of classification into one of the four classes was performed through several rounds superpositions based on either atom-specified or automatic selection of superposing points. Interactions falling into either Class I or II were easier to classify, compared to Classes III and IV. The members of Starter Cluster 1 were directly migrated into Class I without any changes because of the distinctive structural features the interaction. To detect the Class II interactions, which include standard tetraloop/helix interactions and their variations, every single member from Starter Clusters 2 and 3 were first compared to other members within the same cluster, and structures with smaller pairwise RMSD values were put into subclusters. Representatives of each subcluster were then compared other across clusters 2 and 3 for the merging of similar sets. A simplified flowchart describing the process is shown in Figure 3, and the final results are shown in Table 3.

Two functions of the program Chimera -- Ensemble Cluster and Ensemble Match -- were used to generate pairwise RMSD information for Starter Cluster 2 because all the interactions involve 8 nucleotides (4 from the tetraloop and 4 from the two-basepair receptor). A total of 96 backbone atoms from these 8 nucleotides were chosen for calculating the pairwise RMSD values. The first round of applying Ensemble Cluster/Match on the Starter Cluster 2 generated eight subclusters, including three multi-member groups of 22, 3 and 3 respectively, plus six single-member groups. Accordingly, because the three multi-member groups all matched the characters of tetraloop/helix interaction, they were assigned as Class II subclasses 1, 2 and 3, while the 6 single-member groups were temporarily put aside as "side group A" due to the rather large pairwise RMSD values (Figure 3). More rounds of Ensemble Cluster/Match analysis were

performed to further divide the 22 subclass 1 members. The second round applied on the entire subclass 1 generated an 18-member group (subclass 1.1) and 4 single-member "groups" (subclass 1 individual). The third round was applied onto subclass 1.1, and resulted a 7-member group (subclass 1.1.1), a 4-member group (subclass 1.1.2), and 7 single-member groups (subclass 1.1 individual). For subclasses 2 and 3, due to the small size of each ($n<4$), they were kept the same as the result of the first round (Figure 3).

For Starter Cluster 3, the greater variety of their geometries made it difficult to perform Ensemble Cluster/Match onto the same set of atoms for calculating pairwise RMSD values and subgrouping. Therefore, MatchMaker was first used for a crude comparison across Starter Cluster 3 to automatically generate approximate subgroups. At this point, some structures contained nucleotides distant from the actual loop-receptor interaction, and these regions had to be removed before the automatic comparison to avoid artifactual superpositions. Starter cluster 3 could thus be roughly divided into one multi-member group of 9 plus 17 single-member "groups" (Figure 3). Since the RMSD values generated by MatchMaker were not always based on the same set of nucleotides, RMSD values were measured between each of the 26 members of Starter Cluster 3 and existing Class II representatives, to ensure a consistent basis for comparison. The result was in agreement with the first automated grouping, in which all interactions from the group of 9 had a rather smaller RMSD values, and the structures were similar to Class II subclass 1. This group was thus assigned as Class II subclass 1 NTL (non-tetraloop). On the other hand, the 17 individual structures did not resemble any of the Class II representatives, and thus were temporarily put aside as "Side Group B" (Figure 3).

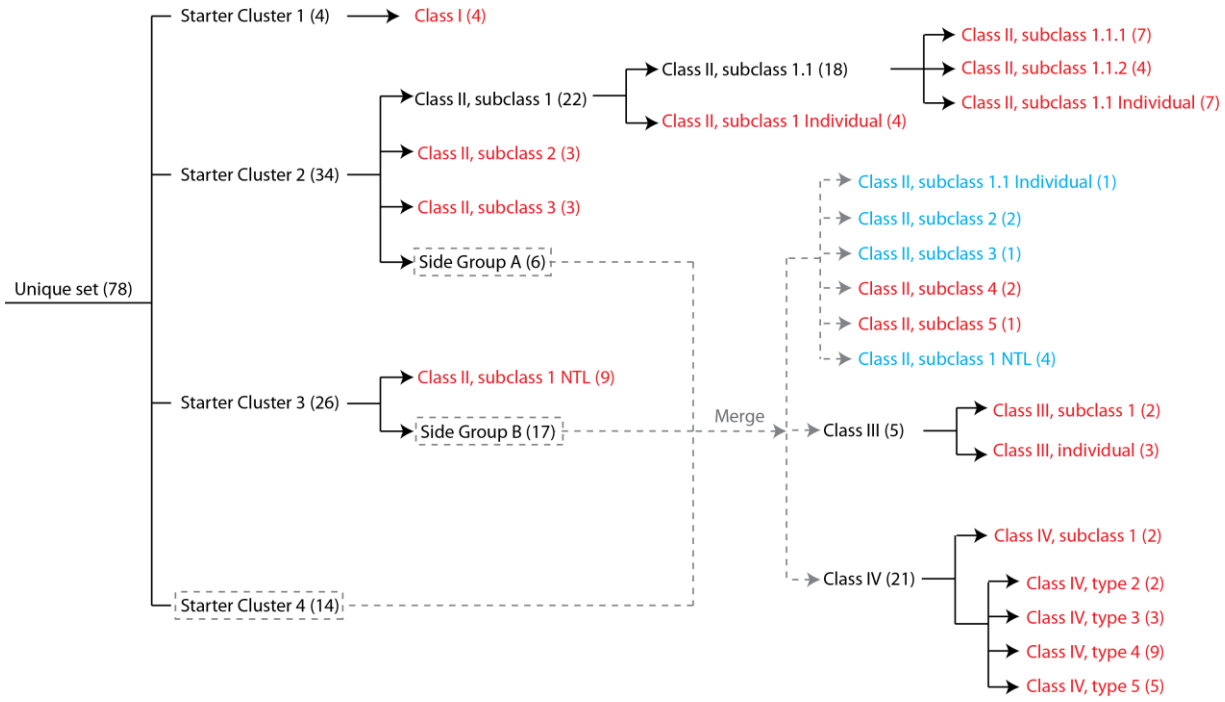


Figure 3: The main progress of classification.

This flowchart reflects how interactions from the four starter clusters were re-assigned into a final class. Solid lines indicate direct assigning of class whereas dashed lines indicate a process of merging and re-matching. Numbers in parentheses correspond to the number of structures inside of each group. Red text indicates the smallest groups that were first assigned a name, while blue text indicates interactions detected later but belonging to existing groups.

Starter Cluster 4 was not analysed by itself. Instead, this cluster was first merged with side groups A and B mentioned above, and similar steps were carried out as for Starter Cluster 3. Two interactions from this set were determined to be Class II, subclasses 4 and one was assigned as Class II, subclass 5. Another 8 structures were found to be similar to several existing classes in Class II, and thus were merged with them (Figure 3, blue text). For the remaining 26 structures, five were assigned as Class III, including a 2-member subclass, and 3 individuals. The last 21

interactions were put into Class IV. Class IV has only one RMSD-based subclass, and the rest are unique structurally based on RMSD criteria. However, the irregular Class IV interactions could be divided into four "types" that are based on structural features rather than RMSD values (Figure 3).

2.3.5 Distribution of Classes

After assigning classes, the distribution of structures used in this study was considered (Figure 4). Out of the 78 loop-receptor interactions, 48 of them are from Class II (~62%), which are the standard GNRA-tetraloop/helix interaction and the variations. The ratio of Class I interactions are about 5% out of the total, which is similar to that of Class III (~6%); the rest is taken by Class IV unique interactions (~27%). Within Class II, subclass 1 is almost 75% of structures and 46% out of all 78 structures, and thus becomes the most prevalent type of Class II interaction.

As the standard tetraloop/helix interaction motif, subclass 1 comprises the majority of all Class II structures. This is partially due to the stability of this configuration. While other classes have fewer contacts, a relatively large number of hydrogen bonds form between the loop and the receptor in the standard type, which stabilises the entire configuration, making it a very common motif for formation of a loop/helix interaction. Meanwhile, although not as common as subclass 1, each of the other subclasses still either actually has or is assumed to have more than one member existing, indicating the possibility of to form these types of interactions in nature. Finally, in addition to the rather small overall ratio of GNRA/11-nt interactions, there were no GNRA/IC3 interactions identified in this data set at all, which was unexpected because there has been experimental data supporting their existence [80].

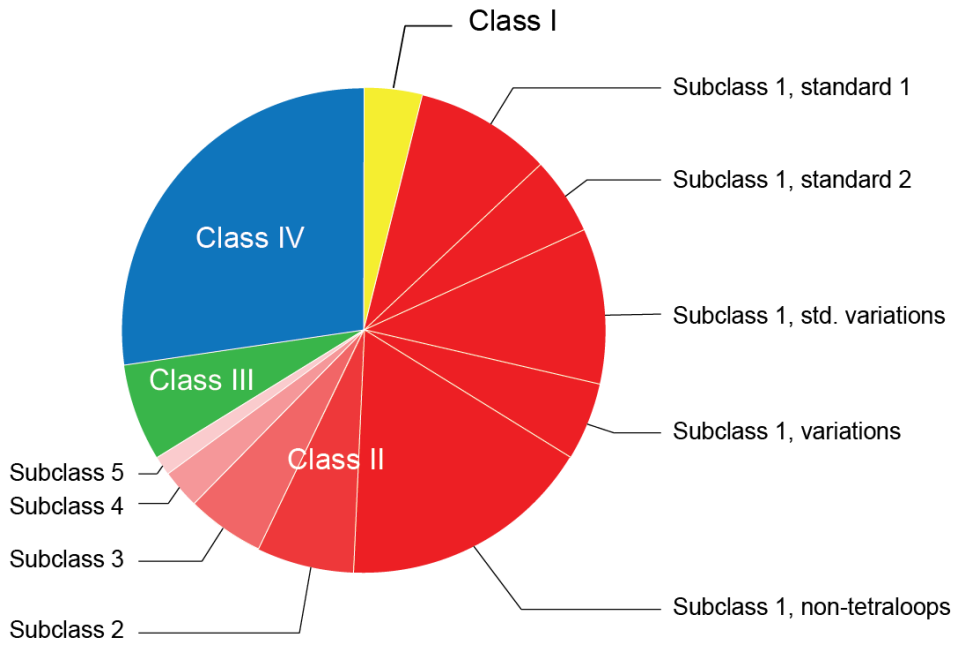


Figure 4: The relative distribution of classes.

Yellow, red, green and blue represent the four main classes, while different shades of red represent different subclasses of Class II.

2.3.6 Class Characterisation

2.3.6.1 Class I

Class I consists of four GNRA/11-nt motif interactions out of the overall 78 structures. They are almost identical in both sequence and hydrogen bonds (Figure 5). Three have identical sequences as the motif consensus (CCUAAG-UAUGG) [64], while the fourth has a one base variation in the adenosine platform (CCUA**C**G-UAUGG, with the deviating nucleotide displayed in bold) (see Appendix B). These four interactions can be superposed very well (Figure 5). The largest pairwise RMSD value among the four members was 0.806 Å based on 156 backbone atoms in

the loop and the receptor. (Because the atom composition of A190 and A191 was interrupted by base modification in the structure 1U6B:189-192, these two positions had to be removed while calculating the pairwise RMSD values for the entire set).

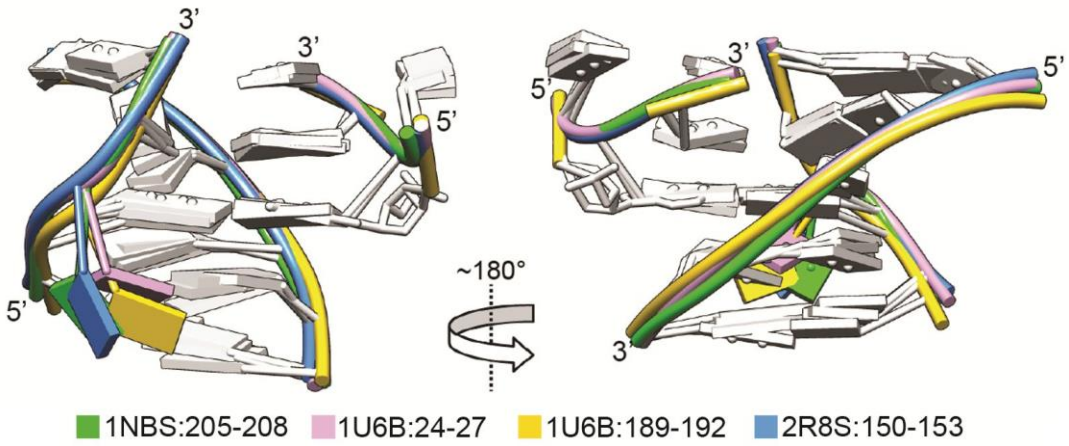


Figure 5: Two rotational views of superposed Class I interactions: GNRA-tetraloop/11-nt motif.

All four Class I structures are superposed based on 156 backbone atoms and are almost identical. The backbone of each individual is shown in a different colour. All bases except the flipping-out uridine are shown in white.

The most significant structural difference among all Class I structures is the position of the bulged uridine in the receptor. Inside structure 1U6B:24-27, the bulged U162 has multiple hydrogen bonds with A150 and G151, which are not seen in other members in Class I (Appendix B). Having or lacking these contacts may cause the different positions of this nucleotide. The surrounding structural units may also affect the formation of the different hydrogen bonds, yet details about the differences and whether they have specific function remained unknown. The

GAAA-11nt interaction is highly conserved in both secondary and tertiary structures among the four examples, and they all correspond to previous 11-nt motif related studies. The only surprising thing is that this type of interaction is absent from ribosomal RNAs, even though more than half of the interactions examined in this study were from rRNAs.

2.3.6.2 Class II

Class II is the largest class containing 48 of the 78 interactions, and their structures fall into five subclasses representing five different ways of contacts between a tetraloop and receptor. Since the loop is not always in the sequence of GNRA, the following text will sometimes annotate the four tetraloop nucleotides by T1-T4 instead of G1-N2-R3-A4.

2.3.6.2.1 Subclass 1

In spite of structures not having tetraloops (NTLs), almost all loops in this subclass have a sequence of GNRA and interact with the minor groove of a helical receptor. The loop bases are approximately coplanar to the receptor basepairs, and nucleotides T3 and T4 form triple basepairs with the receptor; the base of T2 also contacts the receptor via hydrogen-bonds formed with receptor bases as well as between two backbone riboses. As a result, all of these bonds stabilise the entire motif. This subclass can be further divided into several nested subsets: subclasses 1.1.1, 1.1.2, 1.1 (individual), 1 (individual) and 1 (NTL).

Subclass 1.1.1 and 1.1.2 are the most prevalent among all Class II interactions. These two subsets are nearly identical to each other, and were not differentiated until the third round of

grouping analysis (Figure 3, Figure 6). Because of their relative abundance, these two subsets together are considered as the standard tetraloop/helix interaction. However, if one looks at the loop sequences carefully, they can be roughly described as GYGA (1.1.1) and GNAA (1.1.2) respectively (Appendix B). The pairwise RMSD values are generally small (<1.2 Å based on 96 backbone atoms out of 8 nucleotides). Among these pairwise RMSD, the largest values in subclasses 1.1.1 and 1.1.2 are 0.64 Å and 0.82 Å respectively, while the largest value is 1.18 Å if including both subsets for calculation.

Subclass 1.1 (individual) is a subset at the 1.1 level. These structures have similar backbone pathways like those in 1.1.1 and 1.1.2, but have a greater variety in the base positions (Figure 7). The greatest pairwise RMSD value within this subset is 1.581 Å based on 96 atoms out of 8 nucleotides, which is smaller than other subclasses outside the 1.1 level, but is not as small as those in subclasses 1.1.1 and 1.1.2.

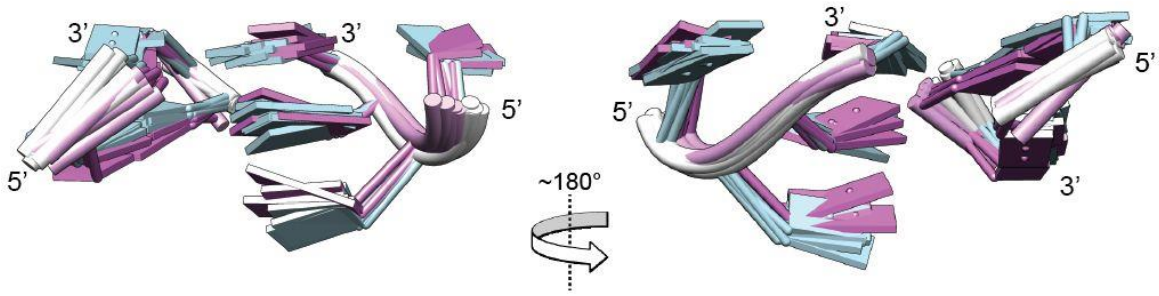


Figure 6: Two rotational views of Class II Subclass 1.1.1/1.1.2 interactions.

All 11 interactions of Class II, subclasses 1.1.1/1.1.2 are superposed based on 96 atom atoms. All 1.1.1 structures (with GYGA tetraloop) have white backbones and blue bases, while all 1.1.2 (with GNAA tetraloop) structures have pink backbones as well as bases.

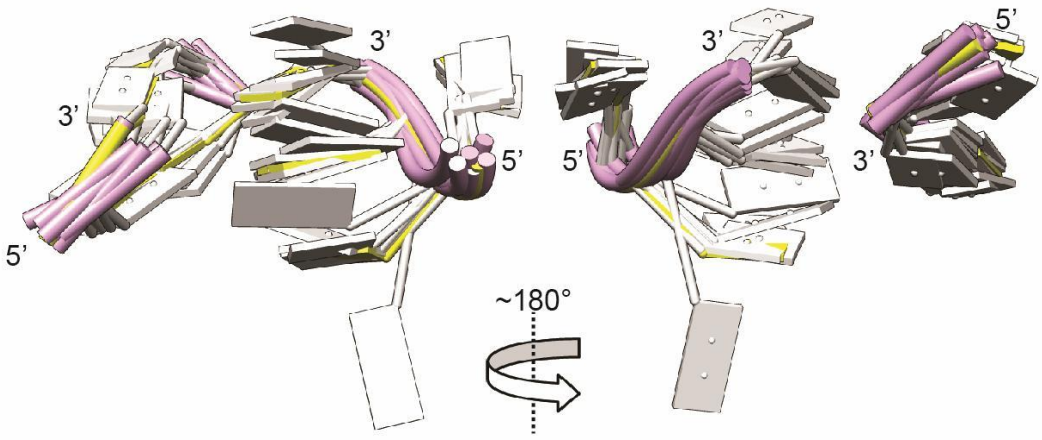


Figure 7: Two rotational views of Class II subclasses 1.1 (Individual) interactions.

The 8 Class II, subclass 1.1 (individual) interactions are superposed to 3IGI:90-93 (subclass 1.1.1) based on 96 backbone atoms. The subclass 1.1 structures have pink backbones and white bases, while 3IGI:90-93 is shown in yellow.

Interactions from Class II, subclass 1 (individual) and subclass 1 (NTL) are outside of the 1.1 level mentioned above. They are similar to the standard tetraloop/helix motif, but with a greater variety. In most cases, the backbone deviates from the standard pathway by different extents, and the bases between the loop and receptor are not coplanar, which results in a reduced number of hydrogen bonds connecting the loop and the receptor (Appendix B). These deviations caused much larger pairwise RMSD values inside subclass 1 (individual), ranging from 1.674 Å to 4.155 Å based on 96 backbone atoms out of 8 nucleotides. However, they were kept in subclass 1 because the structural arrangement agrees with the standard representation (Figure 8). In the case of subclass 1 (NTL) structures due to the different numbers of nucleotides involved for each member of the NTL subclass, it was not possible to calculate the pairwise RMSD values based on the same set of atoms. They belong to subclass 1 because the pathways of their backbones

follow that of a standard interaction, plus there is at least one loop base superposable to the standard representative (Figure 9). Two extraordinary examples exist in this set (Figure 9, green and blue). Compared with other NTL structures, these two have completely different backbone pathways of the loop region, thus resulting a weaker connection between loop and receptor, in which only one loop nucleotide is forming a small number (<3) of hydrogen bonds with the receptor bases. The diversity in this set of NTL interactions suggest that the groupings may be altered in future when more superposable interactions are available.

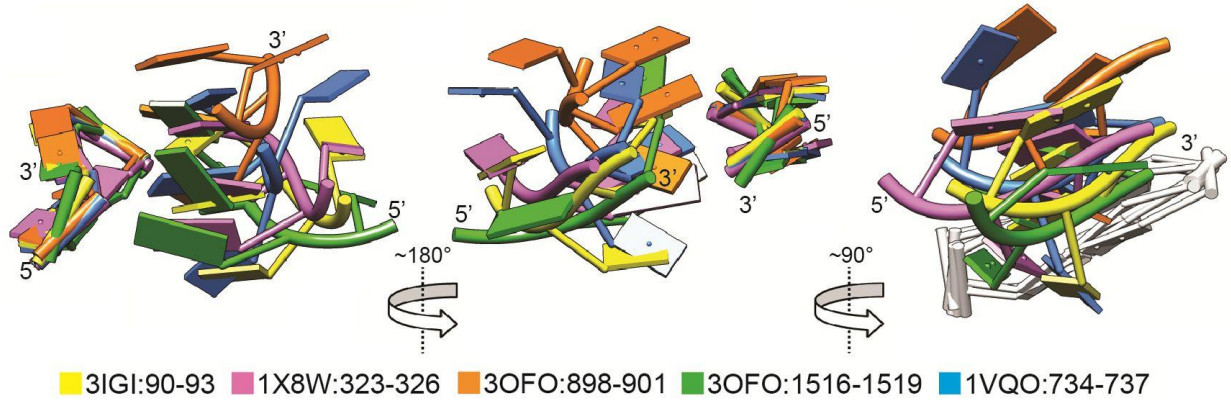


Figure 8: Three rotational views of Class II subclasses 1 (Individual) structures superposed based on the receptor.

Structures are superposed to the standard Class II, subclass 1 representative 3IGI:90-93 based on 96 backbone atoms. Each individual structure is labelled in a different colour. Two views have both the loop and receptor visible (left and middle). A third view is shown with the loop rotated to the front, and the receptors in white (right).

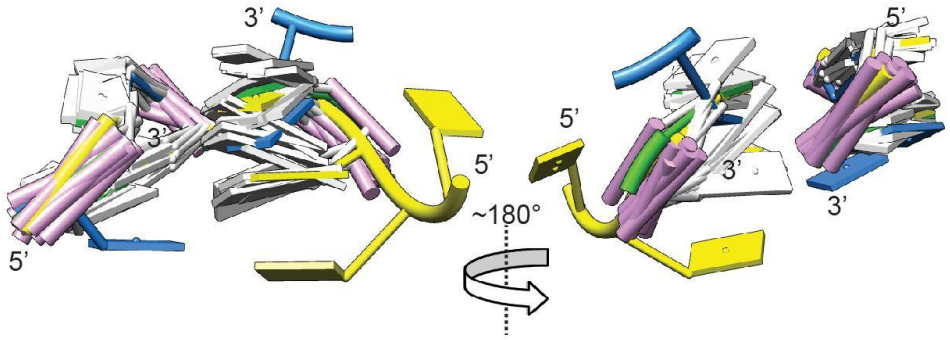


Figure 9: Two rotational views of Class II Subclass 1 (NTL) structures.

All the 13 structures of Class II, subclass 1 (NTL) are superposed to the standard Class II, subclass 1 representative 3IGI:90-93 (yellow), based on 4 nucleotides of the receptor (48 backbone atoms) plus one or two nucleotides of the loop (12 or 24 backbone atoms). Extra nucleotides of the loop that do not belong to the interaction have been removed. Most NTL structures are depicted by pink backbones and white bases, except two extraordinary examples that are in green and blue (see text).

2.3.6.2.2 Subclasses 2, 3 and 4

Subclasses 2, 3 and 4 can be considered to be three different ways for how a tetraloop can interact with a helical receptor. Compared to subclass 1, the contacting interfaces are shifted by different degrees. In subclass 2, T4 forms hydrogen bonds with the backbone of the receptor, with T2 and T3 interacting with the minor groove receptor instead of the T3/T4 scheme in subclass 1 (Figure 10A). Structures in subclass 2 generally have fewer hydrogen bonds compared to those in subclass 1 due to the position shift (Appendix B).

Three out of a total of four subclass 3 structures are GNRA tetraloops interacting with a receptor located at the very end of a stem-loop structure. The conformation of the receptor is affected

greatly by nearby nucleotides and appears skewed compared to a normal minor groove (Figure 10B). The pattern of hydrogen bonds is completely altered from the standard motif, and most contacts occur between T3/T4 and the receptor. One interesting thing is that except for the NTL interaction in this subclass, the other three members all belong to the 6 GNAR tetraloop motifs in helix 6 of signal recognition particles RNA (SRP-RNA). The three members are from different organisms (Table 2), and the sequence identities are very low (data not shown), but they all form a similar overall tertiary structures with the same tetraloop-helix interaction, implying the importance of this specific motif for function of the RNA [101, 102]. The only NTL structure in subclass 3 differs significantly from the other three in this subclass, but it still has some of the main contacts that can be seen in SRP-RNAs (Appendix B). Since the NTL structure is from ribosomal RNA, one may speculate that this region will contribute to some events such as binding and/or recognising other factors.

For subclass 4, the loops have a rotation compared to the standard position of subclass 1, causing the formation of a decreased interaction. For both members of this subclass, there is technically only one loop-receptor contact formed between T2 and the receptor (Figure 10C). Due to the small size of this subclass, it is hard to predict this type of interaction if solely based on the pattern of hydrogen bonds.

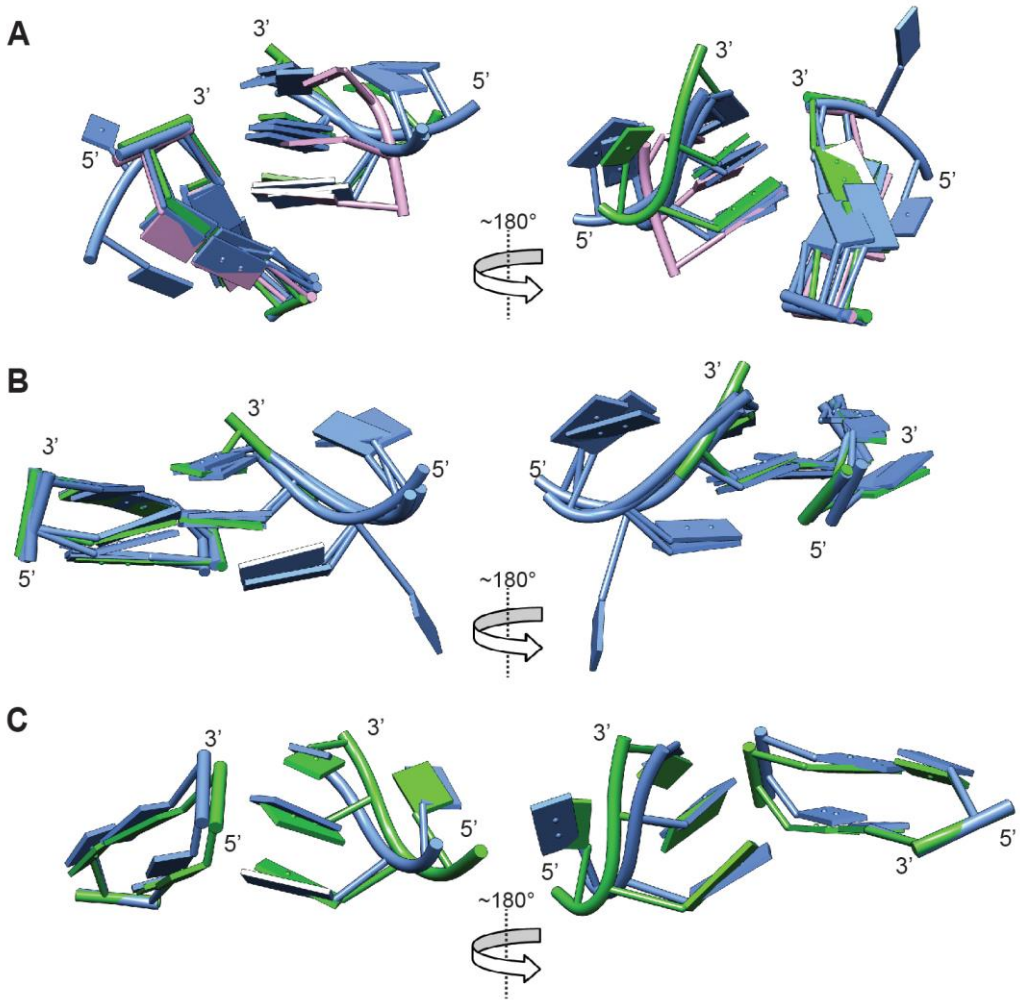


Figure 10: Two rotational views of Class II subclasses 2, 3 and 4 structures.

Panels A), B) and C) depict all members of Class II, subclasses 2, 3 and 4 respectively. Tetraloop interactions are shown in blue, while NTLs are shown in green and pink. Most interactions are superposed based on backbone atoms of the 4-nt loop and the 2-bp receptor, except the NTLs of subclasses 2 and 3, in which only two nucleotides of the loop plus the 2-bp receptor nucleotides are involved.

2.3.6.2.3 Subclass 5

The single structure in subclass 5 was initially put into Starter Cluster 4 along with other novel loop/non-helix interactions. However, two features of the structure are consistent with Class II interactions and it was reassigned as a single-member subclass under Class II. First, the loop is structurally in a nearly standard GNRA-tetraloop geometry (Figure 11), and second, the receptor is a double helical structure. Moreover, the interaction motif is a crucial interaction found across several different classes of group II introns based on its secondary structure, indicating that there are more occurrences of this structure that will fall into this subclass (see section for details) [103]. Subclass 5 has a distinctive arrangement from other Class II interactions mentioned above (see below, Figure 20, Appendix B). Instead of interacting with the minor groove, A370 stacks onto a single base that comes from a bulged 3-nt loop in the context of a double helix. A hydrogen bond between the riboses of the backbone of C372 and one of the receptor strands also exists (not shown). Although the tetraloop looks like it crosses the minor groove and has interactions, potential contacts like those found in other Class II structures are not found by hydrogen bond analysis.

2.3.6.2.4 Comparison of Class II Subclasses

The five subclasses inside Class II depict the different ways that a tetraloop may interact with a double helical receptor. The relative positions of all Class II subclasses are seen from the superposition of different subclass representatives, which are superposed either based on the loop or the receptor. When these structures were superposed based on 48 backbone atoms from the receptor, all receptors are almost identical, while the tetraloops show different degrees of rotation

relative to subclass 1 (Figure 11). When the loops are viewed from the front, the relative positions of all these loops clearly depict their spatial differences (Figure 11A, right). If taking a closer look at the placements of these tetraloops, subclass 2 (green) has a rotation of nearly 90° relative to subclass 1 (yellow), and the loops of subclasses 3 and 4 (orange and pink) are shifted from subclass 1 by about 15 Å but in different orientations. Since the receptor of subclass 5 is unique toward to the others, it was not included in the receptor-based superposition.

For the loop-based superposition based on 48 backbone atoms, and in which subclass 5 is included, a similar result is seen. While all the loops are in the same geometry, the receptors have either rotations or shifts relative to each other (Figure 11B). Again, when viewed from the front, it is clear that the five interactions are distinct (Figure 11B, right). Both the receptor- and loop-based superpositions indicate that various sequences have the ability to form five diverse possibilities in the final loop-receptor configuration.

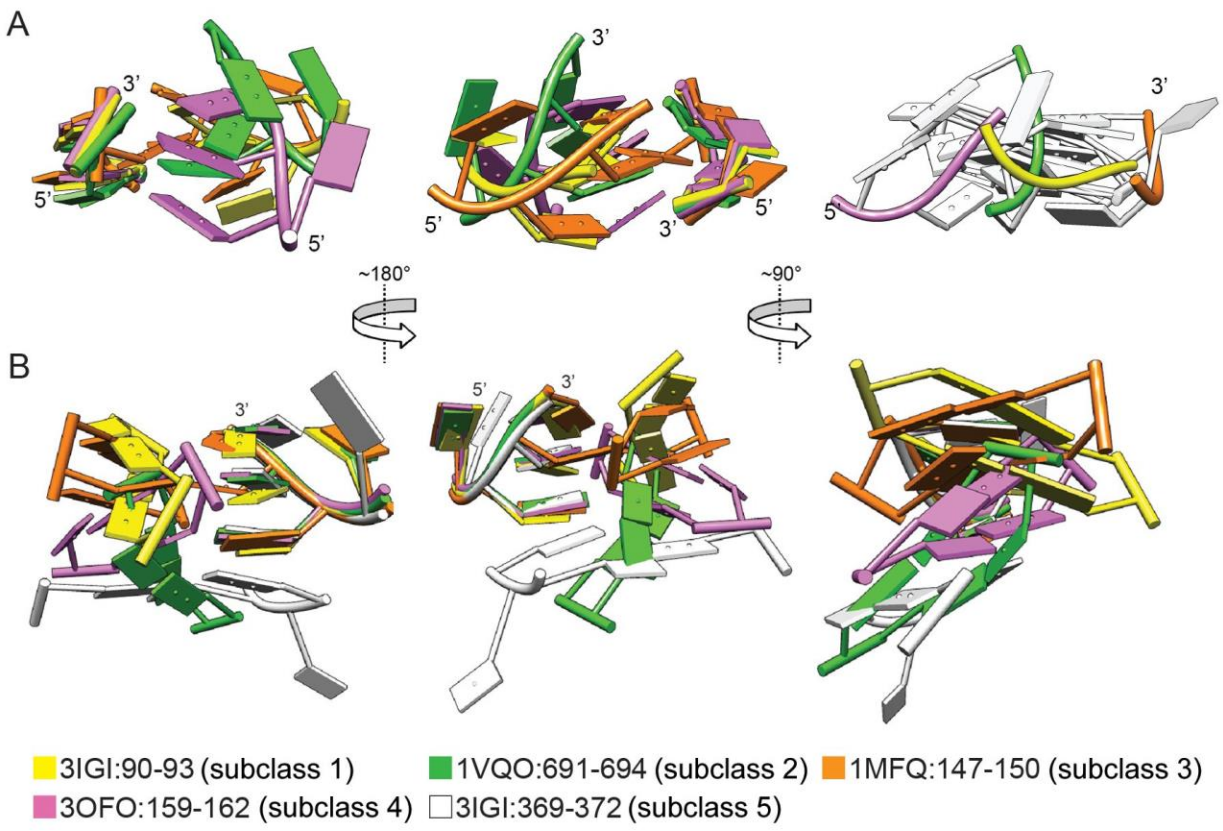


Figure 11: Loop/receptor -based superpositions across Class II subclasses.

Representative interactions were taken from each subclass and superposed to Class II, subclass 1, and each individual is shown in a different colour. Three views are shown for each superposition: both the loop and the receptor are visible in the left and middle views; the right view has either the loops or receptors in the front while the rest elements have been turned to white. ***A)*** *A receptor-based superposition of Class II subclasses 1-4.* ***B)*** *A tetraloop-based superposition of Class II subclasses 1-5.*

2.3.6.3 Class III

Class III consists of five Class II-like structures. Two structures out of the five belong to the same subclass (Figure 12), and they have an RMSD value of 0.78 Å based on 72 backbone atoms from T2, T3 plus the two-basepair receptor. When all these five Class III structures are superposed to the Class II subclass 1 representative, a greater diversity can be seen in terms of

the loop pathways (Figure 13). This superposition was based on 48 backbone atoms from four loop nucleotides that are equivalent to the four in a tetraloop. When all the loops are similar (maximum pairwise RMSD=2.899 Å), the positions of their receptors appear scattered around. It can also be seen that there are some unusual backbone pathways for the loop compared to the Class II representative (Figure 13A, B). On the other hand, if they are superposed based on 48 backbone atoms from the receptors, all of these receptors are normal double helical structures (maximum pairwise RMSD=1.421 Å). However, the backbone pathways of the loops differ (Figure 13C). In brief, similar to the case of Class II, these Class III structures might be considered as other Class II subclasses structurally. Yet they are currently kept outside of Class II, because they are neither in agreement with any of the Class II tertiary arrangement nor having a GNRA sequence in the loop.

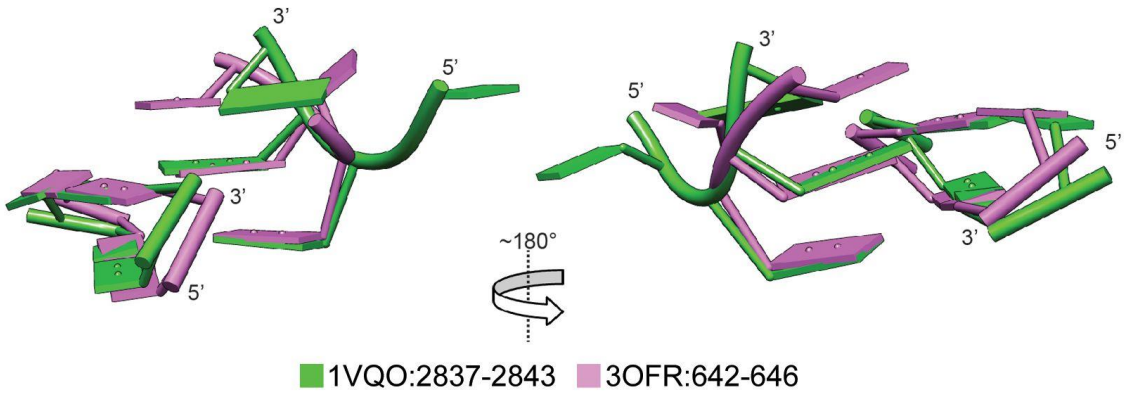


Figure 12: Two rotational views of superposed Classes III, subclass 1 structures.

Class III, subclass 1 contains only two members, and this superposition is based on 48 backbone atoms of the receptor plus 24 backbones of T2 and T3 of the loop.

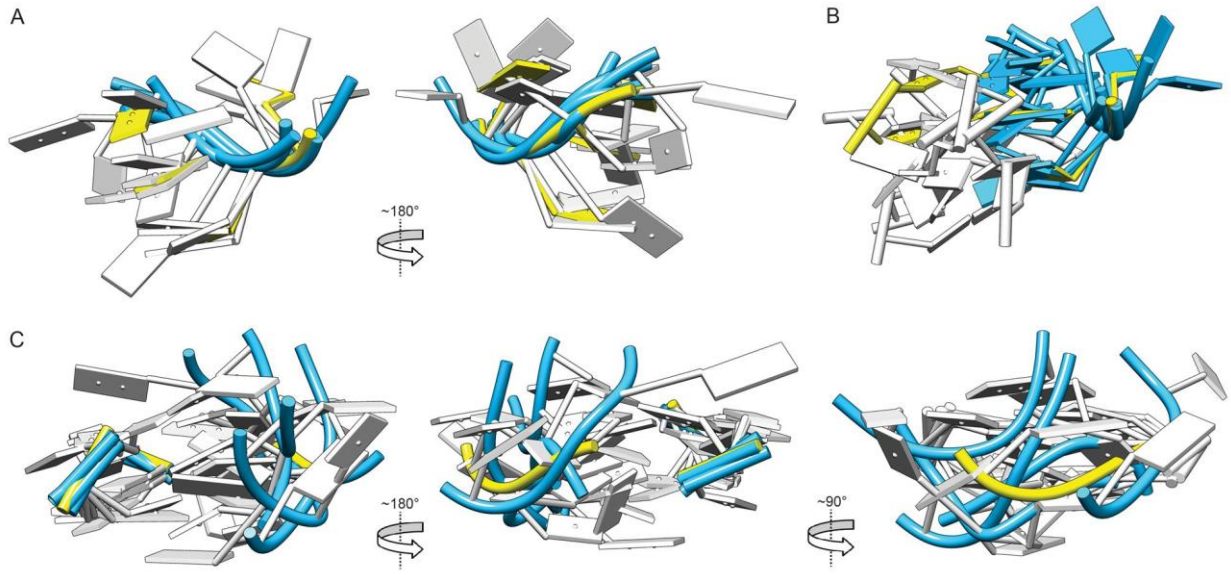


Figure 13: Structural comparisons of Class III structures against Classes II, subclass 1.

All the five Class III structures (blue) are superposed to the Class II, subclass 1 representative 3IGI:90-93 (yellow), based on all backbones of either the loop or the receptor. ***A, B)*** *A loop-based superposition. Only the superposed loops are visible in A), while both the loops and the receptors are visible in B).* ***C)*** *A receptor-based superposition. Left, middle) Two rotational views when the loops and the receptors are both visible; right) a view when the loops are turned to the front.*

Although only a few structures exist in this class, there are still two occurrences falling into the same subset. It would be of interest to collect information about the functions of these interactions, and compare with those standard Class II structures.

2.3.6.4 Class IV

Class IV is a collection of irregular structures that do not fall into the previously mentioned classes. There is one subclass containing two structures that mimic a Class II interaction by the

positioning of two loop nucleotides into where T3 and T4 usually locate in a normal GNRA tetraloop (Figure 14). However, they are not superposable to any of current Class II structures. Also, based on the backbone pathways of their loops, neither can be considered as tetraloop-like. Therefore, these two structures do not belong to either Class II or Class III. Besides, although the two have a RMSD value of 0.6 Å based on 72 atoms from a total of 6 nucleotides including two from the loop and four from the receptor, it is still questionable whether they actually belong to the same type of interaction due to the lack of a common pattern of hydrogen bonds (Appendix B). As a result, these two structures are basically temporarily put in this subclass until more detailed information about their functions becomes available.

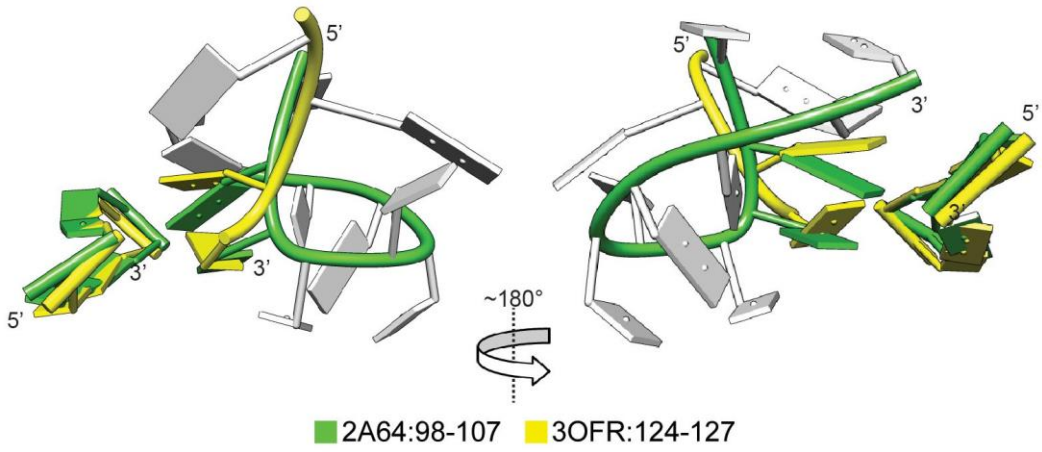


Figure 14: Two rotational views of superposed Class IV, subclass 1 structures.

This superposition is based on all backbone atoms of the 2-bp receptor and two loop nucleotides that correspond to T3 and T4 in a normal tetraloop. The entire backbones of the loop are highlighted in colour, while extra loop nucleotides that do not interact with the receptor are shown in white.

Instead of looking at the superposition and the part of hydrogen bonds, The rest of the 19 unique and irregular interactions technically should all be assigned as "Class IV individual" if following the previous naming rule. Yet here, based on the general structural features, they can be divided into four types (the names start from type 2 to type 5 to reduce confusion while being mentioned together with subclass 1) (Appendix B). Type 2 has two interactions in which a single base of the loop inserts into the minor groove of the receptor helix, stacking onto the receptor base pair (Figure 15A). The single base also interacts with another receptor base in its neighbourhood. Type 3 consists of three interactions in which the loop has several splayed nucleotides that forming a surface and contacting the receptor (Figure 15B). Type 4 contains all other interactions formed between multiple bases of the loop and the receptor, whereas Type 5 covers any interactions that have only a single base of the loop contacting the receptor. In spite of the variety of this entire class, the lack of multiple occurrences that have the same molecular arrangement stalls further analysis of these structures.

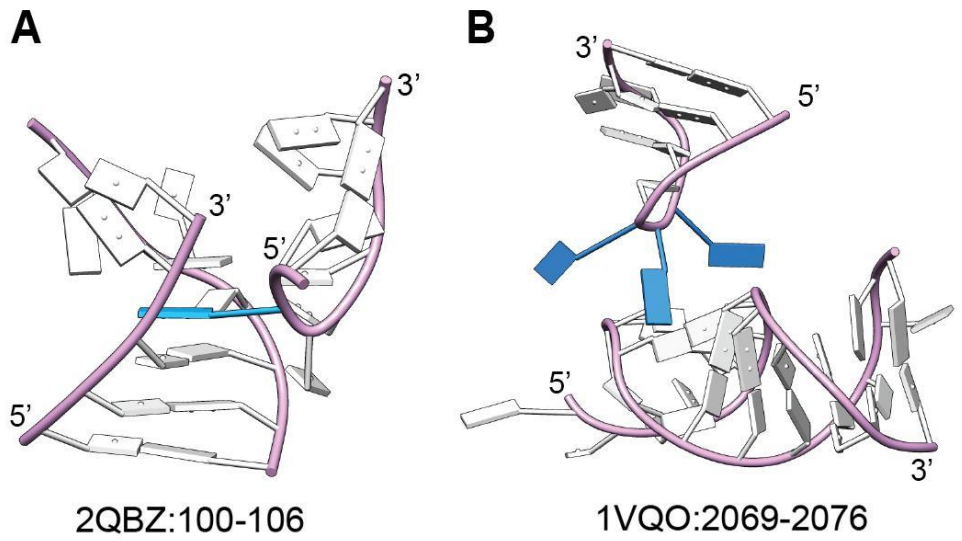


Figure 15: Two examples of Class IV structures showing their unique conformations.

The backbones are shown in pink, and some loop nucleotides that have interaction with the receptor are highlighted in blue. ***A)*** *An interaction in which a single nucleotide of the loop inserts into and interacts with the double helical receptor.* ***B)*** *An interaction in which three nucleotides from loop spread out and interact with the double helical receptor.*

From these unique Class IV structures, one structure 3OFR:1493-1497 shows a vague resemblance to Class I GNRA/11-nt interactions (Figure 16). This structure was first noticed by its bulged adenosine (orange) that flips out of the helix, which can be seen as the equivalent of the bulged uridine in an 11-nt motif. In addition, a few basepairs are exactly the same as those in an 11-nt motif, including a G-C Watson-Crick basepair and an A-U Watson-Crick/Hoogsteen basepair what flank the bulged adenosine, as well as an A-A *trans*-Watson-Crick basepair occurs between the loop and the receptor. These two types of interactions have some common sequences as well. One is the sequence of "AAA" in the loop (Figure 16, yellow) – even though they show different base orientations in the tertiary structures. The other one is the "CUAAG" sequence

found from one of the receptor strands. But instead of an adenosine platform, nucleotides that correspond to the two A's in an 11-nt motif simply stack on each other (Figure 16, green), while the bottom A contacts a nucleotide in the opposite strand. There is no similar base-base contact to hold a tetraloop conformation in this occurrence, and thus causes a more dispersed arrangement of the loop nucleotides. Fewer hydrogen bonds exist between the loop and the receptor, which can be explained by the spatial shifting of this structure comparing to the 11-nt motif. The RMSD value between them based both loop and receptor backbone atoms is rather large (>3 Å), agreeing with their overall structural difference. But if superposition is based only on the top half of the receptor that corresponds the portion above the adenosine platform in an 11-nt motif (Figure 16, blue), they have a RMSD value of 1.345 Å out of 84 backbone atom pairs, which is considerably smaller, implying that the specific conformation of the 11-nt motif has been partially achieved, although not perfectly.

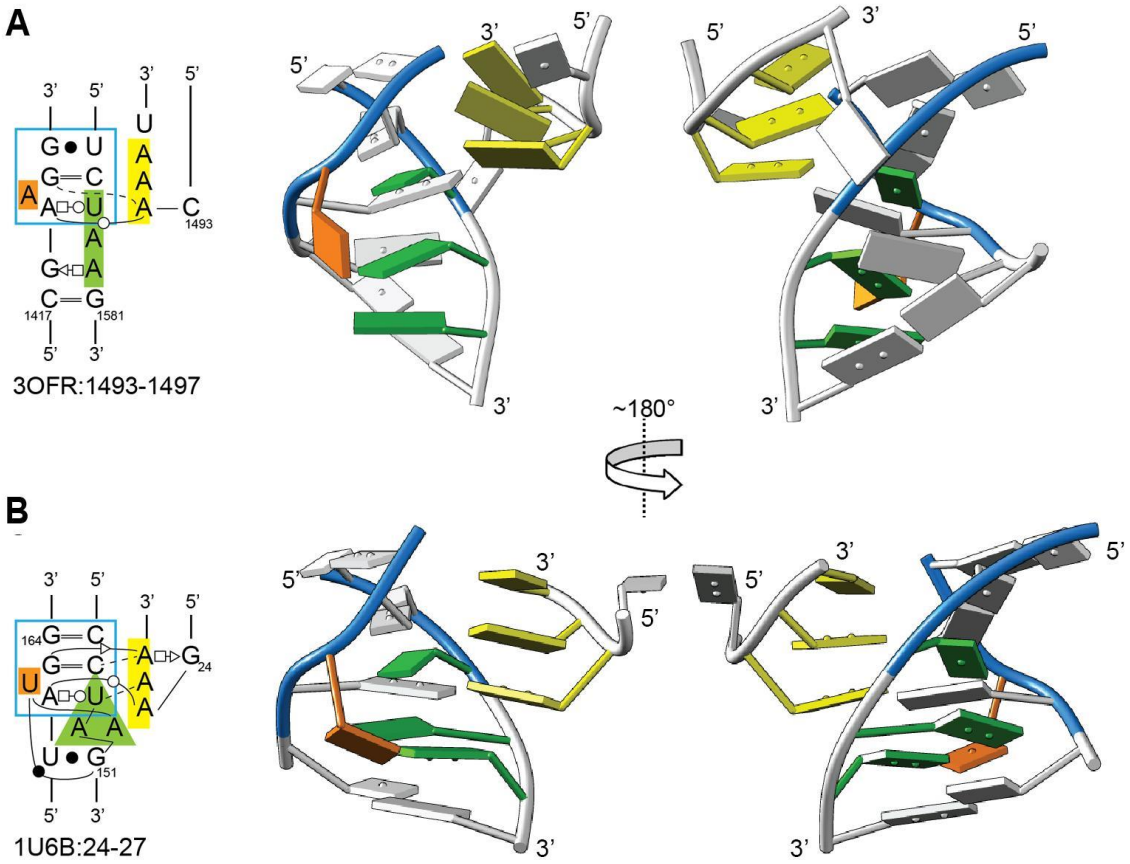


Figure 16: A Class IV structure displays some common features to standard GNRA/11-nt interaction.

This Class IV structure 3OFR:1493-1497 (A) is compared to an example from Class I (B) at both secondary and tertiary levels. Tertiary structures are shown in two rotational views. Comparable regions are highlighted in different colours and they correspond in both of the secondary and tertiary structures.

Another interesting example is a UNCG-tetraloop that interacts with a receptor (1VQO:1770-1773). In this structure, the two-basepair receptor itself is composed of three individual elements: two single nucleotides come together and forms a non-Watson-Crick basepair with a nucleotide from the other strand (Appendix B). The UUCG-tetraloop, on the other hand, contacts the

assembled receptor by forming of two hydrogen bonds. The second loop nucleotide, U1771, forms a U-A WC basepair with the receptor nucleotide A1885, and also forms a single hydrogen bond with nucleotide U2017, which contacts A1885 by a single hydrogen bond. This is so far the only example of an UNCG-tetraloop interaction being revealed in this study. Since exactly the same configuration has been found in *E. coli* LSU (Table 3), it can be inferred that it is a possible way for UNCG-tetraloops to anchor to a distant receptor, although the single hydrogen bonding connection itself is not strong.

2.3.7 Sequence-Structure Correlations

2.3.7.1 Strongly Supported Correlations

A useful purpose of a structural analysis is to identify sequence-structure correlations that may allow the prediction of three-dimensional structures from sequence. In this data set, the most obvious and strongest correlation is the GNRA/11-nt interaction, in which both the loop and receptor sequences are highly conserved (GAAA/CCUAAG-UAUGG). It can be assumed that any loop-receptor will have this specific GAAA/11-nt configuration as long as their sequences are identical or near-identical to the current Class I members. Another strongly correlated example is Class II subclass 3, which has been specified to be the GNAR-tetraloop motifs in SRP-RNAs [101]. Of this type, the GNAR-tetraloop interacts with the very end region of a stem-loop structure in a conserved sequence of CGRAAG (Figure 10B). The third case of sequence-structure correlation is for Class II subclass 5, in which a tetraloop contacting a flipped-out single base from the double helix receptor. The interaction has been proven being present across different classes of group II introns, and thus, it is reasonable to assume that any sequences at

this position in different individual introns will have a similar configuration of the loop and the receptor (see section 2.3.8 for details) [103].

2.3.7.2 GNRA-Tetraloop Specific Analyses

Since the GNRA-tetraloop is one of the major types of RNA tetraloop and an important building block in RNA tertiary structures, it is of interest to look through all GNRA-tetraloops in this study to pull out any possible relationships. If solely considering the loop sequence regardless the overall configuration, there are 31 structures having a GNRA-tetraloop, and 30 of them interact with a double helix receptor, forming into the standard Class II geometry (Appendix C). Although there are some minor variations of the base positions, these 30 GNRA-tetraloops have good agreement in terms of backbone conformation (max RMSD<1.2 Å based on 48 backbone atoms) (Figure 17). Meanwhile, the only one exception out of the 31 is 3OFR:124-127, which is in Class IV subclass 1. It has a GAAA-tetraloop but does not form the common GNRA geometry and have a larger RMSD value (>2.5 Å based on 48 backbone atom) (Figure 17). Of this tetraloop, the first two loop nucleotides (G124 and A125) stretch out along the backbone, and the last two loop nucleotides (A126 and A127) fit into the locations of R3 and A4 as in other common GNRA conformations. The unusual pathway of its loop backbone makes itself looking like an "outsider" while being superposed onto the rest 30 normal GNRA-tetraloops. Despite the one exception, the general trend is that a GNRA-tetraloop will form the conformation seen in Class II.

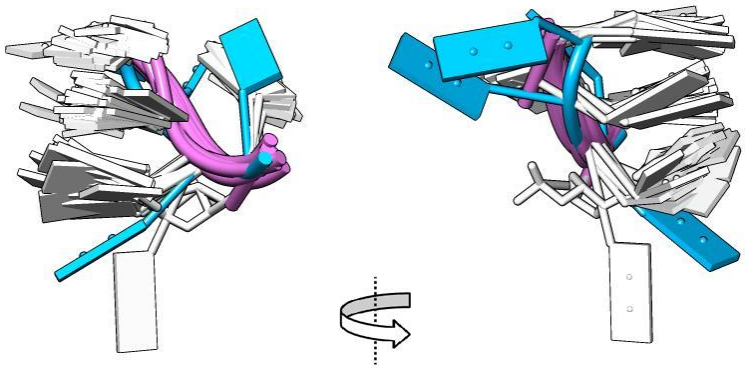


Figure 17: Two rotational views of 31 superposed GNRA tetraloops.

This superposition is based on 48 backbone atoms. Out of 31 GNRA tetraloops, 30 (pink) can be considered identical and are all from either Class I or Class II. The only one exception is 3OFR:124-127 from Class IV (blue).

There are eight possible combinations for a GNRA sequence. Out of the 31 GNRA-tetraloops mentioned above, 15 of them have the sequence of GAAA, 5 have GUGA, 4 have GCGA, 3 have GCAA, 2 have GUAA, 1 has GAGA, and the last has GGAA, while the sequence of GGGA is not seen in any of the structures (Appendix C). One If looks at their receptor sequences, the four GCGA-tetraloops are always associated with AG-CU receptor in Class II subclass 1, while the five GUGA-tetraloops interact with different sequences, but all have Class II subclass 1 structures as well. In contrast, the GAAA-tetraloop interacts with receptors in multiple classes, including Class I, Class II subclass 1, 2 and 4, and Class IV. Consequently, when excluding loops interacting with the 11-nt motif, the GNRA-tetraloops generally have receptors of a double helical structure. Although some types of sequences are more frequent, there is no strong correlation between the sequence of the loop and receptor.

Another attempt was made to look for sequence correlations starting from the receptors. For this, Class IV and Class II subclass 5 structures were excluded because they have irregular receptors, leaving 29 GNRA/helix structures for analysis. The most frequent receptor sequence is AG-CU, with 10 examples. Other common sequences are GG-CC (7), CC-GG (2), CG-CG (2) and AG-CG (2); the rest of receptor sequences do not have multiple examples (Appendix C). The 10 structures having receptor sequence of AG-CU are all from Class II subclass 1, but do not correspond to any specific loop sequence. Likewise, the other receptor sequences do not have specific correlation either. The frequency of occurrence seems not to correlate to certain types of interactions either. While being the most common sequence combination, the GAAA/GG-CC interaction can be seen in multiple structural classes. Consequently, the receptor sequence seems to be completely random, regardless the frequency of appearance. Instead, the exact same sequence may form into absolutely different configurations. A good example showing this is the GAAA/GG-CC interaction coming from Class II subclass 1.1 (individual), Class II subclass 4 and Class IV subclass 1. Three representatives from these classes were selected and superposed (Figure 18). While their receptors composed by two G-C basepairs are identical, the difference among the hydrogen bonding patterns caused the dramatic disagreement of the loop structures.

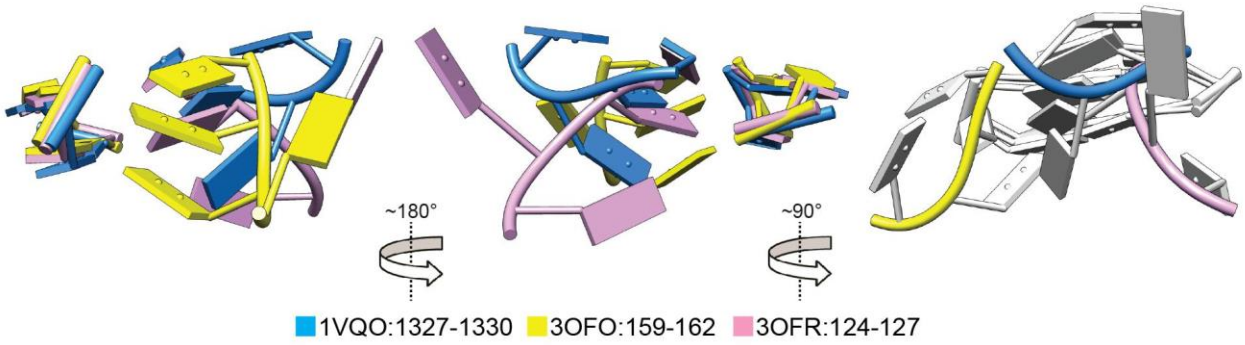


Figure 18: Three rotational views of three distinct structures with the same GAAA/GG-CC sequence.

This superposition is based on 48 backbone atoms of the receptor. Both the loop and the receptor can be seen in the left and middle views; the loops are rotated in the front in the right view while all nucleotides are white except the loop backbones.

It has been reported that GUGA-tetraloops tend to interact with AG-CU receptors, and GUAA-tetraloops tend to interact with GG-CC receptors [64, 104, 105]. Considering the 29 GNRA-tetraloops, they are consistent with the previous correlation, but inexactly. An eight-nucleotide correlation that involves all loop and receptor nucleotides from all the 29 structures was presented by the WebLogo profiles (http://weblogo.berkeley.edu/logo.cgi) [106, 107]. The first sequence motif, GYGA (Y=C/U), has a strong correlation with an A-U basepair (positions "6" and "10") together with a weaker correlation with a G-C basepair (positions "7" and "9") (Figure 19B). The second motif, GNAA, shows the correlation with a G-C basepair (positions "7" and "9"); although the other basepair in the receptor is not fully supported, the existence of G-C basepairs at position "6" and "10" is still visible (Figure 19C).

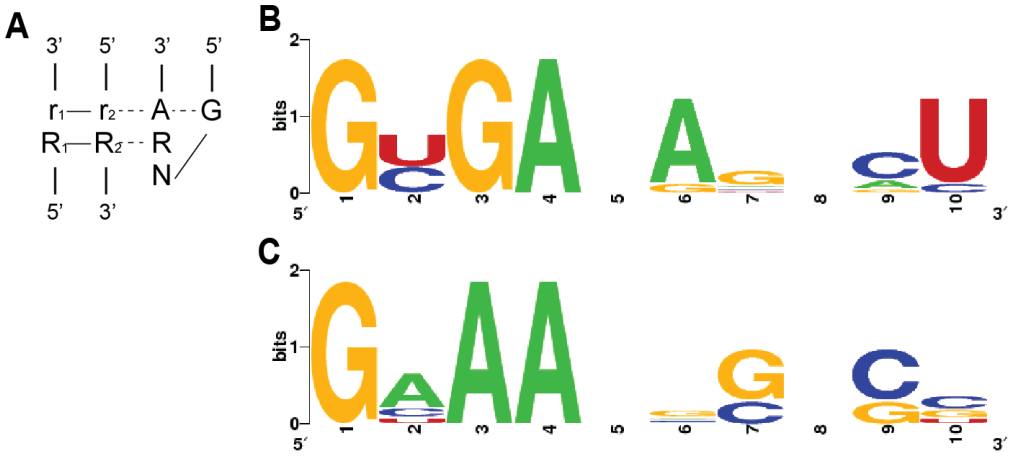


Figure 19: WebLogo profiles for GYGA and GNAA tetraloop interactions.

A) *A representative secondary of a tetraloop-receptor interaction.* ***B, C)*** *Weblogo profiles for interactions that have a loop sequence of GYGA (B) and GNAA (C). Numbers on the x-axis correspond to nucleotides in the secondary structure: 1-4 correspond to GNRA, while 6, 7, 9 and 10 correspond to R_1, r_1, r_2 and R_2 respectively.*

2.3.7.3 The Standard-GNRA-Like Geometry

Interestingly, a few loop structures form a conformation that resembles the standard GNRA-tetraloop even though their sequences do not match the four-nucleotide composition of GNRA. This is the case for the tetraloops UCAA, GAAC and GNAG in Class II subclasses 1 (individual), 3 and 5 respectively (Figure 11B). The second example is for non-tetraloops from Class III (Figure 13), and the third example is the Class II NTLs (Figure 9). Excluding some of the NTLs, most structures have their backbone pathways following that of a standard GNRA-tetraloop, but their base positions differ from class to class due to the different hydrogen bonds formed between loop and receptor. Nonetheless, the recurrence of the GNRA-tetraloop-like configuration explains why this type of interaction forms the majority across all tetraloops. A

further indication is the possibility to utilise the variety of both GNRA and GNRA-like conformations in phylogenetic among RNA species or covariation studies, since the maintenance of a certain configuration is required for an individual RNA molecule to function properly, and the configuration is largely contributed by one or more internal tetraloop-receptor interactions (continued discussion in the next section).

2.3.8 Insight into the Evolution of Tetraloop-Helix Interaction in Group II Introns

The evolution of tetraloop-receptor interactions is always of interest since they play an important role in stabilising RNA tertiary structures. In this section, the evolution of a tetraloop-receptor is considered, using the example of the conserved ζ-ζ' interaction of group II introns.

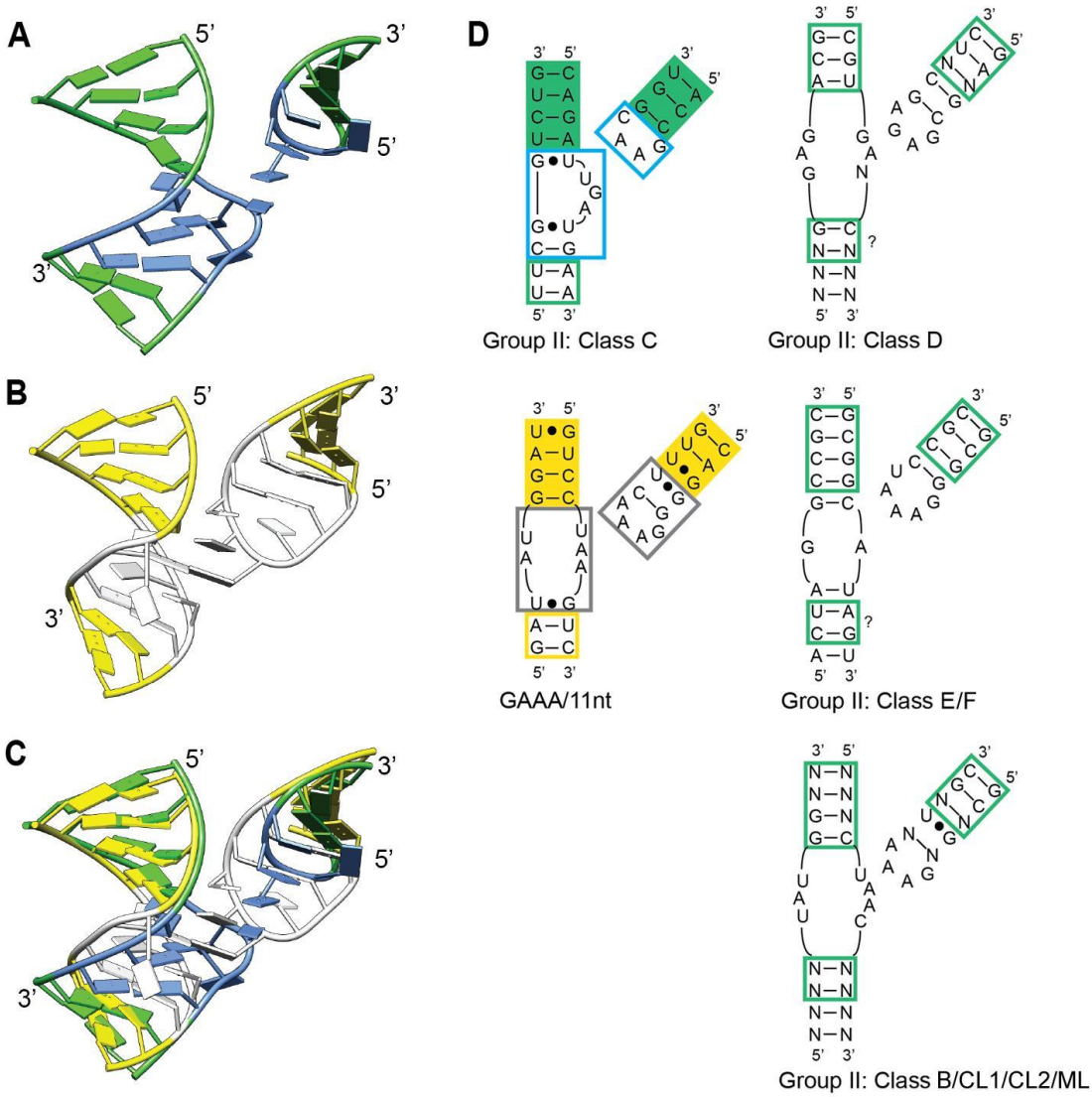


Figure 20: The comparison between the ζ-ζ′ interaction of group II introns and Class I GNRA/11-nt motif interactions.

A, B) *Extended tertiary structures of examples from Class II subclass 5 (A) and Class I (B).* ***C)*** *A superposition of A) and B) based on 168 backbone atoms in the shaded boxes shown in panel D).* ***D)*** *Secondary structures of A) and B), together with several consensus sequences from other group II intron ORF-based classes. Both shaded and open green boxes indicate potential superposable regions across group II introns. Group II class A is excluded due to a small class size (n=3).*

Group II introns are known as large ribozymes, which can splice themselves out of the flanking exons (see 3.1.1 for more details). A functional group II intron has six domains (DI-DVI), and the process of self-splicing requires the correct folding of all these domains, especially domain V, which is the catalytic core. There are several distant connections across domains are crucial to the folding; one of these is the ζ-ζ' interaction that brings DI and DV into proper positions, which has been demonstrated by experimental data (Figure 20) [108]. Although the primary sequence varies in different group II intron sequences, the secondary and tertiary arrangements hardly alter [109]. This ζ-ζ' interaction was first identified as a GNRA/11-nt interaction in the yeast group II intron ai5γ [78]. Therefore, the ζ-ζ' interaction is assumed to be in the same conformation as the GNRA/11-nt interaction, regardless of the group II intron. However, the ζ-ζ' interaction in the group II intron from *Oceanobacillus iheyensis*, which is 3IGI:369-372 in this data set, is not a Class I GNRA/11-nt interactions, but instead is Class II subclass 5 (Figure 20).

It is intriguing to consider why there are two completely different configurations for the same structural contact. However, when 3IGI:369-372 is overlaid onto the GAAA/11-nt interaction 2R8S:150-153, a reasonable explanation can be proposed (Figure 20C). The superposition was calculated based on 168 backbone atoms out of 7 basepairs (Figure 20D, shadow boxes in green and yellow). An RMSD value of 1.741 Å infers the maintenance for the global organisation of adjacent structures, while the interacting centre is totally different. It can be seen that instead of there being the strong and normal interaction of a GNRA-11-nt interaction, the Class II subclass 5 type ζ-ζ' interaction takes place by stretching the internal loop receptor out of the double-helical context, thus making the loop nucleotide A2 stack on to it, and fulfilling its function to bring the loop and receptor together.

It would be expected that other ζ-ζ' interactions will preserve the same geometry, but due to the few number of resolved crystal structures, it is not yet possible to directly compare structures from different classes of group II introns. Alternatively, a comparison using the secondary structural information around the ζ-ζ' interaction was done by the consensus sequence in each of the class (Figure 20D). The consensus structures being used here, however, were simply generated manually by looking for nucleotides that appear by the highest frequency at every position. Then, based on the number of nucleotides of the internal loops, similar classes are combined, including classes E/F, and classes B/CL1/CL2/ML. A loop containing a GNRA sequence is confirmed in all of them with occasional mutations; however, for the receptors, the internal loop is composed by different numbers of nucleotides in the different classes. Overall, all currently known bacterial group II introns are assumed to have an *O. iheyensis*-like or GNRA/11-nt arrangement for the ζ-ζ' interaction, but none of them have the consensus sequence of the GNRA/11-nt interaction. Considering the phylogenetic distance between bacteria and yeast, evolution must have played a role causing the formation of the two distinctive arrangements. As a candidate site for covariation analysis, the ζ-ζ' interaction may bring more information about how a local interaction might evolve in order to maintain the functional global conformation.

CHAPTER THREE: FRET-BASED ANALYSIS OF THE GROUP II INTRON LL.LTRB – THE FIRST STEPS

3.1 Introduction

3.1.1 Group II Introns

Group II introns are large ribozymes that have been found in the genomes of bacteria, archaebacteria, the organellar genomes of fungi, algae, protists and plants, but they are not found in nuclear genomes [110-113]. They can self-splice and reinsert their sequences into other genomic location. The splicing of group II introns requires the correct folding of the intron RNA, in which a protein cofactor is necessary under physiological conditions. This protein cofactor is usually encoded within the intron sequence (intron-encoded protein; IEP), and it helps the intron to fold and then maintains the conformation over the entire splicing event. Furthermore, this protein also allows the intron RNA to reverse splice into specific sites within the genome, reverse-transcribe it into DNA, and integrate the DNA sequence into these sites, which is known as retrohoming. However, without the IEP cofactor, group II introns are still capable to self-splice in an environment with high concentrations of both salt and magnesium ions [114-116].

The typical secondary structure of a functional group II intron consists of six distinguishable domains (DI to DVI) connected by a central wheel (Figure 21) [117]. DI is the largest domain and mainly functions as the scaffold, including several exon binding sites (EBS) that are complementary to the intron binding sites (IBS) that reside in the exon sequence [118]. DIII enhances the intron splicing rate [118]. DIV is not required in splicing but contains the open reading frame (ORF) of the IEP [119]. DV is the most conserved domain and it forms the

catalytic core together with DI, and DVI has the bulged adenosine (bulged A) that initiates the splicing event. Based on the variety in RNA secondary structure, group II introns can be divided into three classes: IIA, IIB and IIC [109]; each class has its own structural elements to recognise the flanking exons [120, 121].

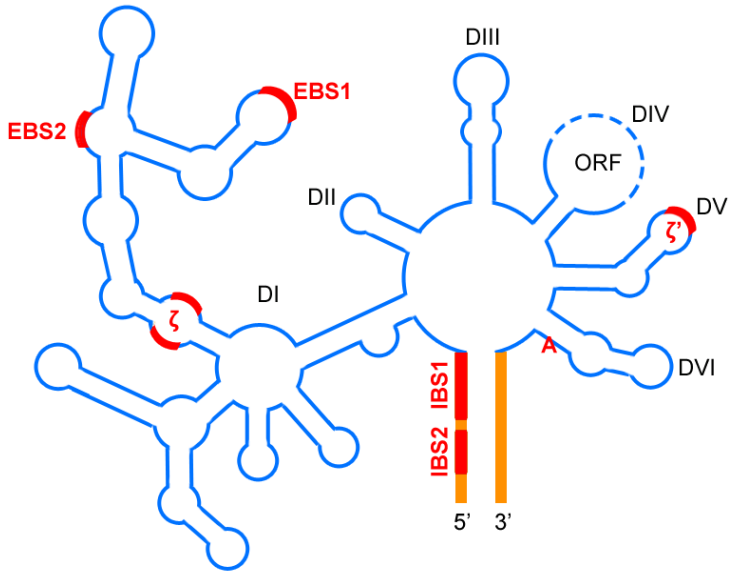


Figure 21: Group II intron consensus secondary structure.

A typical group II intron RNA (blue) folds into six domains (DI-DVI); the dashed line represents where the ORF resides. The bulged A in DVI is shown in red. The bold orange lines represent the flanking exons. Some important sites are highlighted by bold red lines, including IBS1, IBS2, EBS1, EBS2, ζ and ζ'.

The IEP mainly has four domains: a reverse-transcriptase domain (RT), a maturase domain (X), a DNA binding domain (D), and an endonuclease domain (En). These four domains contribute to either splicing or retro-homing. Sometimes, a domain may not be present in a set of introns, and

the relative functions have to be achieved by environmental factors [122, 123]. Based on phylogenetic analysis of the ORF, most identified group II introns can be classified as bacterial A-F, chloroplast-like 1 and 2 (CL1 and CL2), and mitochondrial-like (ML) [124-126]. Interestingly, these ORF based classes are associated with specific RNA based classes mentioned above, for example, the ORF-based Bacterial Class C introns all belong to the RNA-based class IIC [111].

As for group I introns, the splicing mechanism of group II introns involves two transesterification reactions (Figure 22). First, the 2' hydroxyl group of the bulged A in DVI attacks the 5' splice site, and thus the RNA precursor is cleaved into two pieces: the 5' exon and a splicing intermediate lariat made by the intron plus the 3' exon. Second, the 3' splice site breaks, the two pieces of exons that are held by intron-exon base-parings are brought together and ligated, the intron lariat is then released [109]. This reaction, if occurs *in vivo*, is associated with the IEP; after splicing, the intron lariat remain attached to the IEP, and is subsequently reinserted into other locations. If the local tertiary structures are intact, fragments of different intron pieces may still fold and assemble correctly to enable the splicing event. These features, together with some sequence similarity, are like the eukaryotic spliceosome and a few other mobile elements, group II introns are therefore considered to be the ancestor of both spliceosome and non-long terminal repeat (non-LTR) retroelements [111, 127-129].

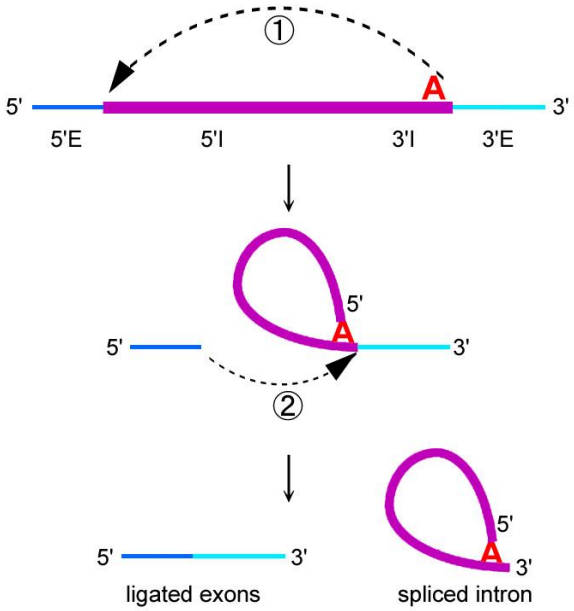


Figure 22: Self-splicing mechanism of group II introns.

The bold and purple line represents the intron RNA. Dark and light blue thinner lines represent its 5' and 3' flank exons. The bulged-A is highlighted in red, and circled numbers of 1 and 2 indicate the two transesterification reactions (Figure adapted from http://webapps2.ucalgary.ca/~groupii/introduction/splicing.html).

3.1.2 Förster Resonance Energy Transfer

Förster (Fluorescence) resonance energy transfer (FRET) is a mechanism involving the non-radioactive energy transfer between two chromophores. The donor chromophore is first excited, and the acceptor chromophore is then excited by the energy that the donor emits. FRET is extremely distance-sensitive, and the FRET efficiency is defined as $E=1/(1+(R/R_0)^6)$, in which E stands for FRET efficiency, R is the donor/acceptor separation distance and R_0 is the Förster distance of the donor/acceptor pair when E equals 50% [53, 130]. To ensure that FRET is

occurring, the emission spectrum of the donor has to overlap with the excitation spectrum of the acceptor (Figure 23). Currently, FRET has been used widely as a tool to measure the dynamical change in distance between two labelled biomolecules within a range of 1-10 nm.

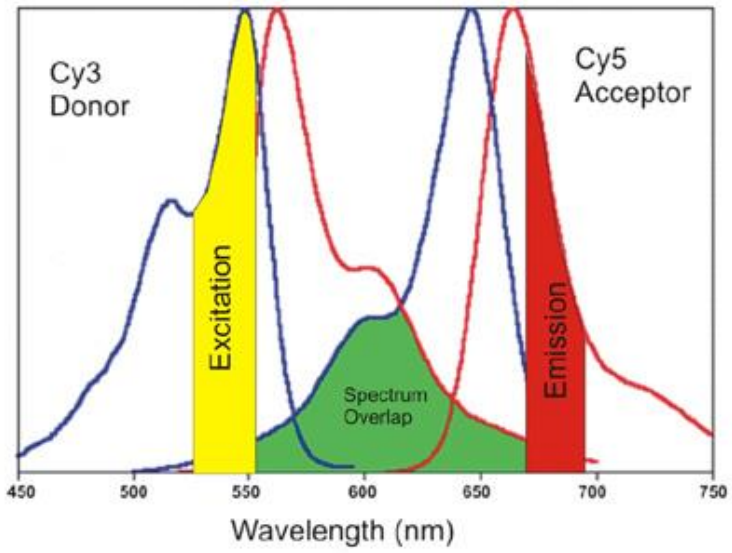


Figure 23: Requirements of fluorophores for FRET to occur.

FRET only occurs when the emission wavelength of the donor and the excitation wavelength of the acceptor overlap. Here, Cy3 and Cy5 are used as donor and acceptor respectively. Their excitation spectrums are in blue and emission spectrums are in red. Green indicates the overlap; yellow and red indicate the possible ranges of excitation and emission when FRET occurs (Figure copied from http://www.biotek.com).

3.1.3 The Aim of Study

In this chapter, a FRET-based experiment was designed to measure distances between selected sites of the group II intron Ll.LtrB in order to test and refine its current tertiary structural model.

The group II intron Ll.LtrB is from *Lactococcus lactis* (sometimes referred to as L.l.I1). It

belongs to the ML class based on the ORF sequence, and is a IIA intron based on the RNA secondary structure. It has many standard properties of group II introns including the splicing and homing mechanisms, and has been studied for years [122, 131-138]. Yet, due to its large size, there is no complete crystal structure describing the entire intron to date. In a previous study, several long-range contact sites for of Ll.LtrB were identified experimentally through cross-linking and were used to build a tertiary structure of this intron. More detailed and accurate spatial arrangements were also found by the crystal structure of a portion of a group II intron from *O. iheyensis* [37, 139]. The *O. iheyensis* intron belongs to class C based on the ORF sequence and is a IIC intron based on the RNA sequence. Because Ll.LtrB is much larger than the crystallized O. iheyensis intron, there are regions present in Ll.LtrB but absent in the *O. iheyensis* intron. This caused some ambiguous arrangements of the modelled tertiary structure of Ll.LtrB, and other sources of experimental information will be required to refine the current model.

A second purpose of the FRET study is to address the dynamic conformational changes that occur before and after the intron undergoes the splicing event. Distance measurements between intron segments can thus contribute to models of conformational changes during splicing. Since this project is at an early stage, only initial experimental data is provided in this chapter. First, different intron constructs are tested for the self-splicing activities *in vitro*, followed by the conjugation of Cy5-hydrazide at the 3' end of intron. A positive control to detect FRET was set up, and the experimental conditions were adjusted. Currently, the main challenge of the experiment is the labelling density on the cpRNA after conjugation. A few possible solutions are discussed and the following experiments are to be extended in future.

3.2 Materials and Methods

3.2.1 Intron Construct and Oligos

Plasmid: The intron construct was built by former members in Zimmerly lab. It contains two identical copies of intron Ll.LtrB sequence in tandem with flanking exons between the tandem copies. Each copy has the intron ORF deleted from positions 1742-3271. The 5' and 3' flanking exons are 215 bp and 190 bp in length respectively. The first copy was inserted into the vector at the restriction site *Sma* I in a pBluescript II KS^+ (pKS^+) vector, and the second copy was inserted at site *EcoR* V. There two copies are linked by a short sequence of 18 bp (Figure 24).

PCR primers: Every pair of PCR primers was designed to amplify a full-length circular-permutated intron DNA (cpDNA). Primers were named by the position number in which a break is created that results the circular permutated full-length intron (Figure 24). The T7 RNA polymerase promoter has been added at the beginning of every forward primer (Table 5).

Annealing Oligos: OligoS (stands for oligo-sense) is complementary to the portion of sequence 2 bp upstream of IBS2. OligoAS (stands for oligo-antisense) has the complementary sequence to OligoS but is 2-nt shorter. OligoS has Cy3 being attached to its 5'-end, while OligoAS has Cy5 being attached to its 3'-end (Table 5).

Table 5: Sequences of PCR primers and other oligos.

Name	Sequence (5' – 3')
219P5	C**TAATACGACTCACTATA**GGTTATGTGTCGATAGAGG
219P3	GAAATTAGAAACTTGCGTTCAGTAAAC
225P5	C**TAATACGACTCACTATA**GTGTCGATAGAGGAAAGTG
225P3	ATAACCGAAATTAGAAACTTGCGTTCA
230P5	C**TAATACGACTCACTATA**GATAGAGGAAAGTGTCTGA
230P3	GACACATAACCGAAATTAGAAACTTGC
OligoS	/Cy3/-CACGATCGACGTGGGTTGCAATCACAAATT
OligoAS	AATTTGTGATTGCAACCCACGTCGATCG-/Cy5/

All primers/ oligos were ordered from IDT. The promoter for T7 RNA polymerase is shown in bold text.

3.2.2 Experiment Set-ups

PCR: Every 50 µl of PCR reaction contained 20 mM Tris-HCl pH 8.4, 50 mM KCl, 5 mM $MgCl_2$, 0.2 mM each dNTP, 1 µM forward primer, 1 µM reverse primer, 1 ng DNA template Ll-450 and 1 µl Taq polymerase. The Taq polymerase was noncommercially prepared and the activity is unknown. Unless otherwise indicated, all reactions followed the same amplification program running in a GeneAmp 9700 thermal cycler (Applied Biosystems):
94 °C, 7 min – [94 °C, 30" – 60 °C, 45" – 72°C, 3min] *35 cycles – 72 °C, 7min – 4 °C hold.
Amplified PCR products were directly used as the template of *in vitro* transcription.

***In vitro* transcription using α-^{32}P-UTP:** Every 20 µl of transcription reaction contained 0.5 µg DNA template, 40 mM Tris-HCl pH 7.5, 12 mM $MgCl_2$, 5 mM spermidine, 50 mM NaCl, 0.5 mM ATP/CTP/GTP, 0.1 mM UTP, 1 µl α-^{32}P-UTP (MP Biomedicals), 5 mM DTT and 1 µl T7

RNA polymerase. The T7 RNA polymerase was noncommercially prepared and the activity was unknown. Samples were incubated at 37 °C for 1-1.5 hours.

***in vitro* transcription with nonradiolabelled UTP:** Every 200 µl of transcription reaction contained 0.5 µg DNA template, 40 mM Tris-HCl pH 7.5, 12 mM $MgCl_2$, 5 mM spermidine, 50 mM NaCl, 0.5 mM NTP, 5 mM DTT and 10 µl T7 RNA polymerase. The T7 RNA polymerase was noncommercially prepared and the activity was unknown. Samples were incubated at 37 °C for 1-1.5 hours. After transcription, samples were heated to 80 °C for 1 min to deactivate T7 RNA polymerase and chilled on ice. In every 100 µl of transcript solution, approximately 170 U of DNase I (Invitrogen) was added and samples were incubated at 37 °C for another 30 minutes to digest the DNA template. RNA transcripts were phenol extracted followed by ethanol precipitation.

Oxidation: Transcript solution was diluted if the final RNA concentration was greater than 0.25 mM. Every 35 µl of reaction contained 10 mM freshly prepared $NaIO_4$ (Sigma-Aldrich) and was incubated at room temperature (RT) for 1-2 hours in dark. To quench the unreacted $NaIO_4$, 1 µl 1.8 M $NaSO_3$ was added to reach a final concentration of 50 mM, and the solution was incubated for an additional 15 minutes. The solution was then directly preceded to the conjugation step.

Conjugation of Cy5: A total of 36 µl oxidised transcript RNA was first mixed with 10 µl of 0.5 M phosphate buffer pH 7.4, and 2 µl of 50 mM Cy5-hydrazide (Lumiprobe) dissolved in DMSO. The sample was incubated at RT for 3 hours followed by another overnight incubation at -20 °C. The Cy5 conjugated sample was purified by passing through a G-25 column followed by ethanol

precipitation.

Calculation of labelling density: The absorbance values of samples at different wavelengths were read on a Beckman DU 530 UV/Vis spectrophotometer. The approximate labelling density was calculated as Equation 1-Equation 3:

Equation 1: Dye:Molecule = $(A_{dye}*\varepsilon_{molecule})/(A_{molecule}*\varepsilon_{dye})$

Equation 2: $A_{molecule} = A_{260} - A_{dye}*CF_{260}$

Equation 3: $\varepsilon_{molecule} = Length*\varepsilon_{base}$

Among these parameters:

A_{dye} is the absorbance of sample at the maximal excitation wavelength (A_{550} for Cy3 and A_{650} for Cy5),

$A_{molecule}$ is the absorbance at 260 nm solely from the nucleic acid; it can be calculated as Equation 2,

ε_{dye} is the extinction coefficient of the dye being conjugated (ε_{Cy3}: 150000, ε_{Cy5}: 250000),

$\varepsilon_{molecule}$ is the extinction coefficient of the DNA/RNA sequence, can be calculated as Equation 3,

A_{260} is the absorbance of sample at 260 nm,

CF_{260} stands for correction factor; $CF_{260(Cy3)}$=0.08, $CF_{260(Cy5)}$=0.05,

Length is for the sample DNA/RNA in unit of basepair (bp) or nucleotide (nt), and

ε_{base} is the average extinction coefficient for different types of nucleotides; $\varepsilon_{DNA\text{-}oligo}$=10000, ε_{RNA}=8250.

Self-splicing: Depending on the concentration, 1-5 µl of radio-labelled sample (>10,000 cpm) was diluted to 25 µl with water. In case of Cy5-labelled RNA, the desired amounts of both RNA and OligoS were first mixed and then divided into aliquots of 25 µl. The 25 µl sample was loaded into a thermal cycler and run by the program:
90°C 1min – 75°C 5min – ramp cool to 43 °C – 43 °C, 25min – 4 °C hold.

When the temperature reached 43 °C, another 25 µl of 2x splicing buffer was added into each sample to reach a final concentration of 40 mM Tris-HCl pH 7.5, 1 M NH_4Cl and 100 mM $MgCl_2$. A buffer without $MgCl_2$ was also used as a negative control. During the incubation step at 43 °C for 25 min, samples were transferred onto ice to stop the reaction. After splicing, the radio-labelled samples were purified by ethanol precipitation and loaded onto a 4% polyacrylamide gel, while the Cy5 labelled samples in aliquot were combined, kept on ice and then transferred to a fluorescence spectrophotometer for scanning.

Oligonucleotide annealing (positive FRET control): 10 µl of each OligoS (10 µM) and OligoAS (10 µM), together with 10 µl of 10x annealing buffer (100 mM Tris-HCl pH 7.5, 500 mM NaCl, 10 mM EDTA) were mixed and diluted to a total volume of 100 µl. Samples were loaded into a thermal cycler followed by running the annealing program:
95 °C, 2min – ramp cool to 25 °C over 45 min – 25 °C, 1.5 min – 4 °C hold.
After annealing, samples were kept on ice and transferred to a fluorescence spectrophotometer for scanning.

FRET measurements: All FRET measurements were performed by a Cary Eclipse fluorescence

spectrophotometer using two scanning methods. One was to scan the excitation wavelength when the emission wavelength was fixed at 670 nm, and the other was to scan the emission wavelength when the excitation wavelength was fixed at 550 nm. The slit length was 5 nm for both methods. Other settings were used as default.

General chemical supplies: All other chemicals were purchased from Sigma-Aldrich, VWR Scientific Products and Fisher Scientific unless otherwise indicated.

3.3 Results and Discussion

3.3.1 Experimental Design, Self-Splicing of Different cpRNAs and Annealing of Oligos

In this experiment, two fluorophores will be attached at specific locations on the intron, and then FRET will be used to measure the distance between the fluorophores. This will test the structure model by seeing whether the two sites are at the expected distance, and it will also allow for detection of conformational changes during the splicing reaction. The RNA to be labelled is a cpRNA of the intron, which will be labelled with Cy5 at the 3' end of the transcript. A Cy3 modification will be located in a DNA oligo that will be annealed to the cpRNA. The intron will be self-spliced under normal in vitro conditions. If the two dyes are brought together during the reaction, FRET can be observed and the distance between the labelled fluorescent dyes on the intron can be calculated (Figure 24A).

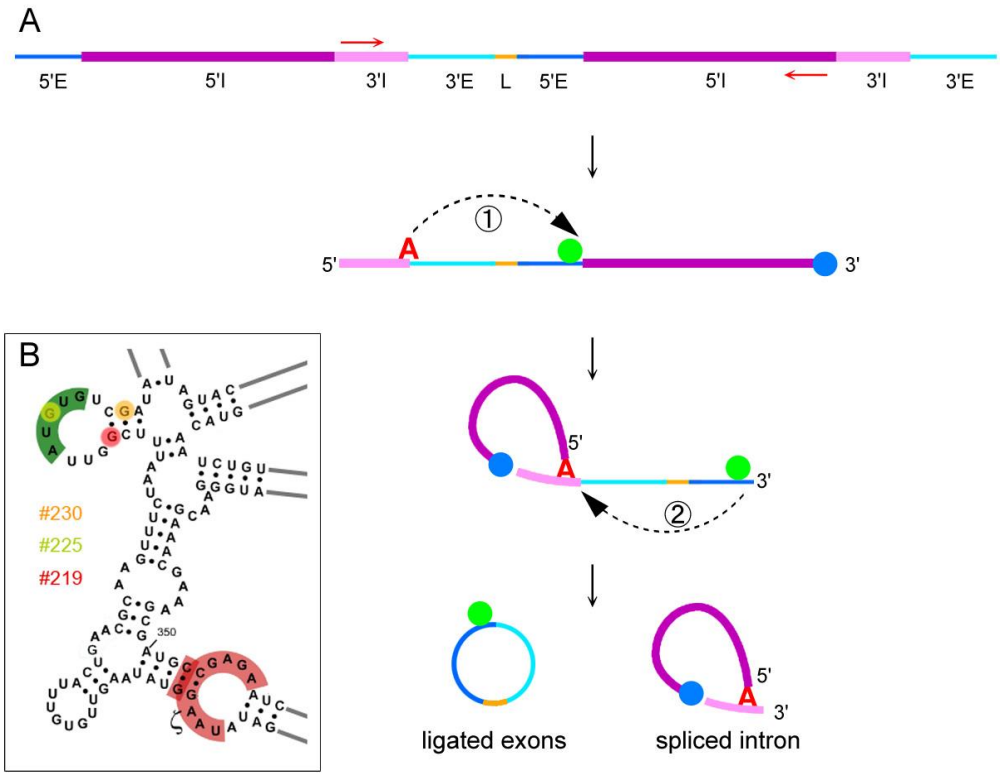


Figure 24: Overview of experiment.

A) *The experiment design. Red arrows indicate one example primer pair that can be used to amplify a circular permutated full-length intron, thus the 5' and 3' portions of the intron are indicated in purple and pink. The flanking 5' and 3' exons are in dark and blue respectively; the linker sequence connecting the two copies is in orange. Cy5 (blue dot) and Cy3 (green dot) will be either conjugated to its 5' end or annealed to the region before its IBS2 site. If the two fluorescent dyes are brought close enough during the intron folding and self-splicing, FRET can be observed. After splicing, the spliced intron and the ligated exons separate and no FRET will be observed.* ***B)*** *The positions at 219, 225 and 230 of Ll.LtrB were chosen to create circular permutations in this study. The EBS2 and ζ regions are highlighted in green and red respectively.*

The site to be tested in the experiment is the IBS2/EBS2 pairing, and will include three circular-permutation positions, 219, 225 and 230. Position 225 is inside the EBS2 sequence while 219 and 230 are adjacent (Figure 24B). The first test in the experiment was to confirm the splicing ability of these cpRNAs. This was done on a small scale with the RNA molecules labelled with

[α-^{32}P]UTP. This construct is different from the wild type intron, because the two flanking exons of the circular-permutated RNA are linked, and so after splicing, the ligated exons will be circular, while the spliced cpRNA will be in a Y-like shape (Figure 24A). The cpRNA-225 construct failed to splice, which may be explained as being due to a collapsed EBS2 loop and interrupted IBS2-EBS2 pairing. The other two constructs spliced properly (Figure 25A).

Because it was not certain whether the oligos would have negative effect on correct intron folding and splicing, another self-splicing experiment was carried out with the addition of excess amounts of OligoS (20 μM) with the other conditions unchanged. Both intron constructs spliced by similar efficiency, indicating that the oligo has little effect on the folding and splicing (Figure 25B).

Finally, to test whether OligoS could anneal to the correct region within the cpRNA, an indirect test was done. OligoS was used as the forward primer in a PCR reaction, together with other reverse primers used in this study. All three PCR amplifications produced DNA fragments at the correct sizes (data not shown). Although these fragments were not confirmed by sequencing, a correct annealing of OligoS, at least on the DNA sequence, could be inferred.

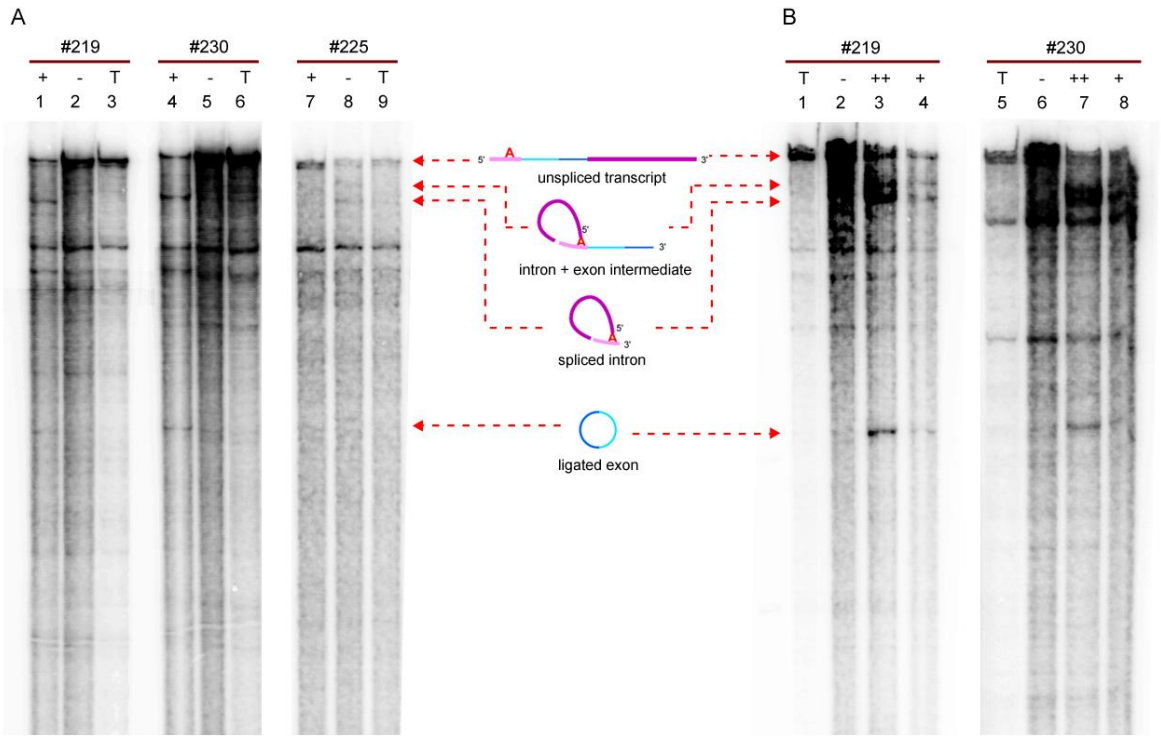


Figure 25: Self-splicing of different cpRNA constructs.

Radio-labelled RNA fragments shown on a 4% polyacrylamide gel. The middle diagram depicts possible products and their appropriate locations on the gel are indicated by red arrows. Any band that is not referred by a red arrow is a non-specific product during transcription. ***A)*** *RNA fragments before and after self-splicing from samples #219 (lane 1-3), #230 (lane 4-6) and #225 (lane 7-9). "+" and "-" indicate whether a positive or negative self-splicing buffer was used, while "T" indicate the RNA transcript.* ***B)*** *RNA fragments before and after self-splicing from samples #219 (lane 1-4) and #230 (lane 5-8) when exceed amounts of oligo DNA existing in the environment. Lanes labelled by "+", "-" and "T" are the same as (A), while "++" indicates the sample was incubated in a positive self-splicing buffer plus an exceed amount of oligo DNA. Results shown in both (A) and (B) have been repeated by multiple independent experiments (n>10).*

3.3.2 The Strategy for Cy5 Conjugation

The method of periodate oxidation was selected to attach the fluorophore tag Cy5 to the RNA molecule. Periodate cleaves adjacent hydroxyl-group containing carbon atoms and thus create two aldehydes groups. The aldehydes groups can spontaneously react with primary amine groups, resulting in conjugated compounds. This method can be applied to RNAs end-labelled at the 3' end, and it has the advantage of being simple and efficient. Also, because DNA nucleotides cannot be oxidised, the chance of getting contamination by labelled DNA nucleotides will be eliminated. Accordingly, the compound Cy5-hydrazide was chosen for conjugation. Although both primary amine- or hydrazide- groups react with aldehydes groups, primary amines will form Schiff bases after the reaction, which can get hydrolysed within a short time and an additional reduction using $NaBH_4$ has to be carried out. However, in case of the hydrazide-group, a rather stable product can be formed by the hydration reaction and thus no reduction is required (Figure 26D) [140].

In initial experiments, samples used in this study had a low density of labelling (<10%, data not shown) under conditions similar to conjugation reactions from the published literature, which were carried for not longer than 3 hours under room temperature (RT) [*e.g.* 140, 141]. To investigate the optimal conjugation condition for this study, different variables were adjusted and labelled samples were measured for labelling density after each single test. Variables introduced included the incubation time, the final concentration of Cy5-hydrazide, the temperature and the total reaction volume. Nonetheless, the labelling density did not increase significantly until a two-step incubation was applied. The first incubation was at RT for 3 hours, followed by the second incubation was under -20 °C overnight. The labelling densities observed after the two-

step incubation could reach 0.6-0.8, while those from samples without the second incubation at -20 °C were generally less than 0.5. This result has been repeated multiple times, but the cause is not clear.

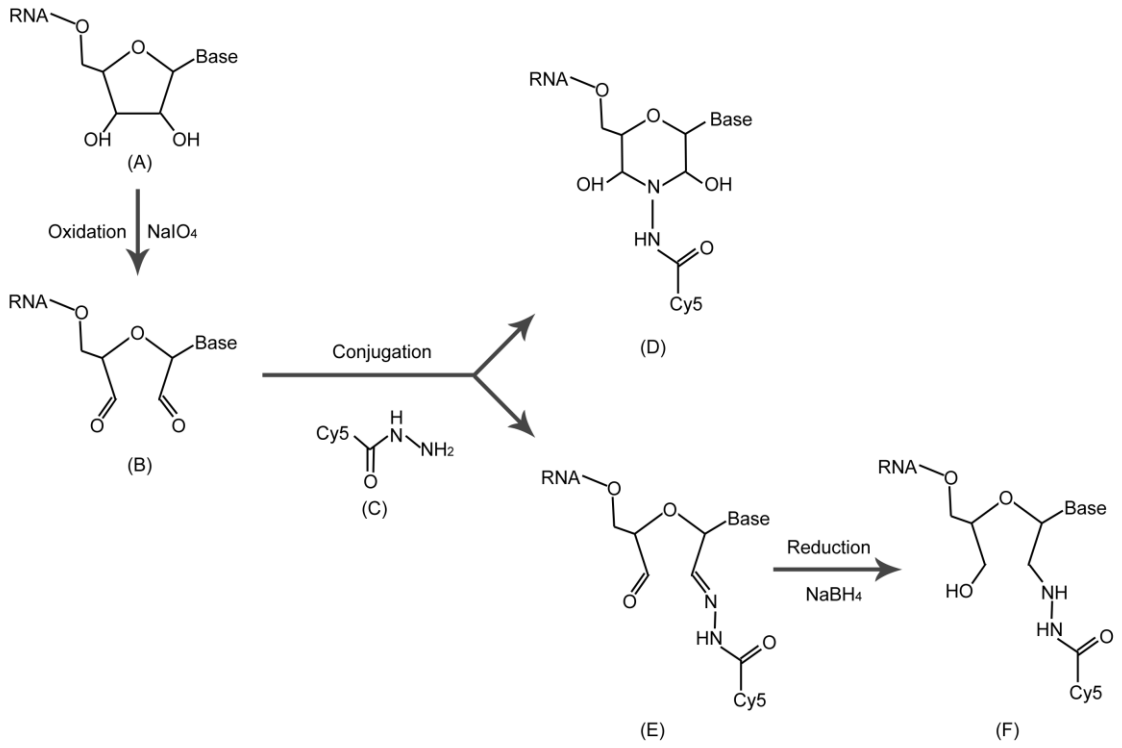


Figure 26: The conjugation between oxidised RNA molecule and hydrazide.

A) *The 3' end of a normal RNA molecule.* ***B)*** *The oxidised 3' end of the RNA molecule.* ***C)*** *Chemical structure of Cy5-hydrazide, where the structure of Cy5 is not shown.* ***D)*** *One possible product of conjugation.* ***E)*** *Another possible product of conjugation, which may be hydrolysed easily.* ***F)*** *The stable product after reducing E).*

3.3.3 Detecting FRET

Because of the low labelling efficiency, a positive control was done to detect FRET between two complementary DNA oligos. Because OligoS-Cy3 and OligoAS-Cy5 are complementary to each other, FRET was expected to be observed when they anneal. Three parallel experiments were prepared with equimolar amounts of each oligo. The first one had only OligoS-Cy3, the second one had only OligoAS-Cy5, and the third had both OligoS-Cy3 and OligoAS-Cy5. Because these oligos had a labelling density of 1.0 (meaning one fluorophore was attached to each oligo), each tube also had equimolar amounts of the dyes. A strong FRET signal could be observed from the tube containing both of Cy3 and Cy5 dyes, whereas other tubes only displayed the normal spectrum for either of the two dyes (Figure 27). The intensities of Cy3 and Cy5 only differ by approximately a factor of 2. This shows the expected result from using Cy5 labelled cpRNA in an experimental sample.

For the experimental sample with a cpRNA, the calculated average labelling density of cpRNA-Cy5 (2μM) was near 0.6 comparing to 1.0 of OligoS-Cy3. Thus, to ensure equal molar amount for both dyes, each single reaction needed 45 μl of cpRNA-Cy5 and 5.5 μl of OligoS-Cy3 (10 μM). Similarly, three parallel experiments were set up, two contained either of the dye attached molecule and the last one contained both of them. Theoretically, similar intensities as the positive control should be observed for both Cy3 and Cy5 regardless whether FRET occurs, but the actual results were rather intriguing to interpret. No FRET was observed from the tube containing both Cy3 and Cy5, and the spectrums were exactly the same as the two negative controls (Figure 28). The intensity of Cy5 (max at 5 a.u.) was dramatically lower than that of Cy3 (max at 70 a.u.). Although it was noticed that the intensity of Cy5-labelled sample decreased by a factor of 2 after

being heated up to 95 °C (data not shown), this about 7 times of difference in dye intensity would still cause the inability to observe FRET.

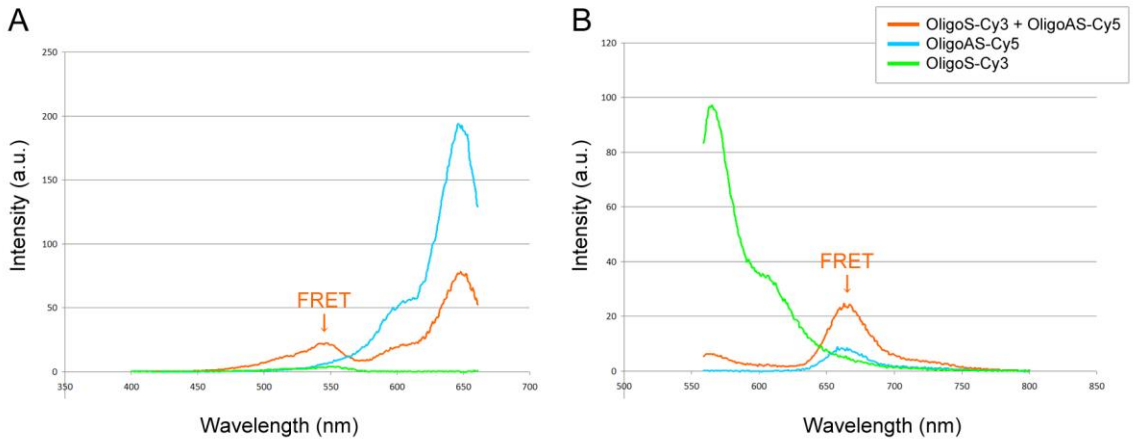


Figure 27: Strong FRET was observed between two complementary DNA oligos.

FRET was observed in the sample with both Cy3 and Cy5 attached oligos after annealing (arrow on orange curves), while negative controls (blue and green) correspond to their normal spectrums. The curves were not scaled and reflect the direct measurements. ***A)*** *The measured excitation spectrum while the emission wavelength was fixed at 670 nm.* ***B)*** *The measured emission spectrum while the excitation wavelength was fixed at 550 nm.*

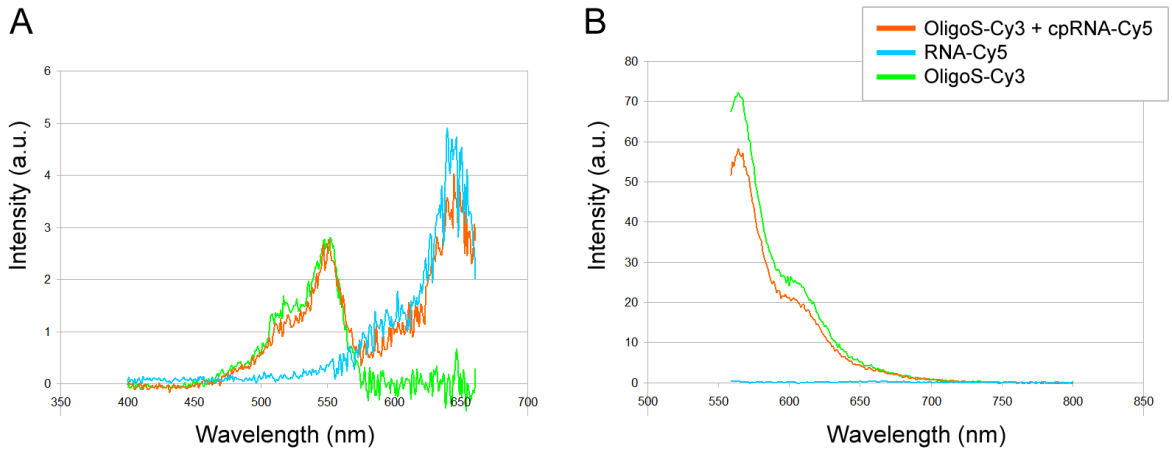


Figure 28: FRET was not observed between sample cpRNA-Cy5 and Cy3 attached oligo, while a large difference between dye intensities exists.

FRET was not observed in either the negative control (blue and green) or the mixture containing cpRNA-Cy5 and oligoS-Cy3 (orange). The curves were not scaled and reflect the direct measurements. ***A)*** *The measured excitation spectrum while the emission wavelength was fixed at 670 nm.* ***B)*** *The measured emission spectrum while the excitation wavelength was fixed at 550 nm.*

A possible reason could be that after conjugation, the sample RNA did not undergo the hydration reaction and form a stable compound (Figure 26D). Instead, it forms a hydrazone linkage and needs to be stabilised by reduction (Figure 26E, F). While being stored at low temperature, this un-reduced compound may be stable and could give a good calculated labelling density, but after the environmental temperature increased, the compound was hydrolysed, causing an observed low intensity. This may then result an overload of Cy3, which could cause photobleaching (destruction of a fluorophore) of Cy5 due to being overexposed in the energy emitted from Cy3. Therefore, the stability of the conjugated RNA molecule after introducing the reducer $NaBH_4$ should be tested next. Moreover, other than the G25 column purification, other methods like high-performance liquid chromatography (HPLC) purification could be applied to further refine

the sample. To reduce the problem of unbalanced amounts of fluorophores, another alternative way to continue this study is to perform single molecule FRET measurements, in which only one pair of donor and acceptor are observed. Finally, in spite of the labelling density of the dye, the efficiency of annealing between the DNA oligo and the RNA sample and how stable the chimeric DNA/RNA complex are to be tested. Nonetheless, the following experiments can only be carried after the confirmation of a more stable cpRNA-Cy5 compound.

In summary, different cpRNA constructs have been confirmed capable to undergo self-splicing, and the addition of excess oligo does not affect the efficiency of self-splicing. The oligo labelled with Cy3 was also confirmed to anneal properly to the cpRNA strand. As for the 3' end-labelling of Cy5 to the cpRNA through conjugation, an acceptable labelling density ranging from 0.6-0.8 was achieved. Finally, a positive FRET control was done and the experimental conditions were adjusted. However, the current problems are: 1) to increase the labelling density for the cpRNA transcript, and 2) to generate a more stable compound after conjugation. Although these results ensure that this experiment can be used to measure the distances between selected sites, problems mentioned here need to be solved first.

CHAPTER FOUR: FINAL SUMMARY

This thesis focused on two topics that were both about RNA tertiary structure. The first topic was mainly about the loop-receptor interactions in RNA structures, while the second topic was more related to obtain experimental data for the refinement of an existing RNA tertiary model.

For the first section, a set of 78 loop-receptor interactions were compiled from a pool of unique resolved RNA crystal structures. Four structural classes together with multiple subclasses were assigned to each of these interactions. The GNRA/11-nt interaction and the GNRA/helical minor groove interaction are the two major types among all these classes, which is consistent with previous studies. In addition, more detailed analyses were done across classes/subclasses, providing more information, such as the possible correlated sequence and structure combination, or the extended variety of interaction between a pair of loop and receptor in different context, as well as a potential interpretation regarding to their evolution. This information may help to build RNA structural models by parallel comparison, or to predict RNA structure from secondary structure. For potential motifs, future experiments may be carried to examine their biological functions, by monitoring effects it may cause after being mutated; or performing covariation analysis to seek for bioinformatic support.

For the second section, a FRET-involved experiment was designed to obtain the dynamic change of group II intron Ll.LtrB during self-splicing. A positive control experiment was used as a reference to determine the dye intensity. A protocol to conjugate Cy5-hydrazyde to the target RNA molecule was developed yet still need to be optimised. Although the experiment was stalled currently by the less-stable Cy5-labelled cpRNA compound, a few possible approaches

were proposed. After examining these approaches, this study can be resumed.

References

1. Crick F. *Central dogma of molecular biology*. Nature. 1970. **227**(5258):561-3.

2. Cech TR, Zaug AJ, Grabowski PJ. *In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence*. Cell. 1981. **27**(3 Pt 2):487-96.

3. Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. *Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena*. Cell. 1982. **31**(1):147-57.

4. Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. *The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme*. Cell. 1983. **35**(3 Pt 2):849-57.

5. Doherty EA, Doudna JA. *Ribozyme structures and mechanisms*. Annu Rev Biophys Biomol Struct. 2001. **30**:457-75.

6. Spirin AS. *The ribosome as an RNA-based molecular machine*. RNA Biol. 2004. **1**(1):3-9.

7. 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**(5481):905-20.

8. Schluenzen F, Tocilj A, Zarivach R, Harms J, Gluehmann M, Janell D, Bashan A, Bartels H, Agmon I, Franceschi F, Yonath A. *Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution*. Cell. 2000. **102**(5):615-23.

9. 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**(6802):327-39.

10. Holley RW, Apgar J, Everett GA, Madison JT, Marquisee M, Merrill SH, Penswick JR, Zamir A. *Structure of a ribonucleic acid*. Science. 1965. **147**(3664):1462-5.

11. Kim SH, Quigley G, Suddath FL, McPherson A, Sneden D, Kim JJ, Weinzierl J, Blattmann P, Rich A. *The three-dimensional structure of yeast phenylalanine transfer RNA: shape of the molecule at 5.5-A resolution*. Proc Natl Acad Sci U S A. 1972. **69**(12):3746-50.

12. Robertus JD, Ladner JE, Finch JT, Rhodes D, Brown RS, Clark BF, Klug A. *Structure of yeast phenylalanine tRNA at 3 A resolution*. Nature. 1974. **250**(467):546-51.

13. Clark BF. *The crystal structure of tRNA*. J Biosci. 2006. **31**(4):453-7.

14. Ladner JE, Jack A, Robertus JD, Brown RS, Rhodes D, Clark BF, Klug A. *Structure of yeast phenylalanine transfer RNA at 2.5 A resolution*. Proc Natl Acad Sci U S A. 1975. **72**(11):4414-8.

15. Krasilnikov AS, Yang X, Pan T, Mondragón A. *Crystal structure of the specificity domain of ribonuclease P*. Nature. 2003. **421**(6924):760-4.

16. Krasilnikov AS, Xiao Y, Pan T, Mondragón A. *Basis for structural diversity in homologous RNAs*. Science. 2004. **306**(5693):104-7.

17. Kazantsev AV, Krivenko AA, Harrington DJ, Holbrook SR, Adams PD, Pace NR. *Crystal structure of a bacterial ribonuclease P RNA*. Proc Natl Acad Sci U S A. 2005. **102**(38):13392-7.

18. Torres-Larios A, Swinger KK, Krasilnikov AS, Pan T, Mondragón A. *Crystal structure of the RNA component of bacterial ribonuclease P*. Nature. 2005. **437**(7058):584-7.

19. Nudler E, Mironov AS. *The riboswitch control of bacterial metabolism*. Trends Biochem Sci. 2004. **29**(1):11-7.

20. Tucker BJ, Breaker RR. *Riboswitches as versatile gene control elements*. Curr Opin Struct Biol. 2005. **15**(3):342-8.

21. Vitreschak AG, Rodionov DA, Mironov AA, Gelfand MS. *Riboswitches: the oldest*

mechanism for the regulation of gene expression? Trends Genet. 2004. **20**(1):44-50.

22. Serganov A, Polonskaia A, Phan AT, Breaker RR, Patel DJ. *Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch*. Nature. 2006. **441**(7097):1167-71.

23. Serganov A, Huang L, Patel DJ. *Structural insights into amino acid binding and gene control by a lysine riboswitch*. Nature. 2008. **455**(7217):1263-7.

24. Garst AD, Héroux A, Rambo RP, Batey RT. *Crystal structure of the lysine riboswitch regulatory mRNA element*. J Biol Chem. 2008. **283**(33):22347-51.

25. Nahvi A, Sudarsan N, Ebert MS, Zou X, Brown KL, Breaker RR. *Genetic control by a metabolite binding mRNA*. Chem Biol. 2002. **9**(9):1043.

26. Cheah MT, Wachter A, Sudarsan N, Breaker RR. *Control of alternative RNA splicing and gene expression by eukaryotic riboswitches*. Nature. 2007. **447**(7143):497-500.

27. Bauer G, Suess B. *Engineered riboswitches as novel tools in molecular biology*. J Biotechnol. 2006. **124**(1):4-11.

28. Dixon N, Duncan JN, Geerlings T, Dunstan MS, McCarthy JE, Leys D, Micklefield J. *Reengineering orthogonally selective riboswitches*. Proc Natl Acad Sci U S A. 2010. **107**(7):2830-5.

29. Verhounig A, Karcher D, Bock R. *Inducible gene expression from the plastid genome by a synthetic riboswitch*. Proc Natl Acad Sci U S A. 2010. **107**(14):6204-9.

30. Perlman PS, Butow RA. *Mobile introns and intron-encoded proteins*. Science. 1989. **246**(4934):1106-9.

31. Lambowitz AM. *Infectious introns*. Cell. 1989. **56**(3):323-6.

32. Dujon B. *Group I introns as mobile genetic elements: facts and mechanistic speculations--a*

review. Gene. 1989. **82**(1):91-114.

33. Belfort M. *Phage T4 introns: self-splicing and mobility*. Annu Rev Genet. 1990. **24**:363-85.

34. Saldanha R, Mohr G, Belfort M, Lambowitz AM. *Group I and group II introns*. FASEB J. 1993. **7**(1):15-24.

35. Valadkhan S. *snRNAs as the catalysts of pre-mRNA splicing*. Curr Opin Chem Biol. 2005. **9**(6):603-8.

36. Vicens Q, Cech TR. *Atomic level architecture of group I introns revealed*. Trends Biochem Sci. 2006. **31**(1):41-51.

37. Toor N, Keating KS, Taylor SD, Pyle AM. *Crystal structure of a self-spliced group II intron*. Science. 2008. **320**(5872): 77-82.

38. Pley HW, Flaherty KM, McKay DB. *Three-dimensional structure of a hammerhead ribozyme*. Nature. 1994. **372**(6501):68-74.

39. Scott WG, Finch JT, Klug A. *The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage*. Cell. 1995. **81**(7):991-1002.

40. Martick M, Scott WG. *Tertiary contacts distant from the active site prime a ribozyme for catalysis*. Cell. 2006. **126**(2):309-20.

41. Ferré-D'amaré AR, Rupert PB. *The hairpin ribozyme: from crystal structure to function*. Biochem Soc Trans. 2002. **30**(Pt 6):1105-9.

42. Zwieb C, Eichler J. *Getting on target: the archaeal signal recognition particle*. Archaea. 2002. **1**(1):27-34.

43. Shan SO, Walter P. *Co-translational protein targeting by the signal recognition particle*. FEBS Lett. 2005. **579**(4):921-6.

44. Ulbrandt ND, Newitt JA, Bernstein HD. *The E. coli signal recognition particle is required*

for the insertion of a subset of inner membrane proteins. Cell. 1997. **88**(2):187-96.

45. Abell BM, Pool MR, Schlenker O, Sinning I, High S. *Signal recognition particle mediates post-translational targeting in eukaryotes*. EMBO J. 2004. **23**(14):2755-64.

46. Schuenemann D, Gupta S, Persello-Cartieaux F, Klimyuk VI, Jones JD, Nussaume L, Hoffman NE. *A novel signal recognition particle targets light-harvesting proteins to the thylakoid membranes*. Proc Natl Acad Sci U S A. 1998. **95**(17):10312-6.

47. Howard EI, Sanishvili R, Cachau RE, Mitschler A, Chevrier B, Barth P, Lamour V, Van Zandt M, Sibley E, Bon C, Moras D, Schneider TR, Joachimiak A, Podjarny A. *Ultrahigh resolution drug design I: details of interactions in human aldose reductase-inhibitor complex at 0.66 A*. Proteins. 2004. **55**(4):792-804.

48. Smyth MS, Martin JH. *x ray crystallography*. Mol Pathol. 2000. **53**(1):8-14.

49. Lukavsky PJ, Puglisi JD. *RNAPack: an integrated NMR approach to RNA structure determination*. Methods. 2001. **25**(3):316-32.

50. Chou JJ, Li S, Klee CB, Bax A. *Solution structure of Ca(2+)-calmodulin reveals flexible hand-like properties of its domains*. Nat Struct Biol. 2001. **8**(11):990-7.

51. Fürtig B, Buck J, Manoharan V, Bermel W, Jäschke A, Wenter P, Pitsch S, Schwalbe H. *Time-resolved NMR studies of RNA folding. Biopolymers*. 2007. **86**(5-6):360-83.

52. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. *The Protein Data Bank. Nucleic Acids Res*. 2000. **28**(1):235-42.

53. Roy R, Hohng S, Ha T. *A practical guide to single-molecule FRET*. Nat Methods. 2008. **5**(6):507-16.

54. Regulski EE, Breaker RR. *In-line probing analysis of riboswitches*. Methods Mol Biol. 2008. **419**:53-67.

55. Strauss B, Nierth A, Singer M, Jäschke A. *Direct structural analysis of modified RNA by fluorescent in-line probing*. Nucleic Acids Res. 2012. **40**(2):861-70.

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

57. Leontis NB, Stombaugh J, Westhof E. *The non-Watson-Crick base pairs and their associated isostericity matrices*. Nucleic Acids Res. 2002. **30**(16):3497-531.

58. Staple DW, Butcher SE. *Pseudoknots: RNA structures with diverse functions*. PLoS Biol. 2005. 3(6):e213.

59. Mans RM, Guerrier-Takada C, Altman S, Pleij CW. *Interaction of RNase P from Escherichia coli with pseudoknotted structures in viral RNAs*. Nucleic Acids Res. 1990. **18**(12):3479-87.

60. Nissen P, Ippolito JA, Ban N, Moore PB, Steitz TA. *RNA tertiary interactions in the large ribosomal subunit: the A-minor motif*. Proc Natl Acad Sci U S A. 2001. **98**(9):4899-903.

61. Doherty EA, Batey RT, Masquida B, Doudna JA. *A universal mode of helix packing in RNA*. Nat Struct Biol. 2001. **8**(4):339-43.

62. Conn GL, Gutell RR, Draper DE. *A functional ribosomal RNA tertiary structure involves a base triple interaction*. Biochemistry. 1998. **37**(34):11980-8.

63. Jaeger L, Michel F, Westhof E. *Involvement of a GNRA tetraloop in long-range RNA tertiary interactions*. J Mol Biol. 1994. **236**(5):1271-6.

64. Costa M, Michel F. *Rules for RNA recognition of GNRA tetraloops deduced by in vitro selection: comparison with in vivo evolution*. EMBO J. 1997. **16**(11):3289-302.

65. Batey RT, Rambo RP, Doudna JA. *Tertiary Motifs in RNA Structure and Folding*. Angew Chem Int Ed Engl. 1999. **38**(16):2326-2343.

66. Woese CR, Winker S, Gutell RR. *Architecture of ribosomal RNA: constraints on the*

sequence of "tetra-loops". Proc Natl Acad Sci U S A. 1990. **87**(21):8467-71.

67. Duszczyk MM, Wutz A, Rybin V, Sattler M. *The Xist RNA A-repeat comprises a novel AUCG tetraloop fold and a platform for multimerization*. RNA. 2011. **17**(11):1973-82.

68. Huang HC, Nagaswamy U, Fox GE. *The application of cluster analysis in the intercomparison of loop structures in RNA*. RNA. 2005. **11**(4):412-23.

69. 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**(8):2342-52.

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

71. Allain FH, Varani G. *Structure of the P1 helix from group I self-splicing introns*. J Mol Biol. 1995. 250(3):333-53.

72. Ennifar E, Nikulin A, Tishchenko S, Serganov A, Nevskaya N, Garber M, Ehresmann B, Ehresmann C, Nikonov S, Dumas P. *The crystal structure of UUCG tetraloop*. J Mol Biol. 2000. **304**(1):35-42.

73. Antao VP, Lai SY, Tinoco I Jr. *A thermodynamic study of unusually stable RNA and DNA hairpins*. Nucleic Acids Res. 1991. **19**(21):5901-5.

74. Leulliot N, Baumruk V, Abdelkafi M, Turpin PY, Namane A, Gouyette C, Huynh-Dinh T, Ghomi M. *Unusual nucleotide conformations in GNRA and UNCG type tetraloop hairpins: evidence from Raman markers assignments*. Nucleic Acids Res. 1999. **27**(5):1398-404.

75. Tuerk C, Gauss P, Thermes C, Groebe DR, Gayle M, Guild N, Stormo G, d'Aubenton-Carafa Y, Uhlenbeck OC, Tinoco I Jr, et al. *CUUCGG hairpins: extraordinarily stable RNA secondary structures associated with various biochemical processes*. Proc Natl Acad Sci U S A. 1988. **85**(5):1364-8.

76. Molinaro M, Tinoco I Jr. *Use of ultra stable UNCG tetraloop hairpins to fold RNA structures: thermodynamic and spectroscopic applications*. Nucleic Acids Res. 1995. **23**(15):3056-63.

77. Pley HW, Flaherty KM, McKay DB. *Model for an RNA tertiary interaction from the structure of an intermolecular complex between a GAAA tetraloop and an RNA helix*. Nature. 1994. **372**(6501):111-3.

78. Costa M, Michel F. *Frequent use of the same tertiary motif by self-folding RNAs*. EMBO J. 1995. **14**(6):1276-85.

79. Lemieux S, Major F. *Automated extraction and classification of RNA tertiary structure cyclic motifs*. Nucleic Acids Res. 2006. **34**(8): 2340-6.

80. Ikawa Y, Nohmi K, Atsumi S, Shiraishi H, Inoue T. *A comparative study on two GNRA-tetraloop receptors: 11-nt and IC3 motifs*. J Biochem. 2001. **130**(2):251-5.

81. Cate JH, Gooding AR, Podell E, Zhou K, Golden BL, Szewczak AA, Kundrot CE, Cech TR, Doudna JA. *RNA tertiary structure mediation by adenosine platforms*. Science. 1996. **273**(5282):1696-9.

82. Ikawa Y, Naito D, Aono N, Shiraishi H, Inoue T. *A conserved motif in group IC3 introns is a new class of GNRA receptor*. Nucleic Acids Res. 1999. **27**(8):1859-65.

83. Young BT, Silverman SK. *The GAAA tetraloop-receptor interaction contributes differentially to folding thermodynamics and kinetics for the P4-P6 RNA domain. Biochemistry*. 2002. **41**(41):12271-6.

84. Ferré-D'Amaré AR, Zhou K, Doudna JA. *A general module for RNA crystallization*. J Mol Biol. 1998. **279**(3):621-31.

85. Geary C, Baudrey S, Jaeger L. *Comprehensive features of natural and in vitro selected*

GNRA tetraloop-binding receptors. Nucleic Acids Res. 2008. **36**(4):1138-52.

86. 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**(16):4755-61.

87. Zhong C, Zhang S. *Clustering RNA structural motifs in ribosomal RNAs using secondary structural alignment*. Nucleic Acids Res. 2012. **40**(3):1307-17.

88. Shen Y, Wong HS, Zhang S, Yu Z. *Feature-based 3D motif filtering for ribosomal RNA*. Bioinformatics. 2011. **27**(20):2828-35.

89. Correll CC, Freeborn B, Moore PB, Steitz TA. *Metals, motifs, and recognition in the crystal structure of a 5S rRNA domain*. Cell. 1997. **91**(5):705-12.

90. Klein DJ, Schmeing TM, Moore PB, Steitz TA. *The kink-turn: a new RNA secondary structure motif*. EMBO J. 2001. **20**(15):4214-21.

91. Clemons WM Jr, Brodersen DE, McCutcheon JP, May JL, Carter AP, Morgan-Warren RJ, Wimberly BT, Ramakrishnan V. *Crystal structure of the 30 S ribosomal subunit from Thermus thermophilus: purification, crystallization and structure determination*. J Mol Biol. 2001. **310**(4):827-43.

92. Tamura M, Holbrook SR. *Sequence and structural conservation in RNA ribose zippers*. J Mol Biol. 2002. **320**(3):455-74.

93. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. *The Protein Data Bank: a computer-based archival file for macromolecular structures*. J Mol Biol. 1977. **112**(3):535-42.

94. Guex N, Peitsch MC. *SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling*. Electrophoresis. 1997. **18**(15):2714-23.

95. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. *UCSF Chimera--a visualization system for exploratory research and analysis*. J Comput Chem. 2004. **25**(13):1605-12.

96. Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wählby A, Jones TA. *The Uppsala Electron-Density Server*. Acta Crystallogr D Biol Crystallogr. 2004. **60**(Pt 12 Pt 1):2240-9.

97. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH. *PHENIX: a comprehensive Python-based system for macromolecular structure solution*. Acta Crystallogr D Biol Crystallogr. 2010. **66**(Pt 2):213-21.

98. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AG, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A, Wilson KS. *Overview of the CCP4 suite and current developments*. Acta Crystallogr D Biol Crystallogr. 2011. **67**(Pt 4):235-42.

99. Emsley P, Lohkamp B, Scott WG, Cowtan K. *Features and development of Coot*. Acta Crystallogr D Biol Crystallogr. 2010. **66**(Pt 4):486-501.

100. Jossinet F, Westhof E. *Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure*. Bioinformatics. 2005. **21**(15):3320-1.

101. Zwieb C. *Recognition of a tetranucleotide loop of signal recognition particle RNA by protein SRP19*. J Biol Chem. 1992. **267**(22):15650-6.

102. Hainzl T, Huang S, Sauer-Eriksson AE. *Structure of the SRP19 RNA complex and implications for signal recognition particle assembly*. Nature. 2002. **417**(6890):767-71.

103. Keating KS, Toor N, Pyle AM. *The GANC tetraloop: a novel motif in the group IIC intron*

structure. J Mol Biol. 2008. **383**(3):475-81.

104. Michel F, Westhof E. *Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis*. J Mol Biol. 1990. **216**(3):585-610.

105. Massire C, Jaeger L, Westhof E. *Phylogenetic evidence for a new tertiary interaction in bacterial RNase P RNAs*. RNA. 1997. **3**(6):553-6.

106. Crooks GE, Hon G, Chandonia JM, Brenner SE. *WebLogo: a sequence logo generator*. Genome Res. 2004. **14**(6):1188-90.

107. Schneider TD, Stephens RM. *Sequence logos: a new way to display consensus sequences*. Nucleic Acids Res. 1990. **18**(20):6097-100.

108. Waldsich C, Pyle AM. *A folding control element for tertiary collapse of a group II intron ribozyme*. Nat Struct Mol Biol. 2007. **14**(1):37-44.

109. Michel F, Umesono K, Ozeki H. *Comparative and functional anatomy of group II catalytic introns--a review*. Gene. 1989. **82**(1):5-30.

110. Bonen L, Vogel J. *The ins and outs of group II introns*. Trends Genet. 2001. **17**(6):322-31.

111. Lambowitz AM, Zimmerly S. *Mobile group II introns*. Annu Rev Genet. 2004. **38**:1-35.

112. Vallès Y, Halanych KM, Boore JL. *Group II introns break new boundaries: presence in a bilaterian's genome*. PLoS One. 2008. **3**(1):e1488.

113. Copertino DW, Hallick RB. *Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns*. Trends Biochem Sci. 1993. **18**(12):467-71.

114. Peebles CL, Perlman PS, Mecklenburg KL, Petrillo ML, Tabor JH, Jarrell KA, Cheng HL. *A self-splicing RNA excises an intron lariat*. Cell. 1986. **44**(2):213-23.

115. Schmelzer C, Schweyen RJ. *Self-splicing of group II introns in vitro: mapping of the branch point and mutational inhibition of lariat formation*. Cell. 1986. **46**(4):557-65.

116. van der Veen R, Arnberg AC, van der Horst G, Bonen L, Tabak HF, Grivell LA. *Excised group II introns in yeast mitochondria are lariats and can be formed by self-splicing in vitro*. Cell. 1986. **44**(2):225-34.

117. Michel F, Ferat JL. *Structure and activities of group II introns*. Annu Rev Biochem. 1995. **64**:435-61.

118. Su LJ, Waldsich C, Pyle AM. *An obligate intermediate along the slow folding pathway of a group II intron ribozyme*. Nucleic Acids Res. 2005. **33**(21):6674-87.

119. Jarrell KA, Dietrich RC, Perlman PS. *Group II intron domain 5 facilitates a trans-splicing reaction*. Mol Cell Biol. 1988. **8**(6):2361-6.

120. Michel F, Costa M, Westhof E. *The ribozyme core of group II introns: a structure in want of partners*. Trends Biochem Sci. 2009. **34**(4):189-99.

121. Pyle AM. *The tertiary structure of group II introns: implications for biological function and evolution*. Crit Rev Biochem Mol Biol. 2010. **45**(3):215-32.

122. Ichiyanagi K, Beauregard A, Lawrence S, Smith D, Cousineau B, Belfort M. *Retrotransposition of the Ll.LtrB group II intron proceeds predominantly via reverse splicing into DNA targets*. Mol Microbiol. 2002. **46**(5):1259-72.

123. Zhong J, Lambowitz AM. *Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription*. EMBO J. 2003. **22**(17):4555-65.

124. Zimmerly S, Hausner G, Wu Xc. *Phylogenetic relationships among group II intron ORFs*. Nucleic Acids Res. 2001. **29**(5):1238-50.

125. Toro N, Molina-Sánchez MD, Fernández-López M. *Identification and characterization of bacterial class E group II introns*. Gene. 2002. **299**(1-2):245-50.

126. Simon DM, Clarke NA, McNeil BA, Johnson I, Pantuso D, Dai L, Chai D, Zimmerly S.

Group II introns in eubacteria and archaea: ORF-less introns and new varieties. RNA. 2008. **14**(9):1704-13.

127. Michel F, Lang BF. *Mitochondrial class II introns encode proteins related to the reverse transcriptases of retroviruses*. Nature. 1985. **316**(6029):641-3.

128. Xiong Y, Eickbush TH. *Origin and evolution of retroelements based upon their reverse transcriptase sequences*. EMBO J. 1990. **9**(10):3353-62.

129. Doolittle RF, Feng DF, Johnson MS, McClure MA. *Origins and evolutionary relationships of retroviruses*. Q Rev Biol. 1989. **64**(1):1-30.

130. Stryer L, Haugland RP. *Energy transfer: a spectroscopic ruler*. Proc Natl Acad Sci U S A. 1967. **58**(2):719-26.

131. Mills DA, McKay LL, Dunny GM. *Splicing of a group II intron involved in the conjugative transfer of pRS01 in lactococci*. J Bacteriol. 1996. **178**(12):3531-8.

132. Mills DA, Manias DA, McKay LL, Dunny GM. *Homing of a group II intron from Lactococcus lactis subsp. lactis ML3*. J Bacteriol. 1997. **179**(19):6107-11.

133. Dunny GM, McKay LL. *Group II introns and expression of conjugative transfer functions in lactic acid bacteria*. *Antonie Van Leeuwenhoek*. 1999. **76**(1-4):77-88.

134. Mohr G, Smith D, Belfort M, Lambowitz AM. *Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences*. Genes Dev. 2000. **14**(5):559-73.

135. Matsuura M, Noah JW, Lambowitz AM. *Mechanism of maturase-promoted group II intron splicing*. EMBO J. 2001. **20**(24):7259-70.

136. Cui X, Matsuura M, Wang Q, Ma H, Lambowitz AM. *A group II intron-encoded maturase functions preferentially in cis and requires both the reverse transcriptase and X domains to*

promote RNA splicing. J Mol Biol. 2004. **340**(2):211-31.

137. Plante I, Cousineau B. *Restriction for gene insertion within the Lactococcus lactis Ll.LtrB group II intron*. RNA. 2006. **12**(11):1980-92.

138. Belhocine K, Mak AB, Cousineau B. *Trans-splicing of the Ll.LtrB group II intron in Lactococcus lactis*. Nucleic Acids Res. 2007. **35**(7):2257-68.

139. Dai L, Chai D, Gu SQ, Gabel J, Noskov SY, Blocker FJ, Lambowitz AM, Zimmerly S. *A three-dimensional model of a group II intron RNA and its interaction with the intron-encoded reverse transcriptase*. Mol Cell. 2008. **30**(4):472-85.

140. Raddatz S, Mueller-Ibeler J, Kluge J, Wäss L, Burdinski G, Havens JR, Onofrey TJ, Wang D, Schweitzer M. *Hydrazide oligonucleotides: new chemical modification for chip array attachment and conjugation*. Nucleic Acids Res. 2002. **30**(21):4793-802.

141. Stevens B, Chen C, Farrell I, Zhang H, Kaur J, Broitman SL, Smilansky Z, Cooperman BS, Goldman YE. *FRET-based identification of mRNAs undergoing translation*. PLoS One. 2012. **7**(5):e38344.

Appendix A: References of the 41 unique RNA crystal structures.

1 Reiter NJ, Osterman A, Torres-Larios A, Swinger KK, Pan T, Mondragón A. (2010) Nature 468:784-9.

2 Krasilnikov AS, Xiao Y, Pan T, Mondragón A. (2004) Science 306:104-7.

3 Kazantsev AV, Krivenko AA, Harrington DJ, Holbrook SR, Adams PD, Pace NR. (2005) Proc Natl Acad Sci U S A 102:13392-7.

4 Krasilnikov AS, Yang X, Pan T, Mondragón A. (2003) Nature 421:760-4.

5 Ye JD, Tereshko V, Frederiksen JK, Koide A, Fellouse FA, Sidhu SS, Koide S, Kossiakoff AA, Piccirilli JA. (2008) Proc Natl Acad Sci U S A 105:82-7.

6 Guo F, Gooding AR, Cech TR. (2004) Mol Cell 16:351-62.

7 Adams PL, Stahley MR, Kosek AB, Wang J, Strobel SA. (2004) Nature 430:45-50.

8 Golden BL, Kim H, Chase E. (2005) Nat Struct Mol Biol 12:82-9.

9 Zhang L, Doudna JA. (2002) Science 295:2084-8.

10 Toor N, Keating KS, Fedorova O, Rajashankar K, Wang J, Pyle AM. (2010) RNA 16:57-69.

11 Chen JH, Yajima R, Chadalavada DM, Chase E, Bevilacqua PC, Golden BL. (2010) Biochemistry 49:6508-18.

12 Rupert PB, Massey AP, Sigurdsson ST, Ferré-D'Amaré AR. (2002) Science 298:1421-4.

13 Chi YI, Martick M, Lares M, Kim R, Scott WG, Kim SH. (2008) PLoS Biol Sep 30;6(9):e234.

14 Shechner DM, Grant RA, Bagby SC, Koldobskaya Y, Piccirilli JA, Bartel DP. (2009) Science 326:1271-5.

15 Xiao H, Murakami H, Suga H, Ferré-D'Amaré AR. (2008) Nature 454:358-61.

16 Klein DJ, Wilkinson SR, Been MD, Ferré-D'Amaré AR. (2007) J Mol Biol 373:178-89.

17 Cochrane JC, Lipchock SV, Smith KD, Strobel SA. (2009) Biochemistry 48:3239-46.

18 Robertson MP, Scott WG. (2007) Science 315:1549-53.

19 Montange RK, Mondragón E, van Tyne D, Garst AD, Ceres P, Batey RT. (2010) J Mol Biol 396:761-72.

20 Gilbert SD, Rambo RP, Van Tyne D, Batey RT. (2008) Nat Struct Mol Biol 15:177-82.

21 Lu C, Smith AM, Fuchs RT, Ding F, Rajashankar K, Henkin TM, Ke A. (2008) Nat Struct Mol Biol. 15:1076-83.

22 Lu C, Ding F, Chowdhury A, Pradhan V, Tomsic J, Holmes WM, Henkin TM, Ke A. (2010) J Mol Biol 404:803-18.

23 Dann CE 3rd, Wakeman CA, Sieling CL, Baker SC, Irnov I, Winkler WC. (2007) Cell 130:878-92.

24 Dixon N, Duncan JN, Geerlings T, Dunstan MS, McCarthy JE, Leys D, Micklefield J. (2010) Proc Natl Acad Sci U S A. 107:2830-5.

25 Edwards AL, Reyes FE, Héroux A, Batey RT. (2010) RNA. 16:2144-55.

26 Gilbert SD, Mediatore SJ, Batey RT. (2006) J Am Chem Soc. 128:14214-5.

27 Huang L, Serganov A, Patel DJ. (2010) Mol Cell. 40:774-86.

28 Serganov A, Huang L, Patel DJ. (2009) Nature 458:233-7.

29 Serganov A, Huang L, Patel DJ. (2008) Nature 455:1263-7.

30 Smith KD, Lipchock SV, Livingston AL, Shanahan CA, Strobel SA. (2010) Biochemistry 49:7351-9.

31 Serganov A, Polonskaia A, Phan AT, Breaker RR, Patel DJ. (2006) Nature 441:1167-71.

32 Thore S, Frick C, Ban N. (2008) J Am Chem Soc 130:8116-7.

33 Dunkle JA, Xiong L, Mankin AS, Cate JH. (2010) Proc Natl Acad Sci U S A. 107:17152-7.

34 Schmeing TM, Huang KS, Kitchen DE, Strobel SA, Steitz TA. (2005) Mol Cell 20:437-48.

35 Kuglstatter A, Oubridge C, Nagai K. (2002) Nat Struct Biol 9:740-4.

36 Hainzl T, Huang S, Sauer-Eriksson AE. (2002) Nature 417:767-71.

37 Wild K, Bange G, Bozkurt G, Segnitz B, Hendricks A, Sinning I. (2010) Acta Crystallogr D Biol Crystallogr 66:295-303.

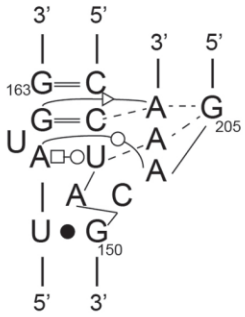
38 Pfingsten JS, Costantino DA, Kieft JS. (2006) Science 314:1450-4.

39 Kieft JS, Zhou K, Grech A, Jubin R, Doudna JA. (2002) Nat Struct Biol. 9:370-4.

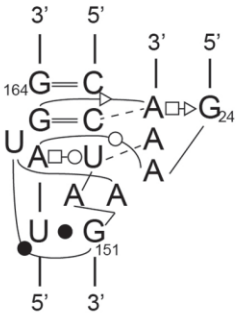
40 Bessho Y, Shibata R, Sekine S, Murayama K, Higashijima K, Hori-Takemoto C, Shirouzu M, Kuramitsu S, Yokoyama S. (2007) Proc Natl Acad Sci U S A 104:8293-8.

Appendix B: Secondary structures of the 78 extracted interactions.

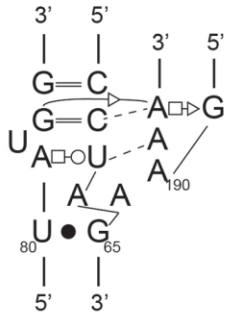
Class I



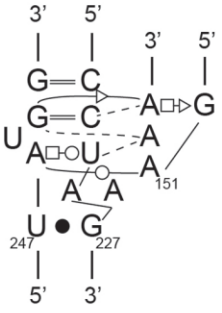
1NBS:205-208



1U6B:24-27

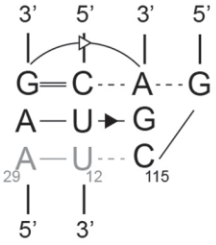


1U6B:189-192

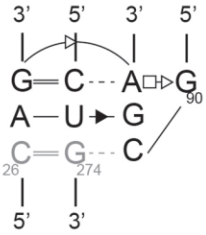


2R8S:150-153

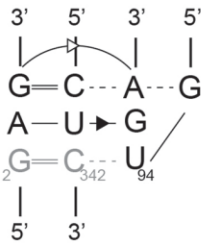
Class II subclass 1.1.1



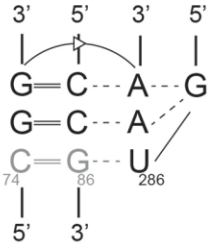
2Z75:114-117



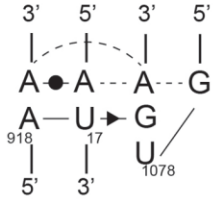
3IGI:90-93



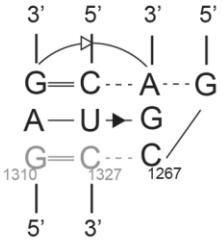
3OK7:93-96



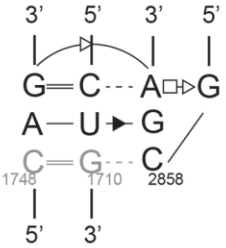
3OK7:285-288



3OFO:1077-1080

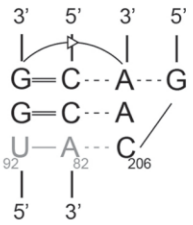


3OFO:1266-1269

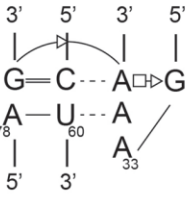


3OFR:2857-2860

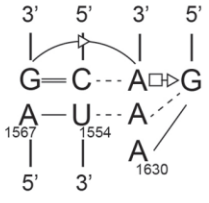
Class II subclass 1.1.2



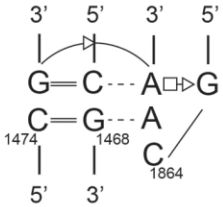
1U9S:205-208



3MXH:32-35

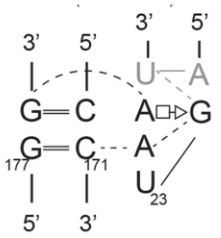


1VQO:1629-1632

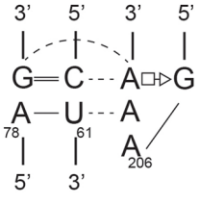


1VQO:1863-1866

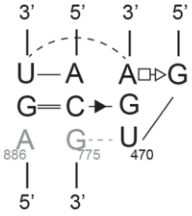
Class II subclass 1.1 (Individual)



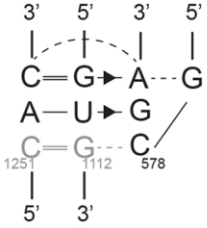
1Y0Q:22-25



1Y0Q:205-208

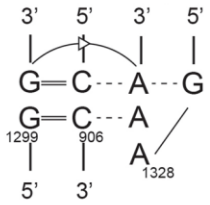


1VQO:469-472



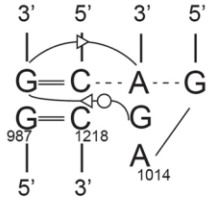
1VQO:577-580

Class II subclass 1 (Individual)

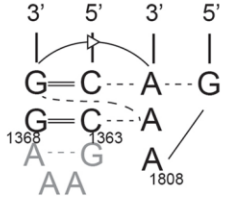


1VQO:1327-1330

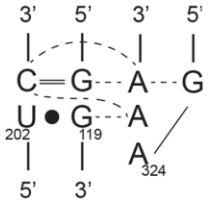
1VQO:2630-2633



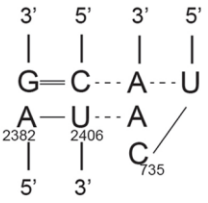
3OFO:1013-1016



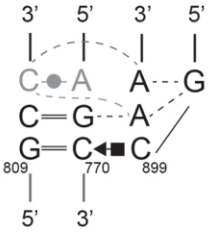
3OFR:1807-1810



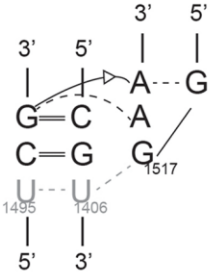
1X8W:323-326



1VQO:734-737

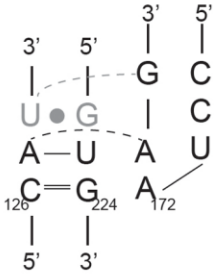


3OFO:898-901

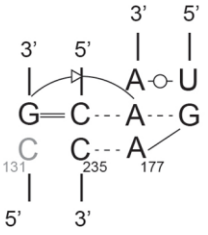


3OFO:1516-1519

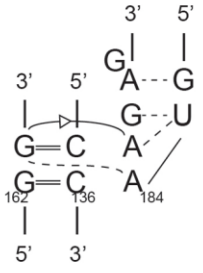
Class II subclass 1 (NTL)



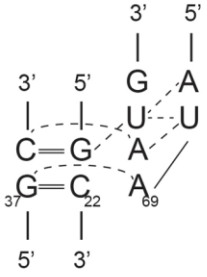
1MFQ:169-174



1NBS:175-179

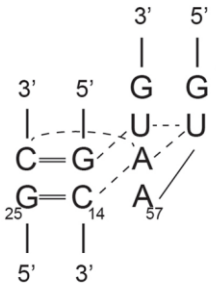


1U9S:182-188

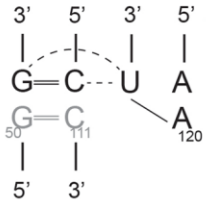


2GDI:67-72

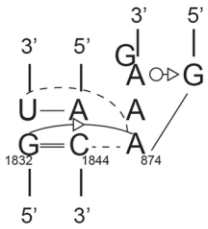
Class II subclass 1 (NTL) -- cont.



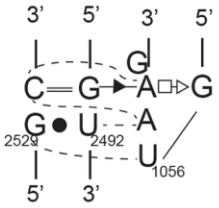
3D2V:55-60



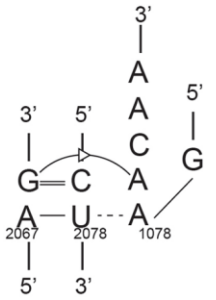
1VQO:119-121



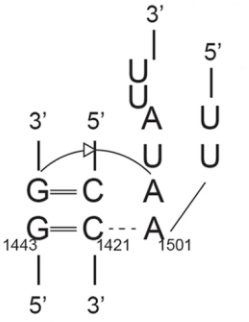
1VQO:873-877



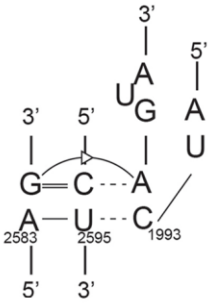
1VQO:1055-1059



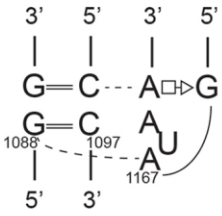
1VQO:1077-1082



1VQO:1499-1506

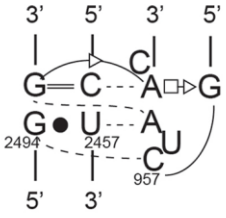


1VQO:1991-1997

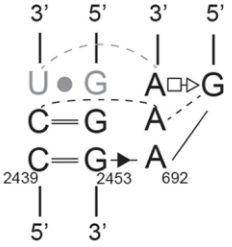


3OFO:1166-1170

Class II subclass 2

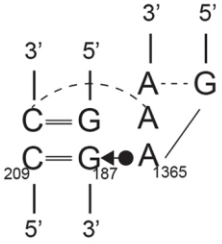


3OFR:956-961

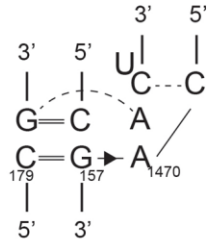


1VQO:691-694

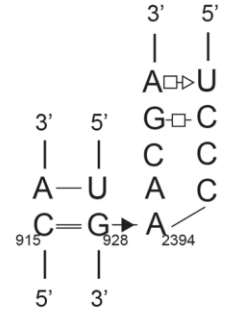
3OFR:630-633



3OFR:1364-1367

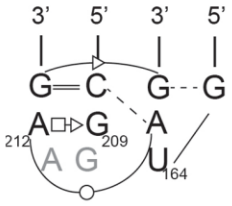


1VQO:1469-1473

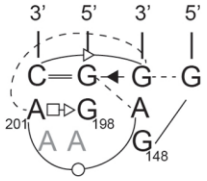


1VQO:2390-2398

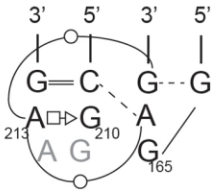
Class II subclass 3



1LNG:163-166



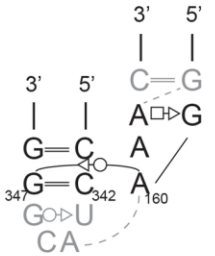
1MFQ:147-150



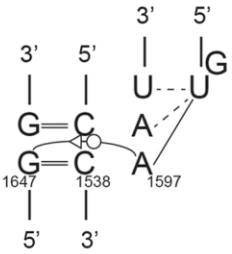
3KTW:164-167

1VQO:2564-2569

Class II subclass 4

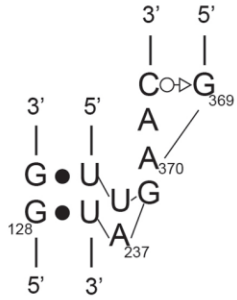


3OFO:159-162



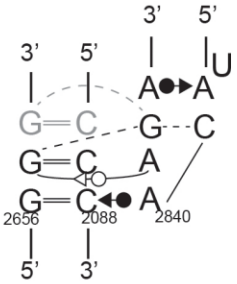
1VQO:1595-1599

Class II subclass 5

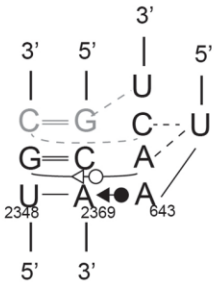


3IGI:369-372

Class III subclass 1

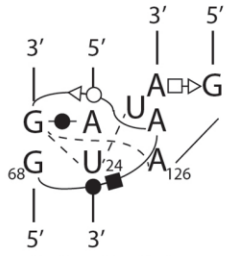


1VQO:2837-2843

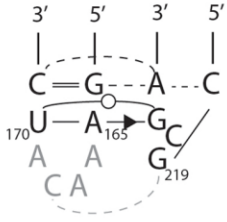


3OFR:642-646

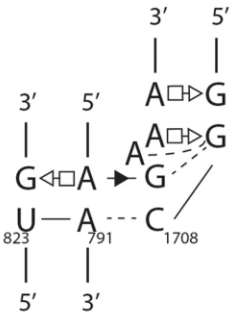
Class III Individual



3DIL:125-129

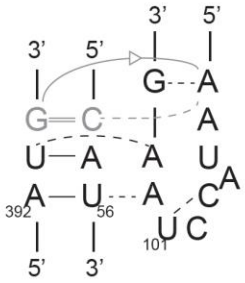


1VQO:218-222

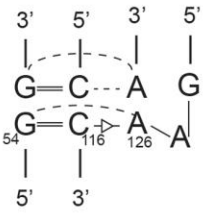


1VQO:1706-1712

Class IV subclass 1

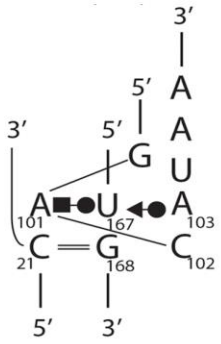


2A64:98-107

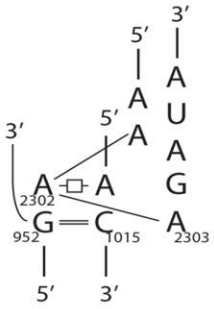


3OFR:124-127

Class IV type 2

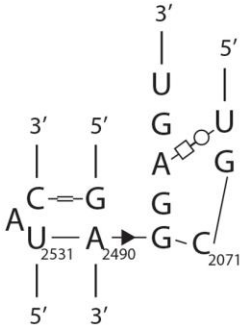


2QBZ:100-106

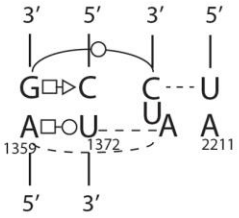


1VQO:2300-2307

Class IV type 3

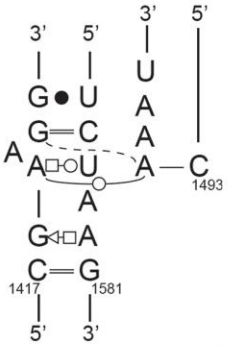


1VQO:2069-2076

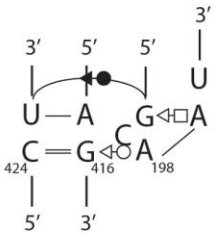


3OFR:2210-2214

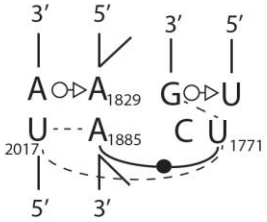
Class IV type 4



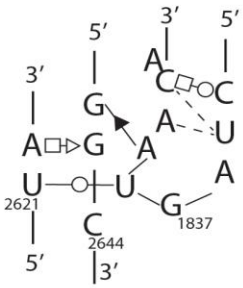
3OFR:1493-1497



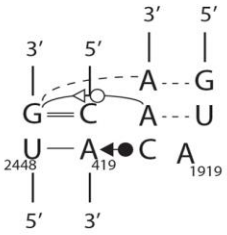
1VQO:196-200



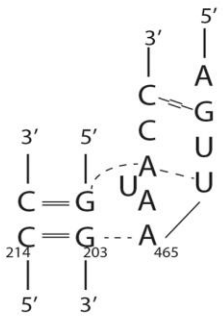
1VQO:1770-1773



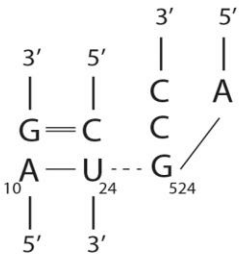
1VQO:1834-1842



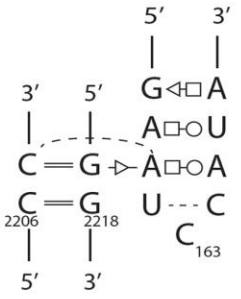
1VQO:1917-1922



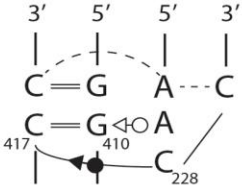
3OFO:461-470



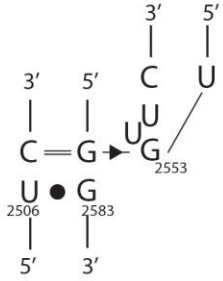
3OFO:523-526



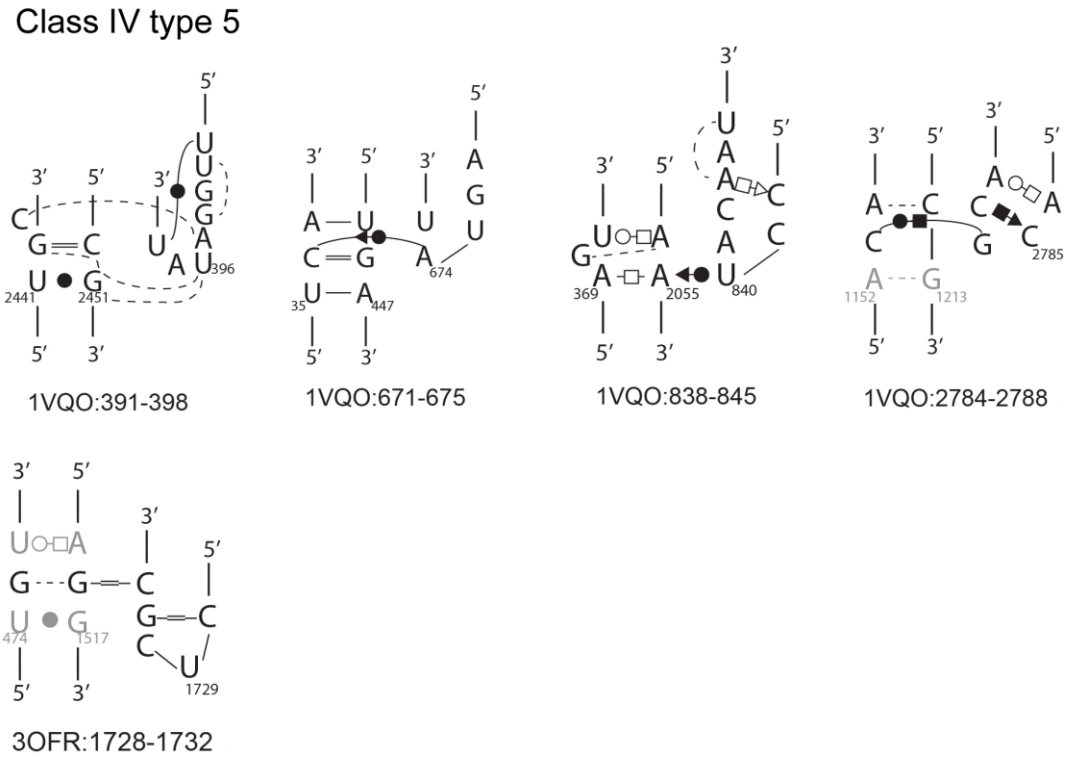
3OFR:159-167



3OFR:226-229



3OFR:2552-2556



The class and subclass (if available) name are shown above the first structure in the class. Gray letters/lines indicate any region that is not a part of the loop or receptor but has interaction(s) with them.

Appendix C: Chart of the loop and receptor sequences.

PDB ID	Loop positions	Receptor positions	Loop sequence	Receptor sequence
Class I				
1NBS	205-208	145-150, 159-163	GAAA	UAUGG/CCUACG
1U6B	24-27	146-151, 160-164	GAAA	UAUGG/CCUAAG
1U6B	189-192	60-65, 80-84	GAAA	UAUGG/CCUAAG
2R8S	150-153	222-227, 247-251	GAAA	UAUGG/CCUAAG
Class II subclass 1.1.1				
2Z75	114-117	10-11, 30-31	GCGA	AG/CU
3IGI	90-93	272-273, 280-281	GCGA	AG/CU
3OK7	93-96	3-4, 340-341	GUGA	AG/CU
3OK7	285-288	75-76, 84-85	GUAA	GG/CC
3OFO	1077-1080	16-17, 918-919	GUGA	AA/AU
3OFO	1266-1269	1311-1312, 1325-1326	GCGA	AG/CU
3OFR	2857-2860	1708-1709, 1749-1750	GCGA	AG/CU
Class II subclass 1.1.2				
1U9S	205-208	80-81, 93-94	GCAA	GG/CC
3MXH	32-35	59-60, 78-79	GAAA	AG/CU
1VQO	1629-1632	1553-1554, 1567-1568	GAAA	AG/CU
1VQO	1863-1866	1467-1468, 1474-1475	GCAA	CG/CG
Class II subclass 1.1 (individual)				
1Y0Q	22-25	170-171, 177-178	GUAA	GG/CC
1Y0Q	205-208	60-61, 78-79	GAAA	AG/CU
1VQO	469-472	773-774, 887-888	GUGA	GU/AC
1VQO	577-580	1110-1111, 1252-1253	GUGA	AC/GU
1VQO	1327-1330	905-906, 1299-1300	GAAA	GG/CC
1VQO	2630-2633	2114-2115, 2470-2471	GUGA	AG/CU
3OFO	1013-1016	987-988, 1217-1218	GAGA	GG/CC
3OFR	1807-1810	1362-1363, 1368-1369	GAAA	GG/CC
Class II subclass 1 (individual)				
1X8W	323-326	118-119, 202-203	GAAA	UC/GG
1VQO	734-737	2382-2383, 2405-2406	UCAA	AG/CU

3OFO	898-901	769-770, 809-810	GCAA	GC/GC
3OFO	1516-1519	1404-1405, 1496-1497	GGAA	CG/CG
Class II subclass 1 (NTL)				
1MFQ	169-174	126-127, 223-224	[GCCU]--AA[GG]	CA/UG
1NBS	175-179	132, 234-235	[UG]--AA[A]	G/C
1U9S	182-188	135-136, 162-163	[GU]--AA[GAG]	GG/CC
2GDI	67-72	21-22, 37-38	[AU]--AA[UG]	GC/GC
3D2V	55-60	13-14, 25-26	[GUA]--AU[G]	GC/GC
1VQO	119-121	50-51, 110-111	[GAA]---U[C]	G/C
1VQO	873-877	1832, 1844	[G]---A[AAG]	G/C
1VQO	1055-1059	2491-2492, 2529-2530	GUAA[G]	GC/GU
1VQO	1077-1082	2067-2068, 2077-2078	[GG]--AA[CAA]	AG/CU
1VQO	1499-1506	1420-1421, 1443-1444	[U]--AA[U]	GG/CC
1VQO	1991-1997	2583-2584, 2594-2595	[AU]--CA[GUA]	AG/CU
3OFO	1166-1170	1088-1089, 1096-1097	[GAU]--AA	GG/CC
3OFR	956-961	2456-2457, 2494-2495	[GCU]--AA[C]	GG/CU
Class II subclass 2				
1VQO	691-694	2439-2440, 2452-2453	GAAA	CC/GG
3OFR	630-633	2401-2403, 2414-2415	GAAA	U[U]C/GG
3OFR	1364-1367	186-187, 209-210	GAAA	CC/GG
1VQO	1469-1473	156-157, 179-180	CAAC[U]	CG/CG
1VQO	2390-2398	915-916, 927-928	[UCCC]-A--[ACGA]	C/G
Class II subclass 3				
1LNG	163-166	208-209, 212-213	GUAG	AG/CG
1MFQ	147-150	197-198, 201-202	GGAG	AC/GG
3KTW	164-167	209-210, 213-214	GGAG	AG/CG
1VQO	2564-2569	2695-2696, 2699-2670	[GC]--AG[AA]	AG/CG
Class II subclass 4				
3OFO	159-162	341-342, 347-348	GAAA	GG/CC
1VQO	1595-1599	1537-1538, 1647-1648	[G]UAAU	GG/CC
Class II subclass 5				
3IGI	369-372	128-129, 234-238	GAAC	-/G

Class III subclass 1				
1VQO	2837-2843	2087-2088, 2656-2657	[UAC]--AA[GA]	GG/CC
3OFR	642-646	2348-2349, 2368-2369	UAAC[U]	UG/CA
Class III individual				
3DIL	125-129	23-24, 68-69	GAA[U]A	GG/AU
1VQO	218-222	164-165, 170-171	CG[C]GA	UC/GA
1VQO	1706-1712	790-791, 823-824	GCG[A]A	UG/AA
Class IV subclass 1				
2A64	98-107	55-56, 392-393	[AAUACCU]**AA**[G]	AU/AU
3OFR	124-127	54-55, 115-116	[GA]**AA**	GG/CC
Class IV type 2				
2QBZ	100-106	21, 167-168	[G]**ACA**[UAA]	C-/UG
1VQO	2300-2307	952, 1014-1015	[A]**AA**[GAU]	G-/AC
Class IV type 3				
1VQO	2069-2076	2490, 2531	[UGC]**G**[GAGU]	U/A
3OFR	2210-2214	1359-1360, 1371-1372	[UA]**A**[U]**C**	AG/CU
3OFR	1493-1497	1418-1421, 1577-1580	[C]**A**[AAU]	A[A]G/CU
Class IV type 4				
1VQO	196-200	415-416, 424-425	[UA]**A**[C]**G**	CU/AG
1VQO	1770-1773	1829, 1885, 2017-2018	[U]**U**[CG]	UA/AA
1VQO	1834-1842	2621-2622, 2642-2643	[CUAG]**UA**[ACA]	UA/G-
1VQO	1917-1922	418-419, 2448-2449	[GUA]**CA**[A]	UG/CA
3OFO	461-470	202-203, 214-215	[AGUU]**A**[AU]**A**[CC]	CC/GG
3OFO	523-526	11-12, 22-23	[A]**G**[CC]	GU/GC
3OFR	159-167	2206-2207, 2217-2218	[AUACCU]**A**[AG]	CC/GG
3OFR	226-229	409-410, 417-418	[C]**CAA**	CC/GG
3OFR	2552-2556	2507, 2581-2582	[U]**G**[UUC]	C/G
Class IV type 5				
1VQO	391-398	2441-2442, 2450-2451	[UUGG]**AUA**[U]	UGC/-CG
1VQO	671-675	36, 446	[AGU]**A**[U]	UC/GA
1VQO	838-845	1369-1371, 2054-2055	[CC]**U**[ACAAU]	A/A
1VQO	2784-2788	1153, 1213	[AC]**G**[CA]	CA/C-
3OFR	1728-1732	1516	[CUCG]**C**	G/G

Receptor sequences are written as "left-strand/right-strand" as in Appendix B. Nucleotides in

square brackets indicate loop nucleotides that do not agree with the GNRA-tetraloop geometry. Dashes indicate "missing" nucleotides in a normal GNRA-tetraloop. Bold text indicate any interacting loop nucleotide from Class IV that is equivalent to a GNRA-tetraloop.

Publication and Copyright

Wu L, Chai D, Fraser ME, Zimmerly S.

Structural variation and uniformity among tetraloop-receptor interactions and other loop-helix interactions in RNA crystal structures.

PLoS One. 2012;7(11):e49225. doi: 10.1371/journal.pone.0049225. Epub 2012 Nov 9.