**Technische Universität München**
Lehrstuhl für biomolekulare NMR-Spektroskopie
Department Chemie

# Structural and functional evolution of the alternative splicing factor LS2 from *Drosophila melanogaster*

Ashish Ashok Kawale
München, 2017

# Technische Universität München

Lehrstuhl für biomolekulare NMR-Spektroskopie
Department Chemie


# Structural and functional evolution of the alternative splicing factor LS2 from *Drosophila melanogaster*


Ashish Ashok Kawale



Vollständiger Abdruck der von der Fakultät für Chemie der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigten Dissertation.

Vorsitzender: Prof. Dr. Bernd Reif
Prüfer der Dissertation: Prof. Dr. Michael Sattler
Prof. Dr. Dierk Niessing

Die Dissertation wurde am 10.01.2017 bei der Technischen Universität München eingereicht und durch die Fakultät für Chemie am 08.02.2017 angenommen.

# Declaration

I hereby declare that parts of this thesis will be published in due course:

Ashish Ashok Kawale, Matthew Taliaferro, Hyun-Seo Kang, Christoph Hartmueller, Ralf Stehle, Arie Geerlof, Christopher Burge, Donald Rio, Michael Sattler. “Elucidating structural evolution of a tissue-specific alternative splicing factor”. (in preparation)

Ashish Ashok Kawale, Ralf Stehle, Michael Sattler. “Biophysical characterization of splicing factor dU2AF50 from *D. melanogaster*”. (in preparation)

# Summary

RNA metabolism is regulated by the interaction of RNA-binding proteins (RBPs) with regulatory RNA elements. RBPs often consists of multiple domains, which are connected by flexible linkers, mediating interactions with nucleic acid elements in order to execute their biological functions. Several studies show that dynamic and cooperative interactions exhibiting domain-domain synergy can be a decisive factor in the regulation of a biological process. Recent progress in structural biological methods enable us to dissect many biologically important protein-RNA interactions at the molecular level.

Alternative splicing is an essential process for eukaryotic gene expression and accounts for the diversity of the proteome while retaining a constant set of genes. This process is orchestrated by a large number of *trans*-acting splicing factors and their interactions with *cis*-regulatory RNA elements. The net combinatorial effect of positive and negative signals to promote or inhibit alternative splicing determines the efficiency of splicing. Hence, in order to regulate this tightly controlled splicing process in a tissue-specific or developmental-stage-specific manner, many alternative splicing factors and their cognate splicing regulatory elements have evolved. Apart from serving as binding sites for *trans*-regulating splicing factors, the ability of splicing regulatory elements to adopt distinct secondary structures adds another degree of regulation. Though a variety of alternative splicing factors, i.e. RBPs and their interactions with *cis*-regulatory RNA elements have been identified, how these RBPs modulate their specificities in order to acquire distinct and diverse functions, is yet unclear.

The *Drosophila melanogaster* LS2 protein is an interesting example for studying evolutionary aspects of alternative splicing regulation. The *Drosophila* U2AF heterodimer, comprised of a 50 kDa large subunit (dU2AF50) and a 38 kDa small subunit (dU2AF38), is essential to recognize the 3’ splice site for constitutive splicing. LS2 has arisen from a gene duplication event of dU2AF50 and is preferentially expressed in the testes of *Drosophila* species. LS2 promotes alternative splicing by preventing dU2AF50 from binding to the polypyrimidine tracts. Despite sharing high sequence similarity (55% identity and 75% similarity), LS2 and dU2AF50 exhibit very different RNA binding specificities (recognition of guanosine-rich and pyrimidine-rich RNAs, respectively), which enables them to subsequently repress or activate target pre-mRNA splicing, respectively. The enrichment of LS2 target transcripts in testes suggests its possible role in testes function, gamete production, and cellular regulation. However, the underlying molecular mechanisms have not been well studied.

This thesis reports a structural and functional analysis of LS2 RNA binding domains along with interdomain linker and their interactions with cognate guanosine-rich RNA sequences. An integrated structural biology approach is employed by combining NMR-spectroscopy with small angle X-ray scattering (SAXS) experiments in solution. The study is complemented by biophysical and biochemical assays. To compare the structural features the dU2AF50 RNA binding domains and their interactions with poly-U RNA are studied, revealing differences between the LS2 and dU2AF50 splicing factors in Drosophila.

Chapter 1 contains an introduction to the biological background of the alternative splicing process as well as the role of the U2AF splicing factor and outlines the evolution of LS2 from

dU2AF50 followed by the role of G-quadruplex RNAs in splicing regulation. The basic principles of NMR and its application to study the structure and dynamics of proteins as well as RNAs are also presented. Chapter 2 summarizes the materials and methods used for molecular biology, biochemical and structural studies.

In Chapter 3, key findings of the thesis are reported. Chapter 3.1.1 describes the structural as well as biophysical characterization of LS2 RNA binding domains (RRM1 and RRM2) and the interdomain linker. Solution NMR data show that both RRM domains of LS2 adopt a canonical βαββαβ topology. Novel and unusual structural features were observed, by an interaction of the RRM1-RRM2 linker with RRM2, and the identification of a novel α-helical region in the LS2-specific linker residues. In Chapter 3.1.2 various guanosine-rich RNAs are investigated to assess their ability to form G-quadruplex folds in the presence of monovalent cations, as potential ligands for LS2. Previously reported high-affinity guanosine-rich RNA ligands for LS2, that were derived from SELEX ('GGX' motif) experiments, were tested and optimized for biochemical and NMR studies. Comprehensive NMR analysis of a 21mer guanosine-rich RNA revealed that it adopts a uniform conformation with dimeric, parallel and three planar topology.

Chapter 3.1.3 describes the interaction studies of LS2 RRM domains and various guanosine-rich ligands. The data show that LS2 recognizes a G-quadruplex RNA structure. The specificity of this interaction involves mainly RRM2, with likely additional contributions from the helical region in the RRM1-RRM2 linker. The LS2 G-quadruplex interaction reported is novel as so far RRM domains are best known for binding to single-stranded nucleic acid sequences. The results show that LS2 RRM2 provides the specificity for recognizing the 21mer RNA G-quadruplex conformation. The information from protein-RNA interaction studies is used to generate LS2 RRM1,2 mutants with impaired RNA binding activity to analyze the contributions of the RNA interaction for functional activity *in vivo*.

Chapter 3.2 reports the characterization of dU2AF50 RNA binding domains (RRM1,2) and their interactions with poly-U RNA. NMR chemical shift-based secondary structure and backbone dynamics data show that both RRM domains adopt a canonical topology. NMR titration data show that both RRM domains participate in RNA binding via conserved RNP sites. On the other hand, ITC studies show that protein-RNA forms 1:1 complex, with a dissociation constant ($K_d$) value similar to that of U2AF65 RRM1,2 and U9 interaction. Altogether, the data indicate that the structural and RNA-binding properties of dU2AF50 are consistent with its human ortholog U2AF65.

In summary, results presented in this thesis report the structures of the RRM domains of the alternative splicing factor LS2 and their interactions with a G-quadruplex RNA structure. Novel structural features associated with the RRM domains and sequence variations in key residues that mediate RNA contacts in LS2 and its paralog U2AF50 explain the drastically different RNA binding preferences for the two proteins. The divergent evolution of the two gene products, originated from gene duplication in *Drosophila* species, presents an intriguing example of paralog proteins for the adaptation of tissue-specific functions.

# Zusammenfassung

Der RNS Metabolismus wird durch die Interaktion von RNS-binden Proteinen (RBPs) mit regulatorischen RNS-Elementen reguliert. RBPs bestehen häufig aus mehreren Domänen, die durch flexible Linker verbunden sind, die sowohl spezifische als auch unspezifische Wechselwirkungen mit Nukleinsäure-Elementen vermitteln, um ihre biologischen Funktionen auszuführen. Mehrere Studien zeigen, dass dynamische und kooperative Wechselwirkungen mit Domänen-Synergie ein entscheidender Faktor für die Regulierung eines biologischen Prozesses sein können. Jüngste Fortschritte in struktur-biologischen Methoden ermöglichten es uns, viele biologisch wichtige Protein-RNS-Interaktionen auf molekularer Ebene zu studieren.

Das alternative Spleißen ist ein essentielles Verfahren für die eukaryotische Genexpression und trägt der Vielfalt des Proteoms bei, während es einen konstanten Satz von Genen beibehält. Dieser Prozess wird durch eine große Anzahl von transaktiven Spleiß Faktoren und deren Wechselwirkungen mit *cis*-regulatorischen RNS-Elementen koordiniert. Der Kombinatorische Effekt der positiven und negativen Signale, die alternativen Spleißen fördern oder hemmen, bestimmt die Effizienz des Spleißens. Daher haben sich, um diesen eng kontrollierten Spleiß Prozess gewebespezifisch oder entwicklungs-stadien-spezifisch zu regulieren, viele alternative Splicing-Faktoren und deren zugehörige Spleiß-Regulationselemente entwickelt. Abgesehen davon, dass sie als Bindungsstellen für transregulierende Spleiß Faktoren dienen, fügt die Fähigkeit der regulatorischen Spleiß-Elemente, unterschiedliche Sekundärstrukturen anzunehmen, eine weitere Regulierungs-Ebene hinzu. Obwohl eine Vielzahl von alternativen Spleiß Faktoren, d.h. RBPs und ihre Wechselwirkungen mit cis-regulatorischen RNS-Elementen, identifiziert wurden, ist noch unklar wie diese RBPs ihre Spezifitäten modulieren, um unterschiedliche und verschiedene Funktionen zu erlangen.

Das *Drosophila* LS2 Protein ist ein interessantes Beispiel für den evolutionären Aspekt der Regulation von alternativem Splicing. Das *Drosophila* U2AF Heterodimer, welches aus einer großen 50 kDa Untereinheit (dU2AF50) und einer kleinen 38 kDa Untereinheit (dU2AF38) besteht, ist essentiell für die Erkennung der 3‘-Splicing Position im konstitutiven Splicing. LS2 ist durch Genduplikation des dU2AF50 Gens entstanden und ist hauptsächlich in den *Drosophila* Hoden exprimiert. LS2 ermöglicht alternatives Splicing, indem es die Interaktion zwischen dU2AF50 und der Polypyrimidinsequenz verhindert. Trotz der starken Ähnlichkeit (55% Sequenzidentität und 75% Sequenzähnlichkeit) haben LS2 und dU2AF50 sehr unterschiedliche RNS-Bindespezifität (Erkennung von Guanosin im einen und Polypyrmidinsequenzen im anderen Fall). Dies ermöglicht es Ihnen das Spleißen der Ziel RNSs zu befördern oder zu unterdrücken. Die Anreicherung von LS2 in Hoden, spricht für seine mögliche Rolle in der Hodenfunktion, Produktion von Gameten und der Zellregulation. Allerdings wurden die zugrundeliegenden molekularen Mechanismen noch nicht gut untersucht.

Diese Arbeit zeigt eine strukturelle und funktionelle Analyse von LS2-RNS-Bindungsdomänen zusammen mit dem interdomänen Linker und deren Wechselwirkungen mit verwandten Guanosinreichen RNS-Sequenzen. Ein integriert struktur-Biologischer Ansatz wird durch die Kombination von NMR-Spektroskopie mit Kleinwinkel-Röntgenstreuung (SAXS) -Experimente in Lösung angewendet. Die Studie wird durch biophysikalische und biochemische Assays ergänzt. Um die strukturellen Eigenschaften zu vergleichen, werden die dU2AF50-RNS-Bindungsdomänen und ihre Wechselwirkungen mit Poly-U-RNS untersucht, was die Unterschiede zwischen den LS2- und U2AF50-Paralogen in Drosophila offenbart.

Kapitel 1 enthält die Einführung in den biologischen Hintergrund des alternativen Spleißprozesses und die Rolle des U2AF-Splicing-Faktors und skizziert die Evolution von LS2 aus dU2AF50 sowie die Rolle von G-Quadruplex-RNSs bei der Spleißregulation. Die Grundlagen der NMR und ihre Anwendung zur Untersuchung der Struktur und Dynamik von Proteinen und RNSs werden ebenfalls vorgestellt. Kapitel 2 umfasst die Materialien und Methoden für die Molekularbiologie sowie für biochemische und strukturelle Studien.

In Kapitel 3 werden die wichtigsten Ergebnisse der Arbeit berichtet. Kapitel 3.1.1 beschreibt die strukturelle und biophysikalische Charakterisierung von LS2-RNS-Bindungsdomänen (RRM1 und RRM2) und dem Interdomain-Linker. Lösungs-NMR-Daten zeigen, dass beide RRM-Domänen von LS2 eine kanonische βαββαβ Topologie annehmen. Neue und ungewöhnliche Strukturmerkmale wurden durch eine Wechselwirkung des RRM1-RRM2-Linkers mit RRM2 und die Identifizierung einer neuen α-helikalen Region in den LS2-spezifischen Linkerresten beobachtet. In Kapitel 3.1.2 werden verschiedene Guanosinreiche RNSs untersucht, um ihre Fähigkeit zur Bildung von G-Quadruplex-Falten in Gegenwart monovalenter Kationen als potenzielle Liganden für LS2 zu untersuchen. Bisher beschreiben hochaffine Guanosin-reiche RNS-Liganden für LS2, die aus SELEX-Experimenten (GGX-Motiv) gewonnen wurden, wurden für biochemische und NMR-Untersuchungen getestet und optimiert. Umfassende NMR-Analyse einer 21mer Guanosin-reichen RNS ergab, dass sie eine einheitliche Konformation in einer dimeren, parallelen und dreifach planaren Topologie annimmt.

Kapitel 3.1.3 beschreibt die Interaktionsstudien von LS2-RRM-Domänen und verschiedenen guanosinreichen Liganden. Die Daten zeigen, dass LS2 eine G-Quadruplex-RNS-Struktur erkennt. Die Spezifität dieser Interaktion umfasst hauptsächlich RRM2, mit wahrscheinlich zusätzlichen Beiträgen aus der helikalen Region im RRM1-RRM2-Linker. Die beschriebene LS2-G-Quadruplex-Wechselwirkung ist neu, da RRM-Domänen am ehesten für die Bindung an einzelsträngige Nukleinsäure Sequenzen bekannt sind. Die Ergebnisse zeigen, dass LS2 RRM2 die Spezifität für die Erkennung der 21mer RNS G-Quadruplex-Konformation liefert. Die Informationen aus Protein-RNS-Interaktionsstudien werden verwendet, um LS2 RRM1,2-Mutanten mit beeinträchtigter RNS-Bindungsaktivität zu erzeugen, um die Beiträge der RNS-Interaktion für die funktionelle Aktivität in vivo zu analysieren.

Kapitel 3.2 beinhaltet über die Charakterisierung von dU2AF50-RNSA-Bindungsdomänen (RRM1,2) und deren Wechselwirkungen mit Poly-U-RNS. NMR-basierte Sekundärstruktur und Backbone-Dynamik-Daten zeigen, dass beide RRM-Domains eine kanonische Topologie annehmen. NMR-Titrationsdaten zeigen, dass beide RRM-Domänen an der RNS-Bindung über

konservierte RNP-Stellen beteiligt sind. Auf der anderen Seite zeigen ITC-Studien, dass Protein und RNS einen 1:1-Komplex mit eine Dissoziationskonstanten (Kd) ähnlich der U2AF65 RRM1,2-U9-Wechselwirkung bildet. Insgesamt zeigen die Daten, dass die strukturellen und RNS-Bindungseigenschaften von dU2AF50 mit dem menschlichen orthologen U2AF65 übereinstimmen.

Zusammengefasst zeigen die Ergebnisse dieser Arbeit die Strukturen der RRM-Domänen des alternativen Spleißfaktors LS2 und dessen Wechselwirkung mit einer G-Quadruplex-RNS-Struktur. Neue Strukturmerkmale, die mit den RRM-Domänen assoziiert sind, und Sequenzvariationen in Schlüsselresten, die RNS-Kontakte in LS2 und seinem Paralog U2AF50 vermitteln, erklären die drastisch unterschiedlichen RNS-Bindungspräferenzen für die beiden Proteine. Die divergierende Evolution der beiden Genprodukte, die aus der Gen-Duplikation in Drosophila-Spezies stammt, präsentiert ein faszinierendes Beispiel für Paralog-Proteine zur Anpassung von gewebespezifischen Funktionen.

# Table of contents


**Summary** ..... 1

**Table of contents** ..... 7

**Chapter 1 Introduction** ..... 11

**1.1 Biological background** ..... 11

**1.1.1 Alternative splicing** ..... 11

**1.1.2 Chemistry behind splicing reaction** ..... 13

**1.1.3 Spliceosome and its assembly** ..... 13

**1.1.4 Regulation of alternative splicing** ..... 15

**1.1.5 U2AF and 3' splice site recognition** ..... 17

**1.1.6 dU2AF50 and evolution of LS2** ..... 19

**1.1.7 G-quadruplex and its role in splicing regulation** ..... 21

**1.2 NMR spectroscopy** ..... 24

**1.2.1 Basic principles of NMR** ..... 24

**1.2.2 The chemical shift** ..... 26

**1.2.3 Relaxation** ..... 26

**1.2.4. Protein NMR** ..... 28

**1.2.5 NMR analysis of RNA** ..... 31

**1.2.6. Protein-RNA interaction by NMR** ..... 32

**1.3 Scope of the thesis** ..... 35

**Chapter 2 Materials and methods** ..... 37

**2.1 Chemicals and consumables** ..... 37

**2.2 Molecular biology** ..... 37

**2.2.1 Bacterial strains** ..... 37

**2.2.2 Plasmids for recombinant protein expression** ..... 37

**2.2.3 Cloning and site-directed mutagenesis** ..... 38

**2.2.4 Transformation and plasmid DNA isolation** ..... 39

**2.3 Protein expression** ..... 40

**2.4 Protein purification** ..... 40

**2.4.1 Protein purification protocol for soluble protein constructs** ..... 41

**2.4.2 Inclusion body purification protocol** ..... 42

**2.4.3. Protein analysis** ..... 42

**2.5 RNA oligonucleotides** ..... 43

**2.6 Biophysical methods** ..... 43
**2.6.1 Isothermal titration calorimetry** ..... 43
**2.6.2 Circular Dichroism** ..... 43
**2.6.3 Static light scattering (SLS)** ..... 43
**2.7 NMR** ..... 44
**2.7.1 NMR experiments** ..... 44
**2.7.2 Structure calculation** ..... 45
**2.7.3 NMR titration analysis** ..... 46
**2.8 SAXS** ..... 46
**2.9 Crystallization trials** ..... 46
**Chapter 3 Results** ..... 47
**3.1 Characterization of poly-G RNA recognition by LS2** ..... 47
**3.1.1 Analysis of LS2 RRM domains and linker** ..... 47
**3.1.1.1 Insights from the sequence analysis** ..... 47
**3.1.1.2 Construct optimization of LS2 RNA-binding domains** ..... 50
**3.1.1.3 Aggregation-prone behavior of LS2 RRM1,2** ..... 52
**3.1.1.4 NMR analysis and structures of LS2 individual RRM domains** ..... 56
**3.1.1.5 Interaction between LS2 Linker and RRM2** ..... 59
**3.1.1.6 Characterization of the LS2 RRM1,2 linker** ..... 61
**3.1.1.7 Solution structure of linker-RRM2** ..... 65
**3.1.1.8 Biophysical analysis of LS2 RRM1,2** ..... 67
**3.1.1.9 Crystallization trials** ..... 68
**3.1.2 G-quadruplex formation by LS2 target RNA** ..... 71
**3.1.2.1 21mer poly-G RNA forms G-quadruplex** ..... 71
**3.1.2.2 NMR shows that G-quadruplex has multiple conformations** ..... 71
**3.1.2.3 Designing Shorter oligonucleotides on the basis of SELEX** ..... 73
**3.1.2.4 Low KCl concentration induces uniform conformation for 21mer and 8mer** ..... 74
**3.1.2.5 Biophysical characterization of 21mer RNA G-quadruplex** ..... 77
**3.1.3 Characterization of LS2 RRM domains and poly G RNA interaction** ..... 83
**3.1.3.1 Interaction of LS2 RRM1,2 with 21mer poly G RNA by NMR** ..... 83
**3.1.3.2 Interaction of individual RRM domains with 21mer by NMR** ..... 85
**3.1.3.3 Interaction of single RRM domains with shorter poly G oligonucleotides** ..... 89
**3.1.3.4 RRM2 specifically interacts with 21mer G-quadruplex structure** ..... 93
**3.1.3.5 SLS analysis of RRM2-21mer complex** ..... 95
**3.1.3.6 Linker-RRM2 interaction with 21mer** ..... 95
**3.1.3.7 LS2 interacts with the uniform conformation adopted by 21mer** ..... 96

**3.1.3.8 G-quadruplex-specific inhibitor disrupts linker-RRM2-21mer complex** ..... 98
**3.1.3.9 Mutational analysis to abolish RNA binding contribution** ..... 99
**3.2 Analysis of dU2AF50 RRM domains and their interaction with poly-U RNA** ..... 103
**3.2.1 NMR analysis dU2AF50 RRM1,2** ..... 103
**3.2.2 NMR titration of dU2AF50 RRM1,2 with U9** ..... 105
**3.2.3 ITC study of dU2AF50 RRM1,2 and U9 interaction** ..... 106
**Chapter 4 Discussion** ..... 109
**4.1 Conserved topology of LS2 and dU2A50 RRM domains** ..... 109
**4.2 Features of LS2 RRM1,2 interdomain linker** ..... 110
**4.3 Significance of the G-quadruplex formation by LS2 target RNA** ..... 112
**4.4 Interaction between LS2 RRM domains and G-quadruplex RNA** ..... 113
**4.5 RNA binding proteins and aggregation** ..... 116
**Conclusion** ..... 117
**References** ..... 118
**Appendix** ..... i
**oligonucleotide sequences** ..... i
**Abbreviations** ..... ii
**Acknowledgement** ..... iv
**Curriculum vitae** ..... v

# Chapter 1 Introduction

## 1.1 Biological background

### 1.1.1 Alternative splicing

Inside the cells of all living organisms, tiny molecular machines are constantly decoding the information encrypted in DNA to synthesize functional proteins. In this process, RNA molecules act as an intermediate by passing on the information from DNA to the ribosomes in order to direct the protein assembly. Most of the prokaryotes follow 'one gene, one polypeptide' rule, as coding sequences of their genes are uninterrupted, with minor exceptions. On the other hand, eukaryotic genes are divided into coding and non-coding regions. The eukaryotic coding regions are called as **exons** and are often interrupted by few to several non-coding sequences called as **introns**. During the process of transcription, these introns are also passed on to the pre-mRNA transcript. In order to have a functional protein, it becomes necessary to remove these unwanted introns from pre-mRNA before it is passed on to the ribosomes. This process of intron removal and thereby joining the concomitant exons to generate mature mRNA is called as '**splicing**' (Figure 1.1, A).

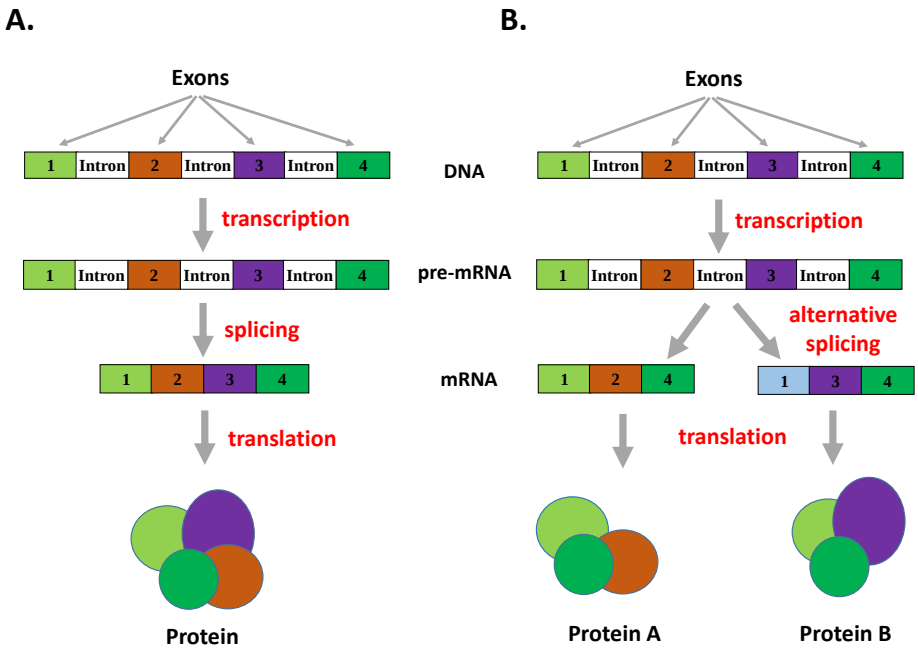



**Figure 1.1 Schematic representation of splicing (A) and alternative splicing (B).** Most eukaryotic genes are divided into exons and non-coding introns. Splicing causes intron removal and concomitant exon joining whereas alternative splicing uses a different combination of exons to produce a variety of mRNA yielding various protein isoforms.

Although the significance of eukaryotic gene interruption is not understood yet, it certainly forms the basis of an interesting phenomenon of '**alternative splicing**' (Figure 1.1, B), in which different combinations of exons are tried, resulting in a variability of splicing patterns. Consequently, such alternatively spliced mRNAs are used by ribosomes to translate proteins with different amino acid composition and often with different biological functions (Raj and Blencowe 2015). Thus, alternative splicing exploits single gene to code for multiple protein isoforms, thereby increasing proteomic diversity (Black 2003). It is estimated that ~100,000 alternative splicing events occur in major human tissues with ~95% of multi-exonic genes being alternatively spliced (Pan, Shai et al. 2008).

As alternative splicing involves the choice of different combinations of exons, it can exhibit a variety of splicing patterns (Figure 1.2). Most commonly represented alternative splicing example is of cassette exons, in which depending on whether a discrete exon is excluded (skipped exon) or included (cryptic exon), alternative splicing can create two different isoforms either without or with cassette exon, respectively (Matlin, Clark et al. 2005). In the mutually exclusive alternative splicing event, the unique exon is selected from multiple available exons. An another common splicing pattern is the intron retention, exhibited in approximately 75% of mammalian genes during developmental stage (Scotti and Swanson 2016). On the other hand, use of alternative 5' and 3' splice sites allow additional or reduction of several nucleotides at 3' or 5' end of the exons respectively, thereby exhibiting exon modification.

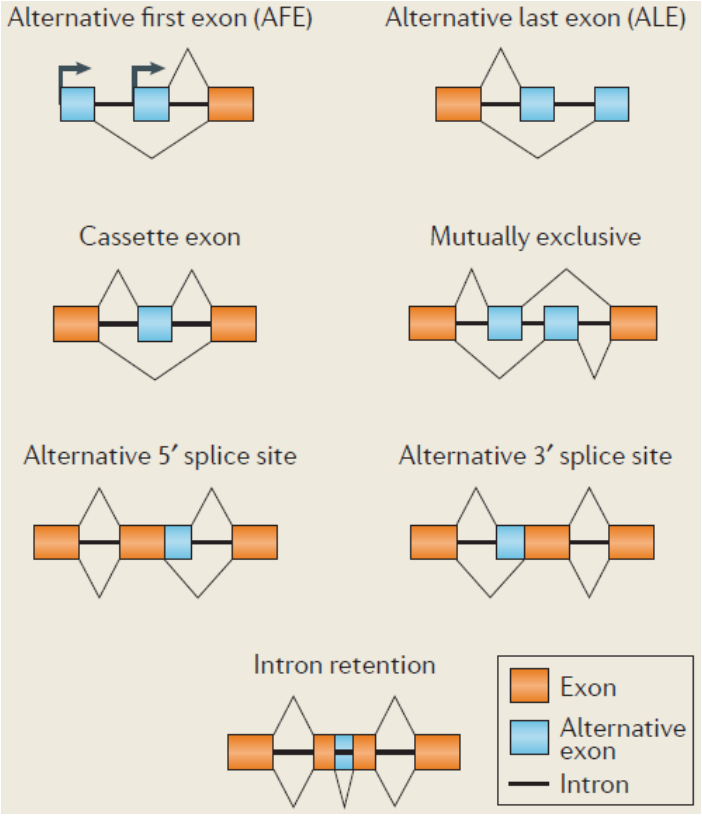



**Figure 1.2 Patterns of alternative splicing.** Depending upon cell type and developmental stage, multi-exon genes can undergo several splicing variations as shown above [Figure adapted from (Scotti and Swanson 2016)].

## 1.1.2 Chemistry behind splicing reaction

Pre-mRNA splicing comprises of two transesterification steps, each involving nucleophilic attack on the terminal phosphodiester bonds of the intron (Figure 1.1). The reaction is initiated by nucleophilic attack by 2'- hydroxyl group of conserved Adenosine of an intron (Branch point site, BPS) on the phosphate group at 5' exon- intron boundary. As a result, 5' exon is cleaved from the intron, leaving detached 5' exon and intron/3'-exon in the lariat form. In the second step, the nucleophilic attack is performed by 3'- hydroxyl of the detached exon on the phosphate of 3'- end of the intron. This results in ligation of two exons with concomitant release of an intron in lariat form.

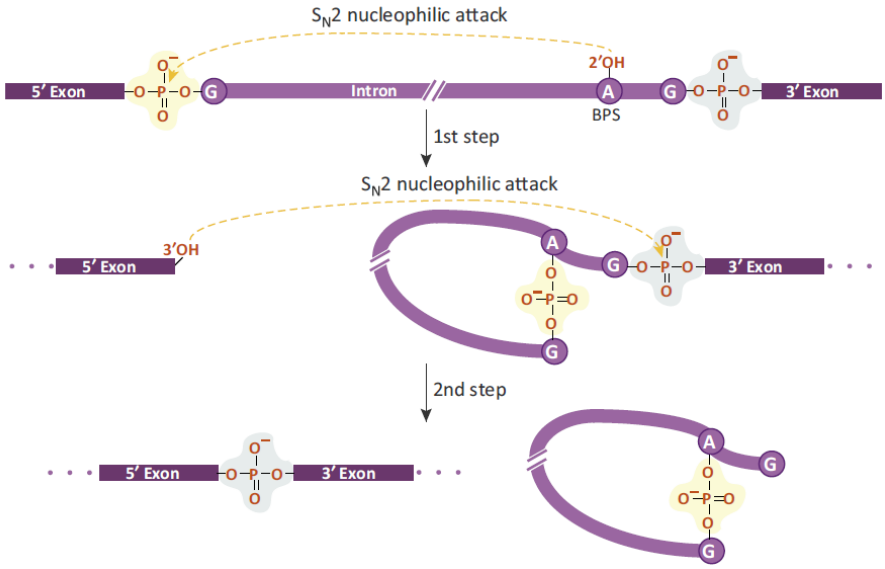



**Figure 1.3 Two transesterification steps of splicing.** In the first step, nucleophilic attack by 2'- hydroxyl group from branch point adenosine results in two reaction intermediates: lariat intron attached to 3' exon intermediate and free 5' exon. The second step results in ligated two exons and released intron lariat. [Figure adapted from (Papasaikas and Valcarcel 2016)].

## 1.1.3 Spliceosome and its assembly

Splicing is performed by the large and complex molecular machinery called as '**spliceosome**', which is found inside the nucleus of the eukaryotic cells. There are two types of spliceosomes present in the eukaryotic cells, the major (U2-dependent) and minor (U12- dependent) spliceosome, respectively (Will, Schneider et al. 1999). Each spliceosome is made up of five small nuclear ribonucleoprotein particles (snRNPs). Major spliceosome comprises of U1, U2, U4, U5 and U6 snRNAs and is responsible for splicing of ~95.5% of all introns (Turunen, Niemela et al. 2013). On the other hand, minor spliceosome is made up of U11, U12, U4atac, U5 and U6atac snRNAs (Scotti and Swanson 2016). During splicing a large number of auxiliary proteins are also associated with spliceosome.

In order to perform splicing reaction correctly, spliceosome has to recognize ends of the intron (5' and 3' splice sites) accurately (Figure 1.4). 5' splice site junction is marked by 9 nucleotide degenerate consensus sequence YAG/GURAGU (where Y is pyrimidine, R is A or G, and the / stands for the actual splice site) (Busch and Hertel 2012). On the other hand, 3' splice site is defined by 3 sequence elements, namely branch point sequence (BPS), the polypyrimidine tract (Py tract), and the 3' intron/exon junction (Busch and Hertel 2012).

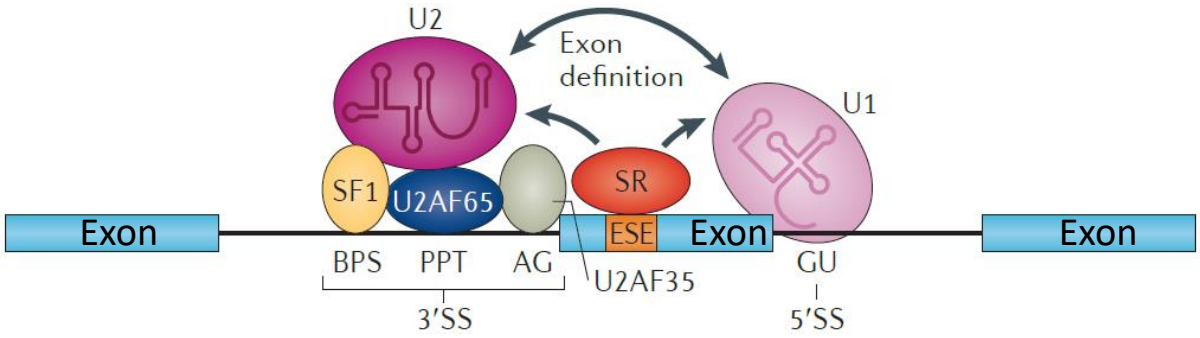



**Figure 1.4 Molecular interactions at 5' and 3' splice sites.** The recognition of 5' splice site is performed by U1 snRNP by interacting with the conserved consensus sequence. On the other hand, U2AF heterodimer and SF1 specifically interacts with Py tract and BPS, respectively and subsequently, recruit U2 snRNP at 3' splice site. [Figure adapted from (Fu and Ares 2014)].

Spliceosome formation proceeds through four distinct complexes, which can be categorized as E complex, A complex, B complex and C complex (Figure 1.5). Initiation of spliceosome assembly starts with the base-pairing of U1 snRNA to the 5' splice site consensus sequence along with binding of the splicing factor 1 (SF1) to the conserved branch point in an ATP-independent manner, which leads to the formation of E' complex. This E' complex can be converted to **E complex** by subsequent binding of U2 auxiliary factor (U2AF) heterodimer (made up of the large subunit and small subunit) to the 3' splice site consensus Py tract and AG dinucleotide (Kent, Ritchie et al. 2005). Replacement of SF1 by U2 snRNP converts ATP-independent E complex into ATP-dependent pre- spliceosome **A complex**. This allows concomitant recruitment of the remaining U4/U6- U5 tri-snRNPs to form B complex, in a reaction catalyzed by the DExD/H helicase Prp28. Series of remodeling and conformational changes, including loss of U1 and U4 snRNPs, results in the conversion of B complex to catalytically **active B complex** (B* complex), resulting in the formation of U2/U6 snRNA structure, which is responsible for catalysis of splicing reaction(Matera and Wang 2014). This activated B* complex then carries out the first transesterification step, generating **C complex**, which contains the free 5'-exon and intron-3' exon lariat intermediate. It then undergoes additional ATP-dependent rearrangements to catalyze the second transesterification step, and produces lariat intron and spliced exons. In the end, U2, U5, and U6 snRNPs are released from the complex, by the action of several ATP-dependent RNA helicases and recycled for additional rounds of splicing (Matera and Wang 2014).

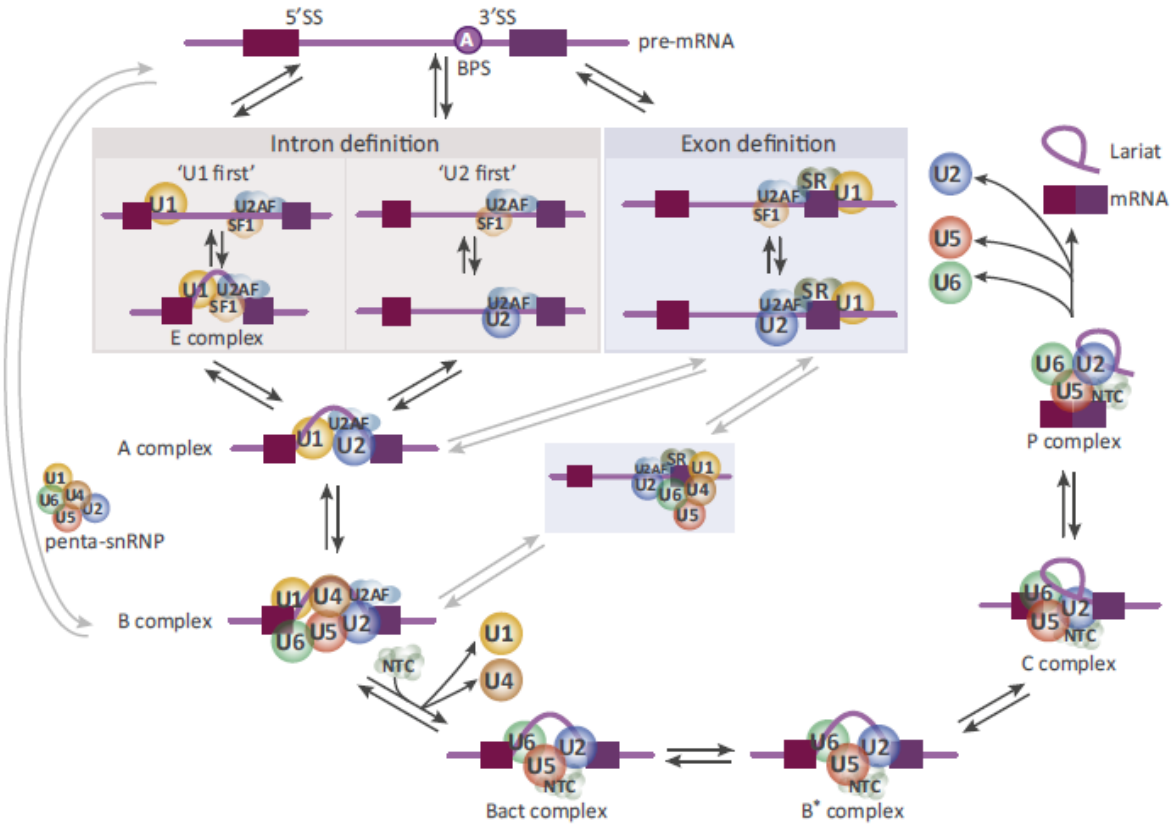



**Figure 1.5 Steps of Spliceosome assembly.** Spliceosome assembly is a stepwise process involving snRNPs binding to the pre-mRNA. Spliceosome assembly starts with the recruitment of U1 and U2AF at 5' and 3' splice sites respectively. It further leads to the formation of various complexes marked by addition, rearrangement and release of different snRNPs, which ultimately performs catalytic steps of splicing. [Figure adapted from(Papasaikas and Valcarcel 2016)].

## 1.1.4 Regulation of alternative splicing

Alternative splicing determines the timing and the location of particular protein isoform production and thereby, regulates many biological activities. For example, alternative splicing has been reported to play a critical role in sex determination of insects (Salz 2011). Hence, change in alternative splicing pattern can have an impact on many cellular activities. Indeed, errors in splicing are attributed to various human genetic disorders as well as cancers (Scotti and Swanson 2016). Hence, a high degree of specificity and fidelity of splicing reaction is necessary for appropriate expression of functional mRNAs and thereby translation of respective protein (Busch and Hertel 2012).

The frequency by which any particular exon is selected or rejected depends on upon the relative ''strength'' of the splice site. This strength corresponds to sequence complementarity of 5' splice site to U1 snRNA as well as longer uninterrupted polypyrimidine tracts at 3' splice

site. In contrast, there are many potential splice sites in the human genome which show great similarity to true splice sites and form pseudo exons (Sun and Chasin 2000). In fact, over-representation of such pseudo exons is often observed than the true splice sites. It turns out that the splice site consensus is generally not sufficient to decide whether a particular site will assemble spliceosome and undergo splicing. Hence, for controlling splicing, *cis*-acting RNA sequences, known as **Splicing Regulatory Elements** (SREs) have emerged.

Depending on the position and the function, SREs are classified into four categories: **exonic splicing enhancers** (ESEs), **exonic splicing silencers** (ESSs), **intronic splicing enhancers** (ISEs) and **intronic splicing silencers** (ISSs) (Chen and Manley 2009). These SREs generally function by recruiting *trans*-acting RNA binding proteins, which can influence the efficiency of spliceosome binding resulting into either enhancement or repression of splice site use. For example, ESEs are mostly recognized by SR protein family, which consists of RS (arginine/serine) domain and at least one RRM domain. In contrast, splicing silencers are reported to recruit hnRNPs (Busch and Hertel 2012). In addition to them, other tissue-specific splicing factors, such as FOX, CELF; NOVA (neuro-oncological ventral antigen and MBNL (muscleblind-like) proteins, functioning equally as splicing repressors and enhancers are also reported (Heyd and Lynch 2011). Mechanisms by which these proteins promote or repress splicing remains largely unclear. It is proposed that splicing activators generally interact with the spliceosome components, facilitating their recruitment to the neighboring splice site whereas splicing repressor sterically hinder the binding of spliceosome components to the nearby splice site and thereby reducing neighboring exon usage.

It is reported that a typical pre-mRNA contains multiple ISE/ESEs as well as ISS/ESSs that involve interactions with multiple regulatory proteins, whose combinatorial actions determine the fate of splicing pattern. Hence, the study of these SREs and their cognate factors on a global scale is necessary to derive a 'splicing code' depicting a set of rules for splicing regulation (Wang and Burge 2008, Barash, Calarco et al. 2010).

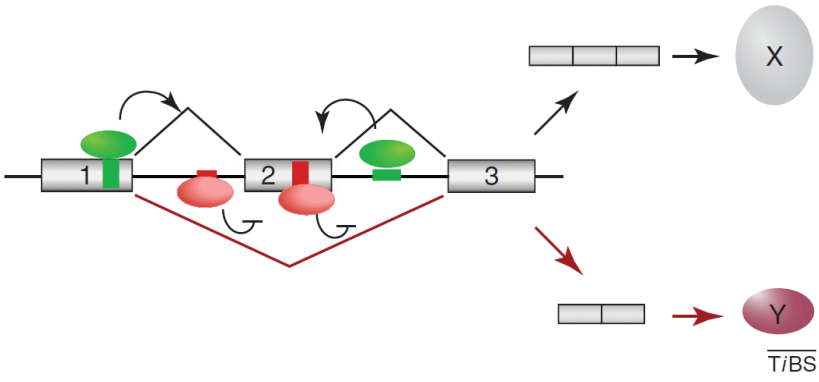



**Figure 1.6 Splicing regulation by Splicing Regulatory Elements (SREs).** *Cis*-acting Splice enhancers (ESE/ISE) together with the *trans*-acting RNA binding proteins (green) are thought to interact with spliceosome components to recruit them to the nearby splice site. In contrast splicing repressors (ESS/ISS) bind to RNA binding proteins (red) which seems to prevent the interaction of spliceosome components with the neighboring splice sites. The combinatorial action of these splicing enhancers and silencers determine the inclusion or exclusion of target exon into a mature transcript. [Figure adapted from (Heyd and Lynch 2011)].

### 1.1.5 U2AF and 3’ splice site recognition

As mentioned in section 1.1.3, U2AF acts in the early stage of spliceosome assembly, during the formation of E complex. U2AF is ubiquitously expressed in all eukaryotic cells and plays a central role during splicing by directly defining ~ 88% of the functional 3’ splice sites in the human genome (Shao, Yang et al. 2014).

U2AF is a heterodimer made up of the Large subunit ($U2AF^{LS}$) and small subunit ($U2AF^{SS}$). $U2AF^{LS}$ binds to Py tract sequence flanked by BPS and 3’ splice site (Zamore, Patton et al. 1992), whereas $U2AF^{SS}$ binds to $U2AF^{LS}$ as well as contacts AG dinucleotide next to intron-exon boundary (Zhang, Zamore et al. 1992, Merendino, Guth et al. 1999). Furthermore, $U2AF^{LS}$ is also known to interact with splicing factor 1 (SF1), which concomitantly recognizes BPS of the pre- mRNA, which is the site for the first step of splicing reaction (Abovich and Rosbash 1997). Following these contacts, the 3’ end of the intron becomes competent for the interaction with U2 snRNP. Thus, U2AF plays a central role in defining 3’ splice site and thereby initiating splicing. $U2AF^{LS}$ on its own can recruit U2snRNP to the branch site, provided Py tract is long enough. Otherwise, additional binding interactions between $U2AF^{SS}$ and adjacent AG are required for stable U2AF association and thereby, U2snRNP recruitment (Moore 2000). Extensive studies on human U2AF heterodimer (made up of U2AF65 and U2AF35) provide detailed atomic-level information about most of the molecular interactions involved defining 3’ splice site (Figure 1.8).

U2AF65 consists of three RNA Recognition Motifs (RRM domains) as well as two peptide motifs, namely RS and ULM (U2AF Ligand Motif). RRM domain is the most abundant RNA binding domain in higher vertebrates by representing about 0.5-1% of human genes (Venter, Adams et al. 2001). Typically, RRM domain is around 90 amino acids long and has βαββαβ topology. It has four-stranded antiparallel β-sheet with two α-helices. Each RRM is characterized by a conserved stretch of aromatic and positively charged residues, called as RNP2 and RNP1, which are located on β1 and β3 respectively and are employed for nucleic acid binding (Daubner, Clery et al. 2013) (Figure 1.7).

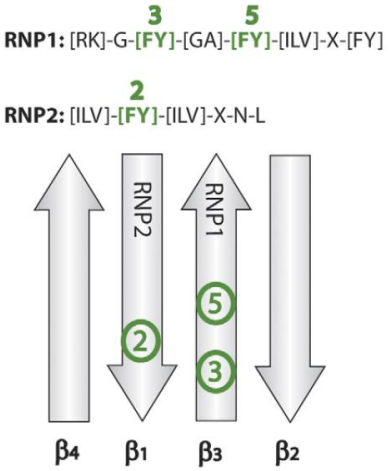



**Figure 1.7 Schematic representation of RNP1 and RNP2 consensus sequences from RRM domain.** [Figure adapted from (Clery, Blatter et al. 2008)].

Biochemical studies showed that out of three U2AF65 RRM domains only two central RRM domains (RRM1 and RRM2) are sufficient for binding to Py tract (Zamore, Patton et al. 1992, Banerjee, Rahn et al. 2003, Banerjee, Rahn et al. 2004). Crystal structure of these two RRM domains lacking the interdomain linker in complex with polyuridine RNA ($U_{12}$) provided atomic details on protein-RNA contacts and showed that uridine recognition is governed by unique hydrogen bonding pattern rather than shape selective recognition. The structure shows that use of flexible side chains for achieving a majority of protein-RNA contacts as well as water-mediated interactions allow, U2AF65 to sustain base substitutions within Py tract sequences found naturally. (Sickmier, Frato et al. 2006).

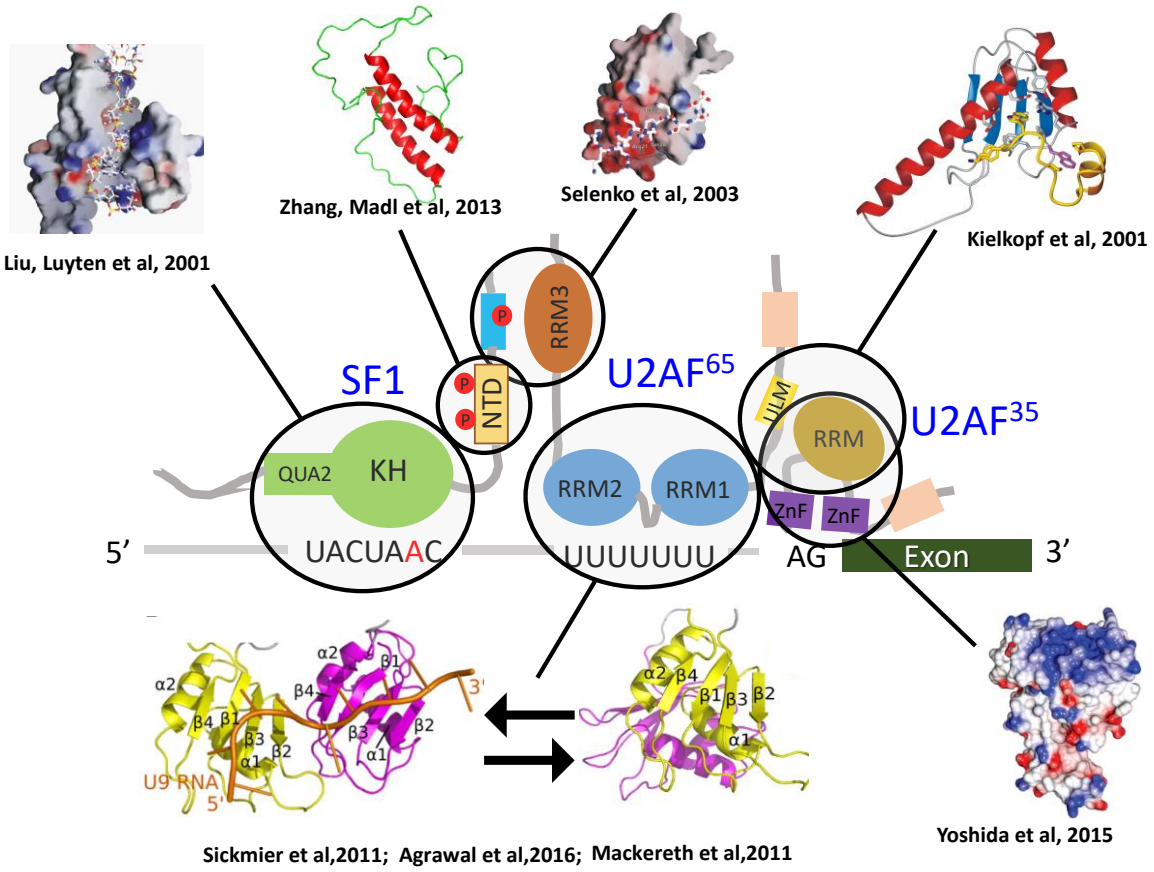



**Figure 1.8 Schematic overview of molecular interactions defining 3' splice site.** U2AF$^{LS}$ recognizes Py tract by using central RRM1 and RRM2 domains (Sickmier, Frato et al. 2006, Mackereth, Madl et al. 2011, Agrawal, Salsi et al. 2016). With third RRM U2AF65 is reported to contact SF1 (Selenko, Gregorovic et al. 2003, Zhang, Madl et al. 2013). SF1 further interacts with branch point adenosine via its KH-QUA2 domain (Liu, Luyten et al. 2001). U2AF$^{SS}$ interacts with U2AF$^{LS}$ via its pseudo-RRM domain (Kielkopf, Rodionova et al. 2001), whereas U2AF$^{SS}$ Zn fingers are reported to recognize AG dinucleotide (Yoshida, Park et al. 2015).

NMR data agreed well with crystallographic studies in the case of protein-RNA contacts. But, it also revealed the conformational dynamics adopted by two RRM domains in presence or absence of strong Py tract. The study reported that tandem domains of U2AF65 can populate two distinct domain arrangements i.e. closed or open conformation depending upon

availability of high-affinity RNA ligand. The molecular rheostat like the model was proposed. According to this model, the equilibrium between the two conformations quantitatively correlates with the strength of the Py tract to the efficiency to recruit U2 snRNP to the intron during spliceosome assembly (Mackereth, Madl et al. 2011). Whereas, recently published crystal structure of extended RRM1,2 with poly-U RNA, shows that the linker residues between two RRM domains as well as N- and C- terminal extensions also play role in RNA recognition (Agrawal, Salsi et al. 2016).

The third RRM of U2AF65 is a pseudo RRM or also called as UHM (U2AF Homology Motif), containing 3 times longer α1-helix in comparison to typical RRM domain. This UHM domain is required for the interaction with N-terminal of SF1 protein (Rain, Rafi et al. 1998, Selenko, Gregorovic et al. 2003). U2AF35 also has a UHM domain, through which it interacts with U2AF65 (Kielkopf, Rodionova et al. 2001), whereas flanking Zn-fingers are proposed to mediate interactions with AG dinucleotide as shown by mutational analysis in combination with crystal structure of apo form of U2AF small subunit from *S. pombe* (Yoshida, Park et al. 2015). In retrospect, RS domains of U2AF heterodimer are known to be required for high-affinity binding to the RNA (Rudner, Breger et al. 1998).

### 1.1.6 dU2AF50 and evolution of LS2

U2AF heterodimer is highly conserved among all eukaryotes, from *S. pombe* to humans (Taliaferro, Alvarez et al. 2011). *Drosophila* homologs of $U2AF^{LS}$ and $U2AF^{SS}$ are called as dU2AF50 and dU2AF38 respectively (Kanaar, Roche et al. 1993). dU2AF50 shares high sequence identity with U2AF65 (human $U2AF^{LS}$) as well as has conserved domain arrangement (Figure 3.1).

Interestingly, testes of the *Drosophila* contain an additional protein called as Large subunit 2 (LS2). LS2 is a paralog of $U2AF^{LS}$, which is found to be evolved from dU2AF50 (Taliaferro, Alvarez et al. 2011). LS2 gene duplication is supposed to have happened between 60 to 250 million years ago. LS2 is found to be specific for *Drosophila* as no LS2 ortholog is detected in mosquitoes or in honeybees. LS2 is reported to be 55% identical and 70% similar to dU2AF50 at the sequence level (Figure 3.1). Lack of intron in LS2 gene and presence of five introns in dU2AF50 gene implies that retro-duplication event by using an RNA intermediate lead to the evolution of LS2.

Surprisingly, LS2 is not a redundant copy of dU2AF50 but rather behaves as a splicing factor with different RNA-binding specificity as well as function (Figure 1.9). LS2 is reported to show a marked preference for guanosine-rich RNA in contrast to pyrimidine-rich RNA recognition by dU2AF50. By doing so, it is reported to function as splicing repressor both in vivo and in vitro. As like dU2AF50, LS2 is also found to interact with the dU2AF38 in an RNA-independent manner, most likely through hydrophobic ULM domain. Electrophoretic gel mobility shift assays (EMSA) show that the LS2/dU2AF38 heterodimer binds G-rich RNA more tightly ($K_d$=150 nM) in comparison to LS2 alone ($K_d$ = 1.9 μM).

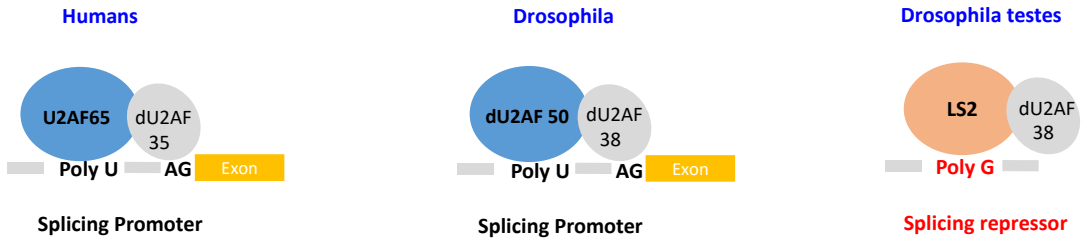



**Figure 1.9 Schematic representation of U2AF65, dU2AF50, and LS2 with their binding partners.** Ubiquitously expressed U2AF$^{LS}$ (U2AF65 in humans, dU2AF50 in *drosophila* recognize poly-U tract and promote splicing. LS2 is a paralog of U2AF$^{LS}$ which is evolved to have different RNA binding specificity (poly-G instead of poly-U RNA) and functions as a splicing repressor in the testes of the *drosophila* (Taliaferro, Alvarez et al. 2011).

In contrast to the ubiquitous expression of the dU2AF50, expression of LS2 is also found to be highly specific for testes of the *Drosophila*. LS2 affected transcripts are also found to have testes- enriched expression, showing involvement in testes function, gamete production, and cellular regulation through phosphorylation (Taliaferro, Alvarez et al. 2011).

LS2 is proposed to perform its function by binding to the *cis*-regulatory poly-G tract, which is located 60 nucleotides upstream of the polypyrimidine site, inhibiting the interaction of U2AF heterodimer with neighboring pyrimidine tract most likely by steric hindrance (Figure 1.10). Another binding site for LS2 is also detected at 120 nucleotides downstream of target exon. It is reported that, when LS2 binds to this poly-G site, it leads to inclusion of the target exon as polypyrimidine tract is freely accessible for U2AF heterodimer because of the absence of LS2 to the neighboring poly-G tract.

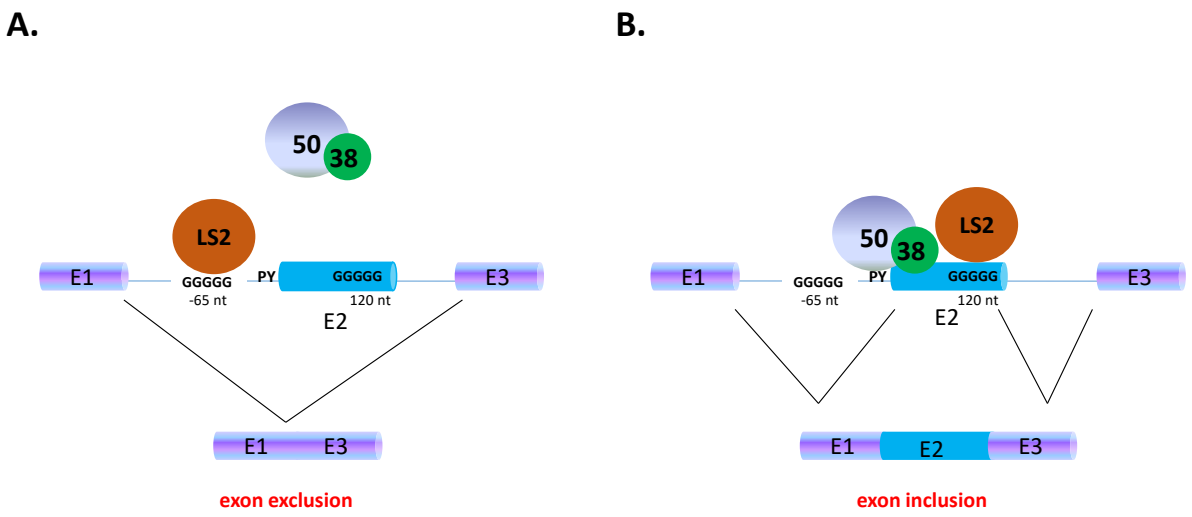



**Figure 1.10 Schematic representation of the mode of action by LS2.** LS2 acts as a splicing repressor when bound to poly-G tract upstream of target exon, by preventing interaction of U2AF heterodimer with pyrimidine tract. On the other hand, target exon is included upon interaction of LS2 with the downstream poly-G tract, as pyrimidine tract is accessible for U2AF heterodimer.

## 1.1.7 G-quadruplex and its role in splicing regulation

Guanosine-rich nucleic acids have the ability to fold into stable, four stranded noncanonical structures called as **G-quadruplexes** (Davis 2004). A G-quadruplex structure contains stacked arrangement of G-quartet, each of which is made of four guanines interconnected by cyclic Hoogsteen hydrogen bonding in a planar arrangement, stabilized by cations (Adrian, Heddi et al. 2012) (Figure 1.11).

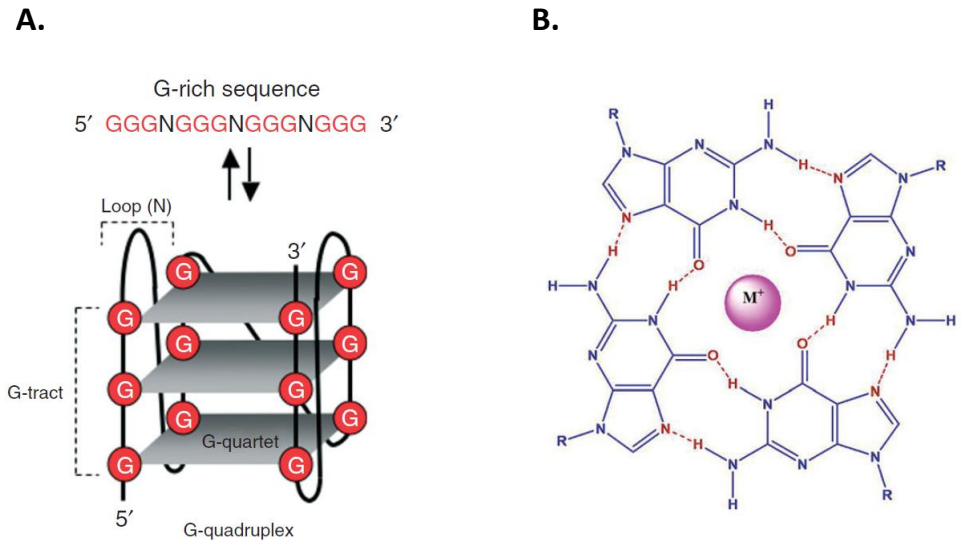



**Figure 1.11 Arrangement of G-quadruplex.** G- quadruplex **(A)** is a stacked arrangement of G-quartets **(B)**, which is a planar structure made up of four guanines and stabilized by a monovalent cation. [adapted from **A**. (Millevoi, Moine et al. 2012) and **B**. (Agarwala, Pandey et al. 2015)].

G-quadruplex topology can be intramolecular, made up of single-stranded DNA/RNA, or intermolecular, formed by two or four separate strands. Apart from this, G-quadruplexes are also known to show structural polymorphism depending upon the nature and concentration of cation, relative direction of the strands (parallel or antiparallel), the glycosidic conformation (syn or anti), the nature and the sequence of the connecting loops, the number of stacked G-quartets as well as the inclusion of bases other than guanine in quartet (Millevoi, Moine et al. 2012). RNA G-quadruplexes can have only a parallel topology because of steric constraints caused by the C' hydroxyl groups in RNA ribose sugars whereas DNA can form either type of structures (Simone, Fratta et al. 2015). Alkali metal ions coordinate G-quadruplex structures by stabilizing the negatively charged oxygen atoms in the individual G-quartet, with decreasing stability of $K^+ > Na^+ > Li^+$ (Simone, Fratta et al. 2015).

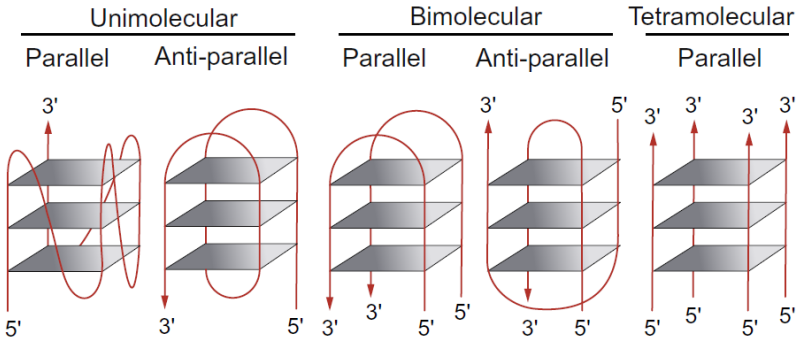



**Figure 1.12 G-quadruplex topology.** Topologies adopted by G-quadruplex structures with respect to a number of strands and their orientations. [Adapted from (Simone, Fratta et al. 2015)].

Bioinformatics studies predicted that the human genome contains over 376,000 G-quadruplex forming sequences (Huppert and Balasubramanian 2005, Todd, Johnston et al. 2005). They are found to be prevalent in telomeres, oncogenic promoters, mutational hotspots, and in a number of non-coding DNAs. On the other hand, G-quadruplex forming RNA sequences are located in introns, 5'and 3' ends of primary transcripts and telomeric RNA (Adrian, Heddi et al. 2012) (Millevoi, Moine et al. 2012). G-quadruplexes were also shown to exist in vivo using in cell-NMR (Hansel, Foldynova-Trantirkova et al. 2013), G-quadruplex-specific antibodies (Biffi, Tannahill et al. 2013, Biffi, Di Antonio et al. 2014) as well as through the discovery of G-quadruplex-specific helicases (Simone, Fratta et al. 2015).

Because of single stranded nature of RNA, RNA G-rich sequences are thought to be more susceptible to fold into a quadruplex structure than DNA, as there is no competition from base-pairing with complementary strand sequences. Recent studies show that RNA G-quadruplexes have a functional role in telomere maintenance, splicing, polyadenylation, RNA turnover, mRNA targeting and translation (Millevoi, Moine et al. 2012).

RNA G-quadruplexes are reported to play an important role in the regulation of alternative splicing. They are proposed to act as a *cis*-regulatory element, as they are discovered in the vicinity of splice sites of a growing number of genes (Millevoi, Moine et al. 2012). For instance, G-quadruplex behaving as intronic splicing silencer is reported in the intron 6 of the human telomerase (hTERT) gene by controlling its own splicing efficiency (Gomez, Lemarteleur et al. 2004).

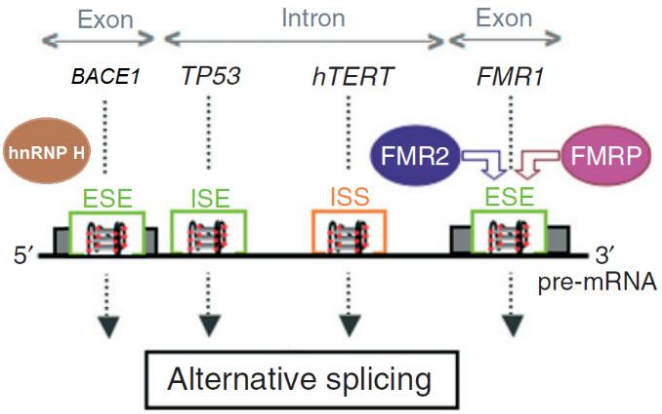



**Figure 1.13 Alternative splicing regulation by G-quadruplex.** RNA G-quadruplex structures formed in introns of TP53 and hTERT genes and reported to behave as intronic splicing enhancers and silencer respectively. On the other hand, G-quadruplex structures are also known to interact with splicing factors (hnRNP H in the case of BACE1 and FMRP/FMR2 in the case of FMR1) and thus, function as exonic splicing enhancer. Figure adapted from (Millevoi, Moine et al. 2012).

The G-quadruplex formation was also reported in the tumor suppressor gene TP53 intron 3. Site-directed mutagenesis of this G-quadruplex forming sequence decreased the excision of the neighboring intron 2 by 30%, suggesting its role as intronic splicing enhancer. Also, change in the topology of this intron 3 G-quadruplex was shown to be associated with increased risk of several common cancers (Marcel, Tran et al. 2011, Millevoi, Moine et al. 2012). The G-quadruplex formation was also reported in the exonic region altering splicing activity. For example, G-rich region in the exon 3 of the β-site amyloid precursor protein (APP) cleaving enzyme 1 (BACE1) gene was reported to form G-quadruplex and concurrently recruit hnRNP H to increase generation of the 501 isoform (Fisette, Montagna et al. 2012). Another example G-quadruplex acting as exonic splicing enhancer is alternative splicing of FMR1 gene, which is shown to be regulated via two independent G-quadruplex structures found in the exon 15 (Didiot, Tian et al. 2008).

In contrast, a recent study claims that RNA G-quadruplex structures are globally unfolded in eukaryotic cells and are exhausted in *E. coli* (Guo and Bartel 2016). The study identifies numerous endogenous regions that form a G-quadruplex fold in vitro. But, with the help of ability of dimethyl sulfate to methylate N7 position of guanosine in single-stranded form and subsequent stalling of reverse transcriptase, the study identifies that G-quadruplex structures are in the unfolded state in the mammalian cells.

## 1.2 NMR spectroscopy

The nucleus of an atom has an intrinsic property, called as Nuclear spin (I), which allows nucleus to behave like a tiny bar magnet. Nuclear spin is a discrete quantity and has a value multiple of 0.5, depending upon a number of protons and neutrons in the nucleus. Thus, if there is an even number of protons and neutrons, then the nucleus has no spin (I = 0). If the sum of a number of protons and neutrons in an atom is odd, then the nucleus has half-integer spin (I = 1/2, 3/2, 5/2). On the other hand, if there are an odd number of neutrons and protons, then the nucleus has an integer spin (I = 1,2,3).

Nuclei with spin value = 1/2 are of specific interest for Nuclear Magnetic Resonance (NMR), as they have the capacity to absorb and re-emit the electromagnetic radiation when immersed into a magnetic field. Examples of such spin half/NMR active nuclei are $^{1}H$,$^{15}N$,$^{13}C$, $^{31}P$.

### 1.2.1 Basic principles of NMR

In the absence of external magnetic field, nuclear spins are randomized and their motion along a Z-axis termed as Nuclear magnetic moment can be described by using following formula,

$$\mu_Z = \gamma I_Z = \gamma \hbar m$$

where, $\mu_z$ = magnetic moment along Z-axis, $\gamma$ = gyromagnetic ratio (ratio of the nucleus's magnetic dipole moment to its angular momentum), $\hbar$ = Plank constant, and $m$ = magnetic quantum number.

As defined by quantum mechanics, the nuclear magnetic moment of a nucleus can adopt on 2I+1 ways to align with an externally applied magnetic field of strength $B_0$. Hence, nuclei with I = 1/2, can have two possible nuclear spin orientations, either with or against the applied field $B_0$. This gives rise to two discrete nuclear spin states separated by certain energy difference and only one transition is possible between two energy levels. In the first state where $m$ = 1/2, is referred to as 'spin-up' or the α state, whereas, the another state with $m$ = -1/2, is referred to as 'spin-down' or the β state (Keeler 2002). The energy difference between two states is defined by a frequency called '**Larmor frequency**', which depends on the gyromagnetic ratio (γ) and the strength of the external magnetic field ($B_0$), as shown below.

$$\Delta E = \gamma \hbar B_0$$

Where γ = gyromagnetic ratio; ħ = reduced Planck constant $B_0$ = external magnetic field strength

The population of energy states can be explained by Boltzmann equation

$$\frac{N\alpha}{N\beta} \approx 1 - \frac{\gamma\hbar B_o}{kT}$$

where Nα, Nβ = populations of individual states, $k$ = Boltzmann constant, T = Temperature

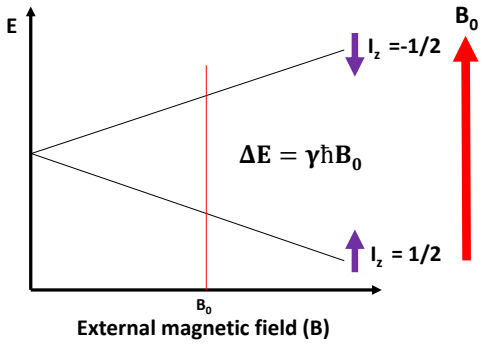



**Figure 1.14 NMR energy levels for spin half nuclei in presence of Magnetic field.** In presence of external magnetic field, spin-half nuclei adopt two different states, characterized by a difference in the energy level.

At equilibrium, the number of spins with α and β state are not equal and thereby build up net magnetic field along the direction of applied field ($B_0$). This gives rise to macroscopically observable Bulk magnetization (M). It could be represented by a vector called magnetization vector with the direction of applied field (z-axis) as described by the vector model (Keeler 2002) (Figure 1.15).

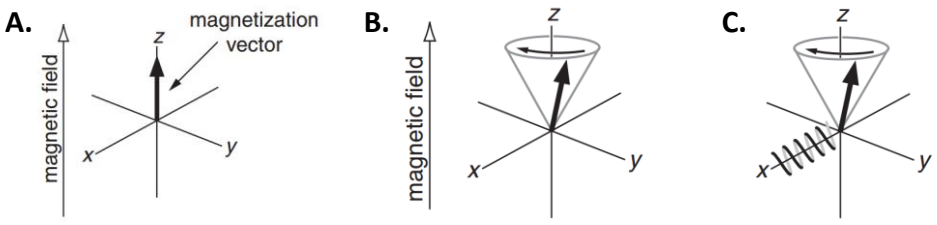



**Figure 1.15 The vector model of NMR. A.** Magnetization vector, which is a net magnetization present in the sample at equilibrium, oriented along the direction of magnetic field. **B.** The precision of the magnetization vector caused by application of radiofrequency pulse along the x-axis. **C.** The precision of the magnetization vector induces current in the coil oriented at x-axis, recording free induction decay. Adapted from (Keeler 2002).

In order to perturb the orientation of the Bulk Magnetization (M) away from the z-axis, electromagnetic field ($B_1$) can be applied in the transverse plane. In an NMR spectrometer, radiofrequency (rf) pulse with Larmor frequency is used along the x-axis, which orients M towards the –y axis. The length of the rf pulse determines the angle of rotation $\alpha$. Once tilted

away from the z- axis the magnetization vector continues to rotate about the direction of the magnetic field, with the Larmor frequency (ω). This precession of the magnetization vector is detected during an NMR experiment using detection coil and called as 'Free Induction Decay' (FID). This FID is a time domain signal representing all the frequencies in the sample. Fourier transformation (FT) can be applied to convert this time domain FID into frequency domain signal which is easier to analyze.

### 1.2.2 The chemical shift

The precise Larmor frequency for the energy transition is dependent on the effective magnetic field experienced by a nucleus. This field is affected by surrounding electrons, resulting into reduced effective nuclear magnetic field. This effect is called as 'electron shielding'. Electron shielding is dependent on the chemical environment of the atom in question. This environment depends on upon the factors such as electronegativity of neighboring atoms, ring currents and the presence of electron withdrawing or donating groups. The unique chemical environment experienced by the nucleus, in turn, reflects the change in the resonance frequency. As a result, nuclei of the same atom type within a molecule resonates with a distinct frequency called as **chemical shift**, because of differences in the electron density in their respective environment. As expected these changes in the frequency are very small in comparison to applied magnetic field and hence described in the units of ppm (parts per million) with respect to a reference compound tetramethylsilane (TMS) for proton NMR. The chemical shift (δ) is expressed in parts per million (ppm) and is defined as,

$$\delta_{\mathrm{ppm}} = \frac{\omega_{\mathrm{atom}} - \omega_{\mathrm{ref}}}{\omega_{\mathrm{ref}}} * 10^{6}$$

where $\omega_{\mathrm{atom}}$ and $\omega_{\mathrm{ref}}$ are the resonance frequencies for the given atom and reference compound (TMS) respectively.

### 1.2.3 Relaxation

As described in the previous section, in presence of applied magnetic field the nuclear spin exists in an energetic equilibrium, which can be easily perturbed upon application of radiofrequency pulses. The return of the spins to their native state which is a net magnetization along z-axis according to the Boltzmann distribution is referred to as a '**relaxation**'.

There are two processes which facilitate the relaxation of the spins. One is the **T1** or **spin-lattice relaxation** in which spins return to the equilibrium by exchanging energy with their surroundings, which can be defined as

$$\mathrm{M_z}\,(\mathrm{t}) = \mathrm{M_0}\,(1 - \mathrm{e}^{-\mathrm{t}/\mathrm{T1}})$$

where $M_z$ (t) is the magnetization as a function of time t and $M_0$ is the equilibrium magnetization.

Whereas, the another process is called as **T2** or **spin-spin relaxation** which involves dephasing of the transverse magnetization (xy plane) and defined as

$$M_{xy}(t) = M_{xy0}(e^{-t/T2})$$

where $M_{xy}$ (t) is the transverse magnetization as a function of time t and $M_{xy0}$ is the initial transverse magnetization.

T1 and T2 depend on the rotational motion (as defined by correlation time, $\tau_c$) of the molecule which in turn depends on upon the molecular size. Small molecules tumble faster and have high T1 and T2 values whereas large molecules such as proteins tumble slowly and have high T1 but low T2 value (Figure 1.16).

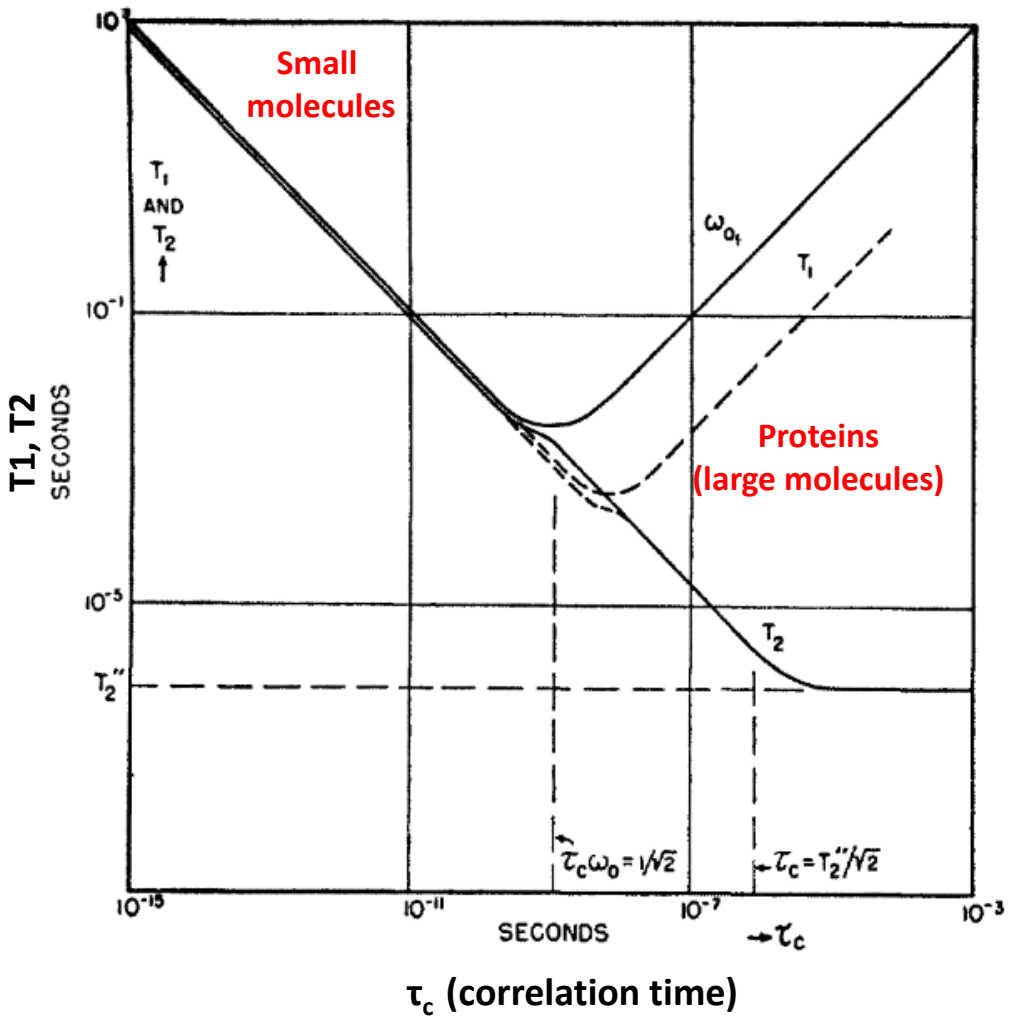



**Figure 1.16 T1 and T2 with respect to correlation time.** Adapted from (Bloembergen, Purcell et al. 1948)

## 1.2.4. Protein NMR

In last few decades, NMR spectroscopy has emerged as a powerful technique to study atomic level information of biological macromolecules in solution. NMR can be successfully utilized to study structure and dynamics of proteins and nucleic acids. Typically, it involves inserting a sample containing biomolecules inside the powerful magnet, which is then excited using radiofrequency waves and the corresponding FID is measured to calculate the distance between nuclei, which in turn could be used to extract the structural and/or dynamics information present in the sample.

Traditionally, NMR was restricted to smaller proteins or protein domains, since with the increase in the size of the protein problem of signal overlap also arises. Also, larger proteins (>30kDa) tend to have faster transverse relaxation rates (R2) (Figure 1.16), which poses a problem for signal detection. The introduction of specific isotope labeling and multidimensional NMR experiments has been successful to circumvent this problem to a large extent. Typically, proteins are expressed recombinantly by growing bacteria in the medium containing $^{13}C$ and $^{15}N$ as the only source of carbon and nitrogen, respectively. Such samples can be used to record heteronuclear multidimensional NMR experiments to provide improved resolution. Additionally, deuteration can also be used, by growing bacteria in a medium containing $^{2}H_2O$ rather than $H_2O$. It results in reducing proton density, which in turn reduces the transverse relaxation rate and thereby, results in sharper spectral lines and resolved signals.

Usually, the first NMR experiment to be recorded on isotope-labelled proteins is a 2D $^{1}H$-$^{15}N$ heteronuclear single quantum correlation (HSQC) spectrum, often referred to as 'fingerprint of the protein'. This experiment shows one peak for each H-N correlation that is present in the protein sample, which mainly includes backbone amide groups of every amino acid (except proline). This experiment shows if the protein sample is folded or not as well as whether further experiments are likely to work on the sample or not.

The next step is to assign the resonances of the backbone and side chains of all the amino acids of the protein. For this, triple resonance experiments are used. In these experiments, $^{1}H$, $^{15}N$, $^{13}C$ containing atomic nuclei of proteins are linked such a way that frequency of the amide proton can be correlated with bonded carbon atoms of each amino acid. Various types of triple resonance experiments (for example HNCA, HNCACB, CBCACONH, HNCOCA) are available, in which typically the magnetization is transferred through Cα, Cβ, CO of the same and/or previous amino acid. For example, in HNCA experiment, magnetization is transferred from amide proton (HN) of the amino acid to the amide nitrogen (N) followed by the Cα of the same (i) and the previous residue (i-1) in the amino acid sequence. As the carbon chemical shifts are characteristics for each amino acid, these chemical shifts along with information regarding protein sequence could be used to assign backbone resonances of the protein. Whereas, to assign side chain atoms TOCSY (total correlation Spectroscopy) experiments are used. TOCSY experiments allow detection of the nuclei that are connected by a chain of couplings, as it provides through bond correlation via a spin-spin coupling. Hence, TOCSY

experiments facilitate the correlation of $^{1}H$ and $^{13}C$ chemical shifts of all the atoms for each amino acid and thereby, resonance assignments of protein side chains.

Once backbone and side chains of the protein are assigned, one needs to record NOESY (Nuclear overhauser effect spectroscopy) experiments to obtain interatomic distance information. The NOE signal is observed between nuclei which are close in space (≤5-6 Å). NOE effect relies on the fact that neighboring spins contribute to the relaxation of a specific spin via dipole-dipole interaction, in a distance-dependent manner ($1/r^6$). Hence, the intensity of the NOE-cross peak can reflect the distance between two atoms. In the case of proteins, $^{13}C$-edited NOESY and $^{15}N$-edited NOESY are recorded to generate cross-peaks for protons (directly bound to $^{13}C$ and $^{15}N$ respectively) to the protons which are close in space but are not necessarily close in the protein sequence. This derived interatomic distance information forms the basis of the structure determination using NMR.

Apart from short-range NOE-derived distance information (≤6 Å), long-range distance information can also be helpful in NMR structure calculation. For this, residual dipolar couplings (RDCs) and paramagnetic relaxation enhancement (PREs) could be employed. RDCs involve the use of anisotropic solutions such as liquid crystalline media to extract relative orientation of two domains a protein, as they provide relative orientations of internuclear vectors independent of distance in between them (Chen and Tjandra 2012). On the other hand, PREs can provide distance information up to 35 Å from the paramagnetic center, which allows drawing conclusions about domain-domain contacts or transient interactions (Clore and Iwahara 2009). PREs result into the enhancement of the relaxation rates of nuclear spins caused by dipolar interactions between a nucleus and the unpaired electron of paramagnetic center owing to the large gyromagnetic ratio of the unpaired electron.

NMR structure calculation protocol performed by using CYANA (Guntert 2004) employs restrained molecular dynamics simulations and simulated annealing. The protocol uses a target function combining different potential energy functions such as bond lengths, angles, electrostatic and van der Waals forces as a force field. Distance information obtained from the NOESY experiments, torsion angle information derived from chemical shifts (Shen and Bax 2013) as well as amino acid sequence information are incorporated during the structure calculation. The protocol begins with approximately 100 randomized structures at high temperature to avoid local energy minima. The temperature is subsequently reduced in the following steps as well as restraints are varied to have minimum violations, resulting in a lower value of target function via simulated annealing. The incorporated NOEs in the structure calculation process could be assigned manually or by using automated NOESY peak assignment protocol provided in CYANA. This protocol combines the use of 3D-structure based filters as well as ambiguous distance restraints along with network anchoring and constraint combination during structure calculation process. Unambiguously assigned NOESY cross peaks are used in an iterative manner to lower the energy of the system as well as the ambiguity of NOE assignments. The output structures can be subjected for explicit solvent refinements to enhance surface electrostatics by using CNS/ARIA program (Linge, Habeck et al. 2003).

The classical NMR structure calculation relies heavily on NOE information. But, the sensitivity of NOESY experiments is relatively lower and hence require protein samples which are stable at higher concentration and/or over a longer period in order survive longer data acquisition. But this is not feasible for all the protein samples. For example, for proteins with aggregation-prone behavior generally, generate low-quality NOESY spectra. CS-ROSETTA *de novo* structure determination protocol circumvents the need of NOESY analysis for the structure calculation (Shen, Lange et al. 2008). Instead, it relies on the fact that chemical shifts and amino acid sequence of the proteins contain information about protein 3D structure. The CS-ROSETTA structure calculation program selects the polypeptide fragments on the basis of chemical shifts and amino acid sequence of the protein. These fragments are then used for *de novo* structure calculation, using a Monte Carlo based assembly process to find compact, low energy folds. The ROSETTA full-atom refinement, involving Monte Carlo minimization and all-atom force field, is then performed to find low energy structures.

One advantage NMR has over other structural biology techniques is that it allows the study of dynamics of protein molecules in solution at residue-specific level. Dynamic properties of proteins can play a key role in its cellular functions ranging from ligand binding to reaction catalysis. In NMR, the spin-relaxation property of nuclear magnetization vector is used to study macromolecular dynamics. Spin relaxation is determined by two factors: longitudinal relaxation (T1) which involves the recovery of magnetization to the equilibrium state after excitation and transverse relaxation (T2), which involves loss of phase coherence of spin components in the transverse plane. Protein internal motions as well as molecular tumbling which ranges from pico- to microseconds time scale can be deciphered using T1 relaxation whereas T2 measurement can reveal information about the chemical exchange. The ratio between T1 and T2 can yield rotational correlation time ($\tau_c$) (time taken by a molecule to rotate by an angle of one radian) and could be used to describe molecular tumbling. Considering globular proteins as a spherical molecules tumbling in the solution, rotational correlation time ($\tau_c$), can be described by stokes law as:

$$\tau_c = \frac{4\pi\eta r^3}{3kT}$$

where $\eta$ is viscosity of solvent, r is the hydrodynamics radius, $k$ is the Boltzmann constant, $T$ is the temperature.

The rotational correlation time (in nanoseconds) depends upon molecular weight and as a rule of thumb, can be thought of approximately half of the molecular weight of the protein (in kDa), given the protein is in aqueous solution at RT. The rotational correlation time from T1 and T2 relaxation time can be approximated by using the equation:

$$\tau_c \approx \frac{1}{4\pi\nu_N}\sqrt{6\frac{T1}{T2} - 7}$$

where $\nu_N$ is the $^{15}N$ resonance frequency in Hertz.

Additionally, {$^{1}$H}-$^{15}$N heteronuclear NOE can be measured to study motions of individual N-H bond vectors. Lower NOE intensity in comparison to the average reflects flexible regions of proteins, such as N and C-termini or flexible loops.

### 1.2.5 NMR analysis of RNA

NMR is also a powerful tool to study nucleic acid structures, dynamics as well as their interactions with respective ligands. In fact, in terms of RNA structures NMR spectroscopy is equally successful, as represented by nearly half of the RNA structures solved so far using NMR.

The RNA sample of interest can be synthesized either chemically or via in vitro transcription using T7 RNA polymerase (Furtig, Richter et al. 2003). Both techniques have their own advantages and disadvantages. The chemical synthesis is well suited for shorter RNAs. It predominantly allows site-specific incorporation of labeled nucleotides along with the possibility to insert non-standard nucleobases, which could be a part of RNA. On the other hand, chemical synthesis could be very expensive and as a result completely $^{13}$C and/or $^{15}$N labeled RNA molecules can be non-affordable. In such cases, in vitro transcription can be used. Nucleotide triphosphates could be used in labeled form, to generate uniformly labeled RNA from DNA template using T7 RNA polymerase. Before conducting NMR experiments, it is necessary to check the purity of the sample. This could be done by running the sample on Polyacrylamide gel electrophoresis under denaturing conditions or by performing High-Performance Liquid Chromatography (HPLC) run.

Assignments of RNA resonances is the first step in order to study the conformation of RNA or to get insights about its interaction with its partner biomolecules. In practice the assignment of RNA resonances is more complicated than of proteins. This complexity comes from the poor chemical shift dispersion of resonances. The presence of only four different types of nucleotides as well as A-form helix as the only dominating secondary structural element, poses similar chemical environment to most of the RNA nucleotides, thereby generating similar chemical shifts with poor dispersion. On the other hand, the presence of hairpins, bulges as well as internal loops can introduce changes in the chemical environments, causing chemical shift dispersion.

In the case of RNA, $^{1}$H 1D NMR spectrum can be very informative. RNA resonances in 10-15 ppm region are the most important, as they function as a ‘fingerprint’ for the RNA structure. These resonances represent imino protons of guanine and uracil residues. Imino protons are generally exchangeable and hence only observed when they are involved in hydrogen bonding by being protected from an exchange with water. Hence, by counting the number of imino proton resonances, it is possible to get information about the number of base pairs (Furtig, Richter et al. 2003). Information on the hydrogen bonding between bases is important for RNA structure determination.

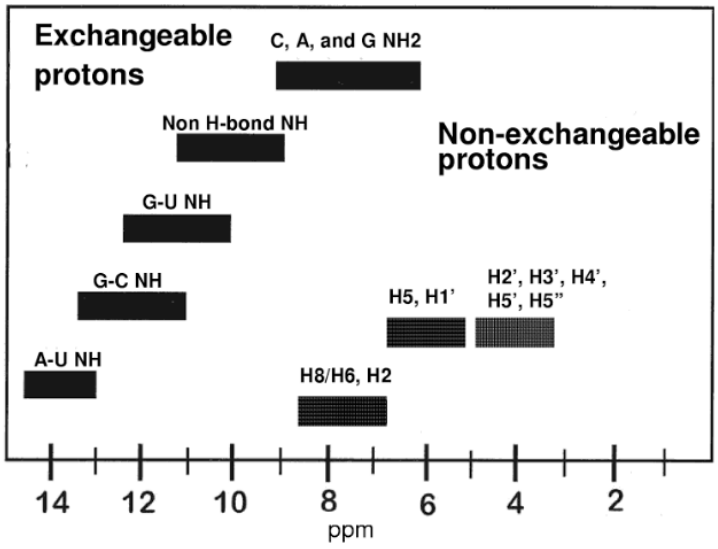



**Figure 1.17 RNA proton resonances.** RNA proton resonances are classified into exchangeable and non-exchangeable. Exchangeable protons in between 10-15 ppm range, are called as a fingerprint of RNA structure, containing information regarding a number of base pairs as well as the pattern. Adapted from (Wüthrich 1986).

It is a common practice to assign proton resonances of AH2, UH5, UH6, CH5, CH6 first. H5 and H6 resonances of uracil and cytosine can be identified using very strong scalar coupling between them observed in a TOCSY experiment. Natural abundance HSQC or HMQC experiments helps to identify AH2 as well as UH5 and CH5 because of distinct carbon chemical shifts. These sharp and distinct signals can be usually distinguishable from aromatic H8/H6. 2D NOESY spectra (with appropriate mixing time) is used to establish sequential assignments using NOE connectivity between H8/H6(n)-H1'(n) and H1'(n)-H8/H6(n+1). H2' proton resonances can be assigned by using DQF-COSY and TOCSY experiments and can be cross-checked from NOESY using strong intra-residue H1'-H2' NOE connectivity. Further assignments of sugar resonances can be achieved by using TOCSY experiments via scalar connectivity of H2'/2"-H3'-H4'-H5'/5". These proton assignments are then confirmed with scalar correlations with carbon and phosphorus nuclei.

## 1.2.6. Protein-RNA interaction by NMR

NMR spectroscopy can be readily used to monitor protein- RNA interaction by using chemical shift perturbation mapping or titration experiments. As chemical shifts depend on the chemical environment of the nucleus, changes in chemical shifts upon addition of ligand can be used to map the interface of a macromolecular complex. This could be achieved by either adding protein into RNA or vice versa. Mostly, series of $^{1}H$-$^{15}N$ HSQC spectrum are recorded on protein sample with increasing concentration of RNA to follow chemical shift perturbations of proteins. In the case of RNA, chemical shift perturbations of imino protons by $^{1}H$ 1D NMR

or H5-H6 cross-peaks of pyrimidines by 2D $^{1}H$-$^{1}H$ TOCSY spectra could by tracked with increasing concentration of protein, to observe the interaction.

During complex formation, the protein, and the RNA are in equilibrium with their free and bound states. The affinity of interaction, termed as the dissociation constant ($K_d$), describes the equilibrium between these states. NMR has three characteristic exchange regimes, which depend on the exchange rate of complex formation, $k_{ex}$, and the difference in the resonance frequencies for free and bound states, $\Delta\nu$ (Dominguez, Schubert et al. 2011). The different exchange regimes are defined as following:

| | |
|---|---|
| $\Delta\nu >> k_{ex}$ | slow exchange |
| $\Delta\nu \approx k_{ex}$ | intermediate exchange |
| $\Delta\nu << k_{ex}$ | fast exchange |

Ligand binding in slow exchange is characterized by two sets of signals, one corresponding to the free state of the protein and other bound to RNA. Slow exchange regime is characteristic of protein-RNA complexes with dissociation constant ranging from 0.5 to 250 nM.

Intermediate exchange induces line broadening of NMR signals upon addition of a partner and thus results in a decrease in signal intensity often beyond detection. Intermediate exchange regime is observed for complexes with a dissociation constant in the range of 400 nM to 2 µM.

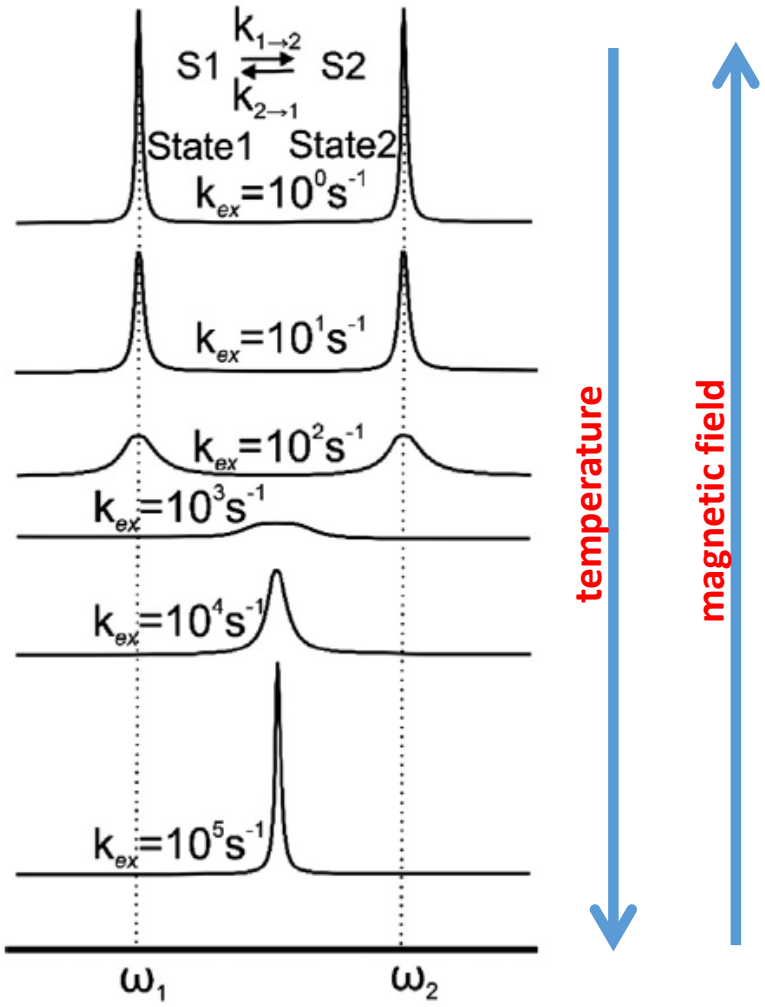



**Figure 1.18 Types of exchange regimes observed for investigating protein-ligand interactions by NMR.** Study of Protein-ligand interaction heavily depends on exchange regime of binding, as it determines the quality of the NMR spectra for the complex. Exchange regimes could be modulated by the change in the temperature and/or magnetic field [Figure adapted from (Gobl, Madl et al. 2014)].

On the other hand, fast exchange regime is characterized by the average chemical shift of free and bound state corresponding to the population of each state resulting in the shift of the signals towards the bound state until saturation. Typically, protein-RNA complexes with dissociation constant higher than 10-100 µM fall under fast exchange regime. Thus, exchange regime for the complex formation turns out to be an important parameter for protein-RNA complex structure determination (Dominguez, Schubert et al. 2011).

Change in temperature can be used to alter the exchange regime, as higher temperature makes exchange faster whereas lowering the temperature results in slower exchange. Also, as the exchange regime depends on the difference in the resonance frequencies of the free and bound states, which in turn depends on the magnetic field, it is also possible to alter the exchange regime by changing magnetic field (Figure 1.18).

## 1.3 Scope of the thesis

Multi-domain RNA-binding proteins and their target *cis*-regulatory RNA sequences lie at the heart of the splicing regulation. Combinatorial action of these elements determine the fate of spliceosomal assembly and thereby splicing reaction. The aim of this thesis is to understand the structural and molecular mechanism of splicing repression by LS2, while it resembles the well-known splicing activator dU2AF50.

An important question is to understand if and how structural and dynamic properties of LS2 RNA binding domains (RRM domains) diverge from dU2AF50 RRM domains in order to be functionally different. It is necessary to check whether LS2 and dU2AF50 RRM domains adopt similar fold, given the high degree of sequence conservation among them. The interdomain linker of LS2 RRM domains is studied regarding its structural or functional significance in comparison to that of dU2AF50.

Given that the guanosine-rich LS2 RNA ligands can potentially adopt G-quadruplex structures various LS2 target sequences designed from SELEX motif sequences are studied by NMR and additional techniques. In order to understand structural and functional features of the interaction, NMR titration studies are conducted between LS2 RRM domains and guanosine-rich ligands. The mode of RNA binding by LS2 RRM domains and dU2AF50 RRM domains to their respective ligands is also compared.

To study the structural features of the LS2-RNA interaction a multi-disciplinary integrated structural biology approach is employed. As a major technique solution NMR-spectroscopy is used, complemented by small angle X-ray scattering (SAXS) and X-ray crystallography. Other biophysical methods such as isothermal titration calorimetry (ITC), circular dichroism (CD), static light scattering (SLS) provide additional information on the interaction.

The structural and biochemical data obtained in this study is used to guide mutational analysis to study the role of the LS2-RNA interaction *in vivo* based on alternative splicing assays.

# Chapter 2 Materials and methods

## 2.1 Chemicals and consumables

Chemicals were purchased from Merck, VWR international, Roth, Roche, Sigma-Aldrich or SERVA unless stated otherwise. Restriction enzymes and related buffers were ordered from New England Biolabs whereas, *Pfu* polymerase and the buffer was purchased from Thermo Fisher Scientific. Medium components for preparing bacterial growth cultures were purchased from Sigma-Aldrich. NMR active isotopically labeled medium components were purchased from CIL, Euriso-top or Sigma-Aldrich. Chromatography column and related materials were purchased from GE Healthcare except for Nickel-NTA resin, which was purchased from QIAGEN. DNA oligonucleotides were synthesized by MWG-Biotech (Eurofins). TEV (Tobacco Etch Virus) protease and SUMO protease (from *S. cerevisiae*) were supplied from Protein expression and purification4 facility, HMGU. TMPyP4 was purchased from Santa Cruz Biotechnology.

## 2.2 Molecular biology

### 2.2.1 Bacterial strains

**Table 2-1 Bacterial strains with their genotypes (adapted from openwetware.org).**

| E. Coli Strain | Genotype |
|---|---|
| **DH5α** | $F^-$ endA1 glnV44 thi-1 recA1 relA1 gyrA96 deoR nupG purB20 ϕ80dlacZΔM15 Δ(lacZYA-argF) U169, hsdR17($r_K^-m_K^+$), $\lambda^-$ |
| **XL1-Blue** | endA1 gyrA96($nal^R$) thi-1 recA1 relA1 lac glnV44 F'[ ::Tn10 $proAB^+$ $lacI^q$ Δ(lacZ)M15] hsdR17($r_K^-$ $m_K^+$) |
| **BL21 (DE3)** | B $F^-$ ompT gal dcm lon $hsdS_B(r_B^-m_B^-)$ λ(DE3 [lacI lacUV5-T7p07 ind1 sam7 nin5]) $[malB^+]_{K-12}(\lambda^S)$ |

### 2.2.2 Plasmids for recombinant protein expression

**Table 2-2 Plasmids for recombinant protein expression.**

| Name | boundaries | Vector | Resistance | Restriction sites | Purification | Source |
|---|---|---|---|---|---|---|
| LS2 | full length | pGEX GST | Amp | - | - | Taliaferro et al, 2011 |
| LS2RRM1,2 | 101-319 | pGEX GST | Amp | - | soluble | Taliaferro et al, 2011 |
| LS2RRM1,2 ΔLinker | 101-319 Δ205-240 | pGEX GST | Amp | - | soluble | Taliaferro et al, 2011 |

| | | | | | | |
|---|---|---|---|---|---|---|
| LS2RRM1,2 | 110-319 | pET Sumo | Amp | NcoI, XhoI | inclusion body | this study |
| LS2RRM1,2 ΔLinker | 110-319 Δ205-240 | pET Z | Kan | NcoI, XhoI | inclusion body | this study |
| LS2RRM1 | 110-204 | pET Z | Kan | NcoI, XhoI | soluble | this study |
| LS2RRM2 | 242-319 | pET MBP | Kan | NcoI, XhoI | soluble | this study |
| LS2RRM1,2 | 110-325 | pET Sumo | Amp | NcoI, XhoI | inclusion body | this study |
| LS2 Linker-RRM2 | 204-325 | pET Sumo | Amp | NcoI, XhoI | inclusion body | this study |
| LS2 exRRM1 | 101-221 | pET Sumo | Amp | NcoI, XhoI | inclusion body | this study |
| LS2 ULM | 53-82 | pET MBP | Kan | NcoI, XhoI | soluble | this study |
| LS2 URRM1,2 | 53-325 | pET Sumo | Amp | NcoI, XhoI | inclusion body | this study |
| LS2 RRM1,2,3 | 110-449 | pET Z | Kan | NcoI, XhoI | inclusion body | this study |
| dU2AF50 | full length | pGEX GST | Amp | - | - | Taliaferro et al, 2011 |
| dU2AF50 RRM1,2 | 92-286 | pET Z | Kan | BspHI, XhoI | soluble | this study |
| dU2AF50 exRRM1,2 | 85-290 | pET Z | Kan | BspHI, XhoI | soluble | this study |
| 50 ULM | 37-64 | pET Z | Kan | NcoI, XhoI | soluble | this study |
| dU2AF38 | full length | pRSET-a | Amp | - | - | Taliaferro et al, 2011 |
| dU2AF38 UHM | 43-148 | pET Z | Kan | NcoI, XhoI | inclusion body | this study |

## 2.2.3 Cloning and site-directed mutagenesis

All the above-mentioned constructs were subcloned by standard cloning methods as described (J. Sambrook 2001). The steps include PCR amplification of the gene of interest, restriction digestion of PCR product as well as vector DNA and subsequent ligation of DNA fragments. Agarose gel electrophoresis was used to separate DNA fragments and purified using Wizard SV Gel and PCR Clean-Up System (Promega).

**Table 2-3 PCR reaction mixture.**

| Components | Volume |
|---|---|
| **Template (100ng/µL)** | 0.5 µL |
| **dNTPs (25 mM)** | 0.5 µL |
| **Forward primer (100 pmol)** | 0.5 µL |
| **Reverse primer (100 pmol)** | 0.5 µL |
| **Pfu reaction buffer+$MgSO_4$ (10x)** | 5 µL |
| **$H_2O$** | 42.5 µL |
| **Pfu DNA polymerase (2.5 U/µL)** | 0.5 µL |

**Table 2-4 PCR thermocycling program.**

| Temperature (°C) | Time (mins) | No. of cycles |
|---|---|---|
| **95** | 3 | 1 |
| **95** | 0.5 | 35 |
| **56** | 0.5 | |
| **72** | 1 | |
| **72** | 10 | 1 |
| **10** | ∞ | |

Site-specific mutagenesis was carried out by PCR amplification using primers containing mutated sequence flanked by overlapping nucleotides. The primers were designed according to Quick-change II XL site-directed mutagenesis kit manual from Agilent technologies.

**Table 2-5 Site-specific mutagenesis reaction mixture.**

| Components | Volume |
|---|---|
| **Template (100ng/µL)** | 0.5 µL |
| **dNTPs (10 mM)** | 1 µL |
| **Forward primer (100 pmol)** | 0.1 µL |
| **Reverse primer (100 pmol)** | 0.1 µL |
| **P*fu* reaction buffer+$MgSO_4$ (10x)** | 5 µL |
| **DMSO** | 2.5 µL |
| **$H_2O$** | 40.8 µL |
| **P*fu* DNA polymerase (2.5 U/µL)** | 1 µL |

**Table 2-6 PCR program for site-specific mutagenesis.**

| Temperature (°C) | Time (mins) | No. of cycles |
|---|---|---|
| **95** | 3 | 1 |
| **95** | 0.8 | 18 |
| **58** | 0.8 | |
| **68** | 12 | |
| **68** | 7 | 1 |
| **10** | ∞ | |

## 2.2.4 Transformation and plasmid DNA isolation

Ligation products were transformed into chemically competent *E. coli* strains of DH5α or XL-1 Blue as described (J. Sambrook 2001). Plasmid DNA was isolated from overnight grown cells by using Pure Yield Plasmid Maniple System kit (Promega). All plasmids were verified by sequencing via GATC Biotech.

## 2.3 Protein expression

Respective plasmids were used to transform BL21(DE3) cells and further selected using appropriate antibiotic agar plates. For expression, bacterial cells were then grown into 1L of LB medium or minimal M9 medium (see Table 2-7). For induction, 1mM IPTG was used at an OD 600 nm of ~0.6 followed by incubation at 16 $^{o}$C (soluble protein purification) or 37 $^{o}$C (Inclusion body purification) (see Table 2-2) for 16 h on a rotary shaker (220 rpm). Cells were harvested by centrifugation at 4000 rpm for 20 minutes at 4 $^{o}$C, which were either directly used for protein purification or stored at -20 $^{o}$C till purification.

**Table 2-7 Bacterial growth medium along with their composition.**

| Medium | Component | Amount per liter |
|---|---|---|
| **Lysogeny broth (LB) rich medium** | Tryptone | 10 g |
| | NaCl | 10 g |
| | yeast extract | 5 g |
| | antibiotic | 50 mg |
| **M9 minimal medium** | $Na_2HPO_4$ | 6 g |
| | $KH_2PO_4$ | 3 g |
| | NaCl | 0.5 g |
| | $^{15}NH_4Cl$ | 0.5 g |
| | Glucose | 4 g |
| | Or U-[$^{13}$C]-glucose | 2 g |
| | $MgSO_4$ (1M) | 1 ml |
| | $CaCl_2$ (1M) | 0.3 ml |
| | Biotin | 1 mg |
| | Thiamine | 1 mg |
| | 100x trace elements stock* | 10 ml |
| ***100x trace elements stock** | EDTA, pH 7.5 | 5 g |
| | $FeCl_3$. $6H_2O$ | 0.83 g |
| | $ZnCl_2$ | 84 mg |
| | $CuCl_2$. 2 $H_2O$ | 13 mg |
| | $CoCl_2$. $6H_2O$ | 10 mg |
| | $H_3BO_3$ | 10 mg |
| | $MnCl_2$. $6H_2O$ | 1.6 mg |

## 2.4 Protein purification

Depending on the solubility and/or the degradation tendency of protein constructs, either soluble or denaturing purification protocol was used for the purification. Both purification protocols, as well as buffers used for the respective purification, are listed below.

### 2.4.1 Protein purification protocol for soluble protein constructs

Cell pellets were resuspended in 40 mL loading buffer supplemented with AEBSF hydrochloride inhibitor, DNase1 (1 mg/ml) and Magnesium sulfate (10 mM). Cells were lysed using 3 cycles of 3-minutes of sonication using 52% power output. The lysate was centrifuged for 45 minutes at 19000 rpm at 4 °C.

**Table 2-8 Buffers for soluble protein purification protocol.**

| Buffer | Component | Concentration |
|---|---|---|
| **Load buffer** | Tris-HCl pH 8.0 | 20 mM |
| | NaCl | 500 mM |
| | Imidazole | 5 mM |
| | β-mercaptoethanol | 1 mM |
| **Wash buffer** | Tris-HCl pH 8.0 | 20 mM |
| | NaCl | 500 mM |
| | Imidazole | 25 mM |
| | β-mercaptoethanol | 1 mM |
| **Elution buffer** | Tris-HCl pH 8.0 | 20 mM |
| | NaCl | 500 mM |
| | Imidazole | 250 mM |
| | β-mercaptoethanol | 1 mM |
| **Dialysis buffer** | Tris-HCl pH 8.0 | 20 mM |
| | NaCl | 150 mM |
| | β-mercaptoethanol | 1 mM |
| **Gel filtration buffer** | Tris-HCl pH 8.0 | 20 mM |
| | NaCl | 150 mM |
| | β-mercaptoethanol | 1 mM |
| **NMR buffer** | Potassium phosphate buffer pH 6.5 | 20 mM |
| | NaCl | 50 mM |
| | DTT | 5 mM |

The cleared lysate was loaded twice onto $Ni^{+2}$-NTA resin column pre-equilibrated with load buffer. 30 ml of wash buffer was used to remove nonspecifically bound proteins from the column. Bound protein was then eluted with 30 ml of elution buffer, which was further dialyzed overnight using dialysis buffer along with TEV protease (approximately 2mg) to cleave $His_6$ bound solubility tag from the target protein. The TEV protease, cleaved solubility tag, and uncleaved protein were separated by applying the sample again onto the $Ni^{+2}$ resin. The flow-through containing free target protein was further concentrated up to 5 ml using an Amicon Ultra-15(Millipore) concentrator and gel filtration was performed using Superdex 75 16/60 column in gel filtration buffer for further purification. Finally, samples containing pure protein were buffer exchanged to NMR buffer using Amicon concentrator.

## 2.4.2 Inclusion body purification protocol

During inclusion body purification protocol, after sonication and subsequent centrifugation, cell pellet containing inclusion bodies was resuspended in 50 ml of denaturing load buffer. Sonication with 52% power output was used to dissolve inclusion bodies completely. Centrifugation at 19,000 rpm for 45 minutes was performed to obtain a clear solution. This clear solution was further loaded onto $Ni^{+2}$ resin and followed by washing with 30 ml of denaturing wash buffer. Elution was performed with at least 60 ml of denaturing elution buffer to prevent concentration dependent aggregation during a subsequent refolding process performed overnight using dialysis buffer. Sumo protease was used to cleave the $His_6$ bound Sumo tag at RT for 30 minutes. Subsequent steps were same as mentioned for soluble protein purification protocol.

**Table 2-9 Buffers for Inclusion body protein purification protocol.**

| Buffer | Component | Concentration |
|---|---|---|
| **Denaturing load buffer** | Tris-HCl pH 8.0 | 20 mM |
| | NaCl | 500 mM |
| | Imidazole | 5 mM |
| | β-mercaptoethanol | 1 mM |
| | Urea | 8M |
| **Denaturing wash buffer** | Tris-HCl pH 8.0 | 20 mM |
| | NaCl | 500 mM |
| | Imidazole | 25 mM |
| | β-mercaptoethanol | 1 mM |
| | Urea | 8M |
| **Denaturing elution buffer** | Tris-HCl pH 8.0 | 20 mM |
| | NaCl | 500 mM |
| | Imidazole | 250 mM |
| | β-mercaptoethanol | 1 mM |
| | Urea | 8M |

## 2.4.3. Protein analysis

The success of purification steps, as well as purity of the sample, were monitored by SDS-PAGE as described (Laemmli 1970). 15% polyacrylamide gels were used for electrophoretic protein band separation, followed by Coomassie-Blue staining method to detect protein bands. Nanodrop was used to measure protein concentration using absorption wavelength at 280 nm. To check the absence of contaminating nucleic acids, the ratio of absorption at wavelengths 260 nm to 280 nm was used. Samples with ratio $A_{260}/A_{280}$ below 0.6 were used for data collection.

## 2.5 RNA oligonucleotides

All RNA oligonucleotides were purchased in lyophilized form from IBA and were dissolved in sterile water.

**Table 2-10 RNA oligonucleotide sequences.**

| No. | Name | Sequence 5'-3' |
|---|---|---|
| **1** | 21mer | GGGGAGGAGGGGGGCGUAUGA |
| **2** | 14mer | GGGUGGUGGAGGGG |
| **3** | 8mer | GGUGGUGG |
| **4** | 5mer | GGUGG |
| **5** | U9 | UUUUUUUUU |
| **6** | Non-G 7mer | CGUAUGA |

## 2.6 Biophysical methods

### 2.6.1 Isothermal titration calorimetry

PEAQ- ITC (Malvern instruments) was used for calorimetric titrations performed at 298 K. protein samples were dialyzed extensively in the buffer used for the experiment. The same buffer was used to resolubilize the RNA and to provide the baseline as required. The sample cell was filled with 200 µL of 20 µM RNA whereas injection syringe contained 70 µL of 200 µM protein. Titration consisted of 19 injections of 2 µL with a 2.5-minute spacing between each injection. Data was fitted to one-site binding model using Microcal PEAQ-ITC Analysis software.

### 2.6.2 Circular Dichroism

The CD spectroscopy was performed using Jasco J-180 spectrometer using 1mm pathlength cuvette. The experiments were conducted using wavelength scan of 200-350 nm, 3 number of scans with 1 sec response time and a 2 nm bandwidth. 300 µL of 60 µM RNA was used for the measurements carried out at RT.

### 2.6.3 Static light scattering (SLS)

SLS experiments were performed using Superdex 75 16/60 size-exclusion column (GE Healthcare). Viscotek Tetra Detector Array (Malvern instruments) was used to detect refractive index (RI) as well as light scattering by the sample components. The column was pre-equilibrated with the buffer used and sample volume of 100 µL was loaded on to it with

a flow rate of 0.5 ml/min. For calibration 4 mg/ml BSA was used. For data analysis, OmniSEC software was used.

## 2.7 NMR

### 2.7.1 NMR experiments

All the NMR experiments were recorded on 500, 600, 750, 800, 900 MHz Bruker Avance NMR spectrometers, equipped with cryogenic (except for 750 MHz) triple resonance gradient probes. Backbone and side chain assignments were performed using experiments as described in (Sattler, Schleucher et al. 1999). Aromatic side chains were assigned as described by (Yamazaki, Formankay et al. 1993) and stereospecific assignments were carried out using the method described by (Neri, Szyperski et al. 1989). For spectra processing NMRPipe/Draw (Delaglio, Grzesiek et al. 1995) or Topspin 3.2 (Bruker software) were used whereas for analysis Sparky (T. D. Goddard and D. G. Kneller, SPARKY 3, University of California, San Francisco) was used.

**Table 2-11 summary of NMR experiments.**

| | **assignments** | **experiments** | **Experimental conditions** |
|---|---|---|---|
| **LS2RRM1** | Backbone | $^{1}$H-$^{15}$N HSQC<br>HNCA, HNCACB | 250 µM protein in 20 mM potassium phosphate buffer pH 6.5, 300 mM NaCl, 5 mM DTT at 298 K |
| | Side chain | HCCH-TOCSY,<br>H(CC)(CO)NH,<br>(H)CC(CO)NH,<br>$^{1}$H-$^{13}$C HSQC<br>(Hβ)Cβ(CγCδ)Hδ<br>(Hβ)Cβ(CγCδCε)Hε | 250 µM protein in 20 mM potassium phosphate buffer pH 6.5, 300 mM NaCl, 5 mM DTT at 298 K |
| | NOESY | $^{15}$N-edited NOESY<br>Aliphatic $^{13}$C NOESY<br>Aromatic $^{13}$C NOESY | 250 µM protein in 20 mM potassium phosphate buffer pH 6.5, 300 mM NaCl, 5 mM DTT at 298 K |
| **LS2RRM2** | Backbone | $^{1}$H-$^{15}$N HSQC<br>HNCA, HNCACB | 500 µM protein in 20 mM potassium phosphate buffer pH 6.5, 50 mM NaCl, 5 mM DTT at 298 K |
| | Side chain | HCCH-TOCSY,<br>H(CC)(CO)NH,<br>(H)CC(CO)NH,<br>$^{1}$H-$^{13}$C HSQC<br>(Hβ)Cβ(CγCδ)Hδ<br>(Hβ)Cβ(CγCδCε)Hε | 500 µM protein in 20 mM potassium phosphate buffer pH 6.5, 50 mM NaCl, 5 mM DTT at 298 K |
| | NOESY | $^{15}$N-edited NOESY<br>Aliphatic $^{13}$C NOESY<br>Aromatic $^{13}$C NOESY | 500 µM protein in 20 mM potassium phosphate buffer pH 6.5, 50 mM NaCl, 5 mM DTT at 298 K |

| | | | |
|---|---|---|---|
| **LS2Linker -RRM2** | Backbone | $^{1}$H-$^{15}$N HSQC<br>HNCA,<br>HNCACB,CBCA(CO)NH | 1 mM protein in 20 mM potassium phosphate buffer pH =6.5, 50 mM NaCl, 5 mM DTT at 298 K |
| | Side chain | HCCH-TOCSY,<br>H(CC)(CO)NH,<br>(H)CC(CO)NH,<br>$^{1}$H-$^{13}$C HSQC | 1 mM protein in 20 mM potassium phosphate buffer pH 6.5, 50 mM NaCl, 5 mM DTT at 298 K |
| | NOESY | $^{15}$N-edited NOESY<br>Aliphatic $^{13}$C NOESY<br>Aromatic $^{13}$C NOESY | 1 mM protein in 20 mM potassium phosphate buffer pH 6.5, 50 mM NaCl, 5 mM DTT at 298 K |
| **RRM1,2** | Backbone | $^{1}$H-$^{15}$N HSQC<br>HNCA, HNCACB<br>$^{15}$N-edited NOESY | 150-300 µM protein in 20 mM potassium phosphate buffer pH 6.5, 300 mM NaCl, 5 mM DTT at 288 K |
| **dU2AF50 RRM1,2** | Backbone | $^{1}$H-$^{15}$N HSQC<br>HNCACB, CBCA(CO)NH<br>(H)CC(CO)NH<br>$^{15}$N-edited NOESY | 800 µM protein in 20 mM potassium phosphate buffer pH 6.5, 50 mM NaCl, 5 mM DTT at 298 K |

Backbone Amide $^{15}$N relaxation data was recorded at 500 MHz and 298 K. Steady state heteronuclear {$^{1}$H}-$^{15}$N NOE spectra were recorded with and without 3s of $^{1}$H saturation. Relaxation rates and error calculations were determined according to (Farrow, Muhandiram et al. 1994). Whereas, to determine $^{15}$N T1 and T1rho relaxation time interleaved pseudo-3D HSQC-based experiments with different relaxation delays was used. Data analysis was performed by using Sparky software. $^{15}$N R2 rates were determined from R1ρ as described by (Massi, Johnson et al. 2004).

$$\mathrm{R2} = \frac{\mathrm{R1}\rho - \mathrm{R1}\cos^2\theta}{\sin^2\theta}$$

and

$$\theta = \arctan\left(\frac{\omega 1}{\Omega}\right)$$

where $\Omega$ is resonance offset and $\omega 1$ is the spin-lock frequency.

## 2.7.2 Structure calculation

The solution structure of LS2 RRM2 and LS2 linker-RRM2 were generated using Cyana 3.0 structure calculation program with the automated NOESY assignment protocol (Guntert 2004). Intramolecular NOEs were used as a distance restraint, whereas, chemical shift information was used to generate backbone torsion angle restraints using TALOS+ (Shen, Vernon et al. 2009). Structure determination protocol uses these restraints during molecular dynamics and simulated annealing run using a target function which combines different potential energy functions. To determine LS2 RRM1 structure CS-ROSETTA structure

calculation protocol was used (Shen, Lange et al. 2008), which also take into account few evident long-range NOEs in between β-strands.

### 2.7.3 NMR titration analysis

For RNA titration, either uniformly $^{15}N$ or $^{15}N$-$^{13}C$ labeled sample was used. The initial protein concentration was 50 µM-100 µM. $^{1}H$-$^{15}N$ HSQC as well RNA 1D with Watergate was recorded upon stepwise addition of RNA (in $H_2O$), to follow Chemical shift perturbation/ Intensity loss. Typical RNA titration series used steps of 0, 0.1, 0.2, 0.4 up to 1,2 or 4 molar equivalents of RNA to protein. Chemical shift perturbations (CSP) were calculated according to:

$$CSP = \sqrt{\left(\delta H_{ppm}\right)^2 + \left(0.2 * (\delta N_{ppm})\right)^2}$$

where,

- $\left(\delta H_{ppm}\right)$is the proton chemical shift change

- $(\delta N_{ppm})$is the nitrogen chemical shift change

## 2.8 SAXS

SAXS experiments were recorded on a Rigaku BioSAXS1000 instrument along with Rigaku HF007 microfocus rotating anode with a copper target (40 kV, 30 mA). Sample measurement was performed in 4/8 900 second frames compared to check for radiation damage, averaged and solvent subtracted by the SAXSLab software (v3.0.2). Sample concentration were 1,2,3 mg/ml. Data processing was performed using ATSAS software suit (Petoukhov, Franke et al. 2012). The experiments were performed in 20 mM potassium phosphate buffer pH 6.5, 300 mM NaCl, 5 mM DTT at 4 °C.

## 2.9 Crystallization trials

Crystallization trials were performed by using commercial screens from QIAGEN and Hampton Research at the X-ray crystallography platform located at Helmholtz Zentrum München. Crystal screens were incubated at either 4 °C or RT.

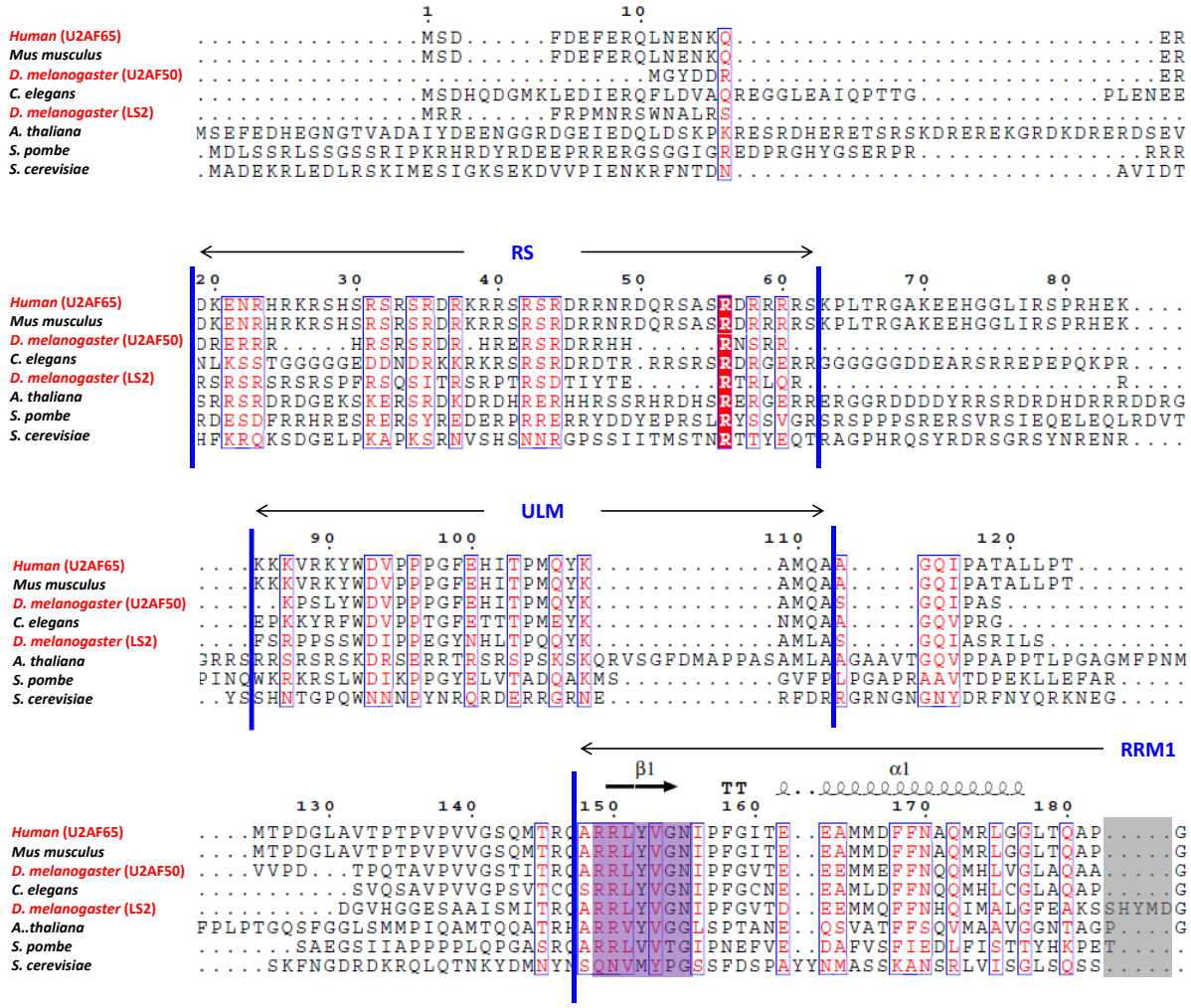
# Chapter 3 Results

## 3.1 Characterization of poly-G RNA recognition by LS2

### 3.1.1 Analysis of LS2 RRM domains and linker

#### 3.1.1.1 Insights from the sequence analysis

To understand functional aspects of any molecular system, careful examination of sequence and concomitant structure is necessary. To perform sequence analysis multiple sequence alignment of LS2 was carried out with U2AF$^{LS}$ from other eukaryotic species (Figure 3.1). The information from sequence alignment was used to define domain boundaries as well as to find out peculiar differences in the amino acid composition of LS2.



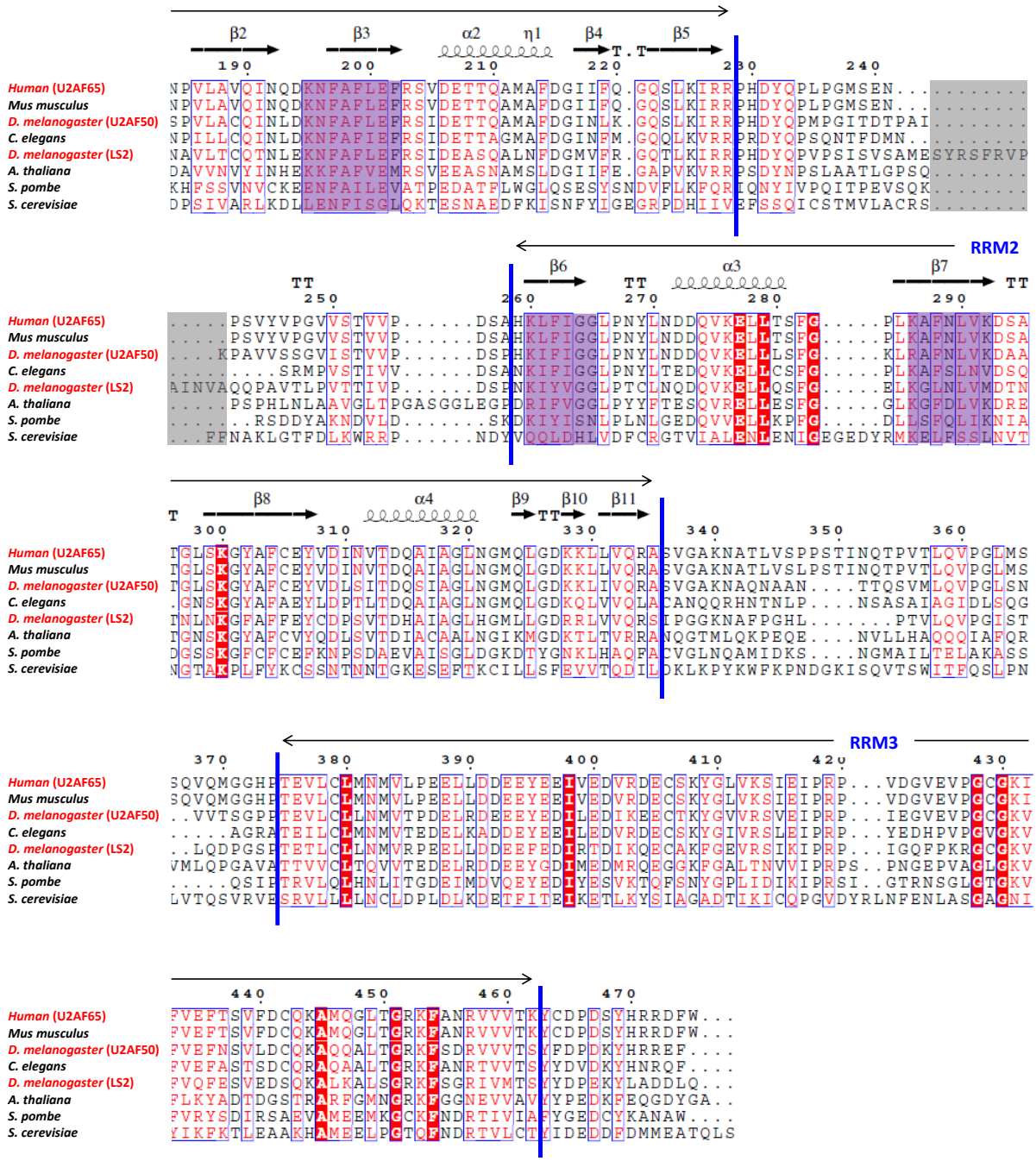
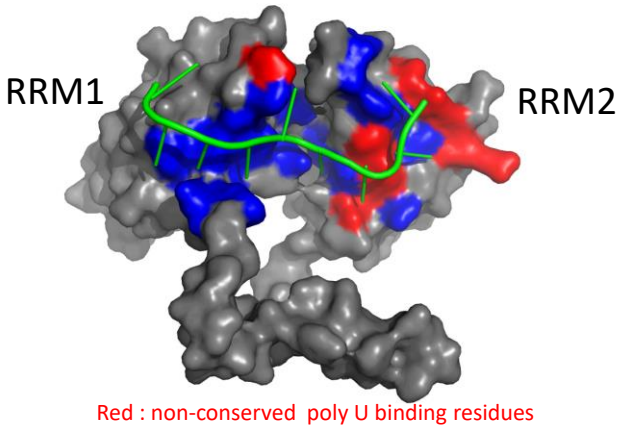



**Figure 3.1 Multiple sequence alignment of U2AF$^{LS}$ and LS2.** U2AF$^{LS}$ is well conserved among all eukaryotic species and it consists of three RRM domains and RS, ULM peptide motifs. LS2 (a paralog of U2AF$^{LS}$ from Drosophila testes) also shows high degree domain conservation with U2AF$^{LS}$, except longer loop between α1 helix and β2 strand as well as a longer linker between RRM1 and RRM2. Violet boxes highlight RNP sites of RRM1 and RRM2 whereas gray boxes indicate additional residues in LS2 which are absent in U2AF$^{LS}$ paralogs from eukaryotic species. Sequences were aligned using Clustal Omega (Sievers, Wilm et al. 2011) and analyzed using ESPript 3 (Robert and Gouet 2014). Sequence conservation is shown by red characters on a white background and if the residues are strictly conserved in the column, they are represented by a white character on a red background.

Apart from conserved RNP sites on both RRM domains, multiple sequence alignment identified two additional potential RNA binding patches in LS2 amino acid sequence. The loop connecting α1 helix and β2 strand contains additional five amino acids, which also includes one histidine and one tyrosine residue. Additionally, there is neighboring lysine residue, which together forms a longer loop which could potentially interact with RNA nucleotides.

On the other hand, the linker between RRM1 and RRM2 is notably long in comparison to eukaryotic U2AF$^{LS}$ paralogs. More importantly, it has a patch of aromatic and charged residues which have potential to interact with RNA nucleotides. It would be interesting to check if these LS2-specific amino acids have a tendency to adopt any secondary structure and/or play any role in RNA binding.

Given the conserved domain arrangement of LS2, one could speculate that, in order to discriminate between poly-G and poly-U RNA nucleotides, RNA binding residues of LS2 RRM domains should differ from poly-U binding residues of U2AF65. In order to check the extent of conservation of poly-U binding residues in LS2, U2AF65 RNA-binding residues [obtained from (Sickmier, Frato et al. 2006)] were mapped on LS2 RRM1,2 homology model. This model was generated using SWISS-MODEL webserver (https://swissmodel.expasy.org/) (Arnold, Bordoli et al. 2006) and U2AF65 RRM1,2 U9 bound structure as a template (2YH1) (Figure 3.2). Comparison of LS2 RRM1,2 with U2AF65 RRM1,2 bound to U9 RNA, showed that poly-U binding residues are more conserved on LS2RRM1, whereas on LS2RRM2, many poly-U binding residues are substituted by other amino acids. This suggests that RRM1 may be non-specific in nature whereas RRM2 might be specific for poly G RNA and thus, could avoid poly-U RNA. This analysis is consistent with U2AF65 RRM domains, whereas it is reported that RRM1 is more promiscuous for cytosine containing Py tracts than RRM2 (Jenkins, Agrawal et al. 2013)




Red : non-conserved poly U binding residues
Blue : conserved poly U binding residues

Homology model of RRM1,2 based on U2AF65 poly U bound structure
(generated using Swiss-modeller)

**Figure 3.2 Conservation of poly-U binding residues on LS2.** Sequence analysis shows that poly-U binding residues are mainly conserved (mapped in blue) in RRM1. On the other hand, many substitutions (mapped in red) for poly-U binding residues were observed on RRM2, indicating that RRM2 potentially governs specificity for poly G RNA.

### 3.1.1.2 Construct optimization of LS2 RNA-binding domains

Various LS2 constructs were subcloned and tested in search of improved sample purity as well as stability at RT (Figure 3.3).

Although the full-length protein (449 amino acids, 50.63 kDa) showed good expression in *E. coli*, the majority of the protein was observed in the inclusion body. In order to have protein size suitable for NMR experiments, the focus was shifted to minimal RNA-binding domains.

LS2RRM1,2 construct [(101-319), 24.45 kDa] was tested first, which yielded pure protein. But, upon removal of solubility tag, the protein was found to aggregate rapidly at RT, as seen by the increase in the turbidity and by the thick pellet formation upon centrifugation. In order to improve protein solubility Microscale thermophoresis (MST) Assay was employed. Combinations of different conditions including various buffer types and the pH (Tris-HCl, HEPES, pH 6-9), various types and concentrations of salts (100-500 mM NaCl or KCl) as well as additives such as (15%-50% glycerol or 50 mM arginine/glutamic acid) were screened. Protein seemed to show marginally better solubility at high concentration of glycerol as well as salt. But unfortunately, these conditions are not suitable for performing NMR experiments.

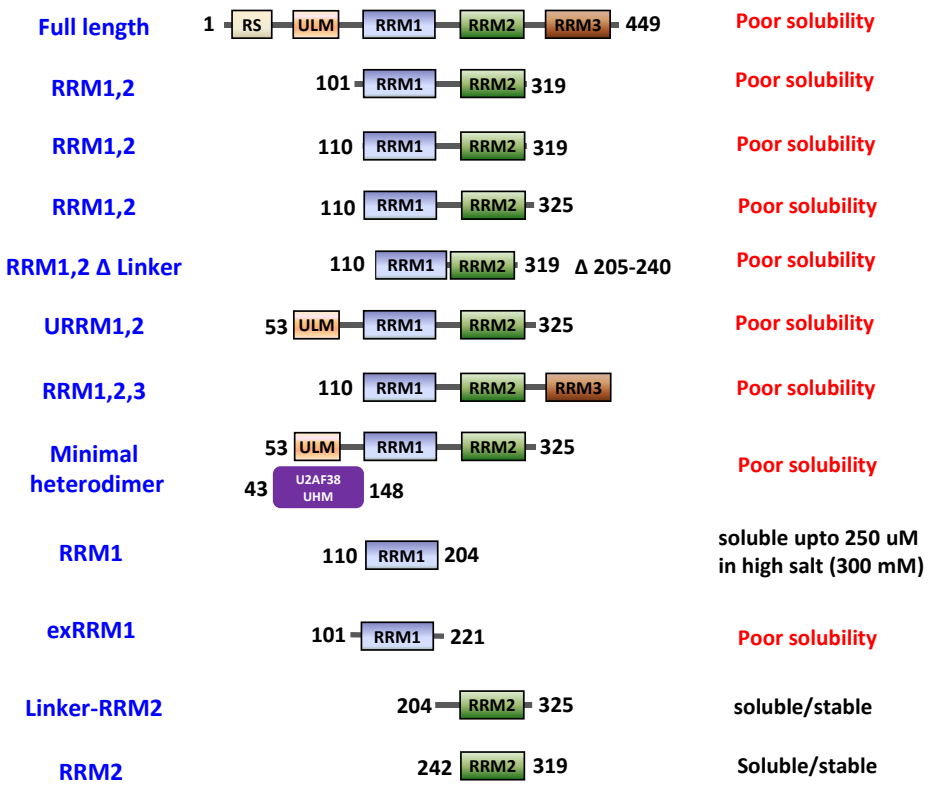



**Figure 3.3 Schematic representation of LS2 constructs.** With the exception of single domains and linker-RRM2, all the LS2 constructs showed poor solubility, characterized by the aggregation of protein samples at RT.

Nevertheless, $^{1}H$-$^{15}N$ HSQC spectrum was recorded on 100 µM protein sample. Dispersed proton amide peaks along the wide ppm range showed that protein is folded (Figure 3.4). But the low signal to noise of NMR signals, caused by protein precipitation, indicated that the sample stability should be improved for further data measurement.

To check whether deletion of hydrophobic residues located at the N-terminal or an inclusion of extra residues at C-terminus has any impact on protein stability constructs such as RRM1,2 (110-319) and RRM1,2 (110-325) were subcloned and protein stability was checked. Both constructs also showed aggregation-prone behavior. $^{1}H$-$^{15}N$ HSQC spectra were recorded on these constructs (Figure 3.4). Additionally, for crystallization trials, deleted linker construct (110-319 without 205-240) was subcloned, but also showed limited solubility as like previous LS2 RRM1,2 constructs.

In addition, longer constructs including N-terminal ULM domain or C-terminal RRM3 domain were also tried but also did not improve the protein solubility. To check if the interaction of dU2AF38 is necessary for the stability of the protein, the ULM-RRM1,2 construct was mixed with UHM domain of dU2AF38. This minimal heterodimer also showed high aggregation tendency, indicating that RRM1,2 is inherently aggregation prone.

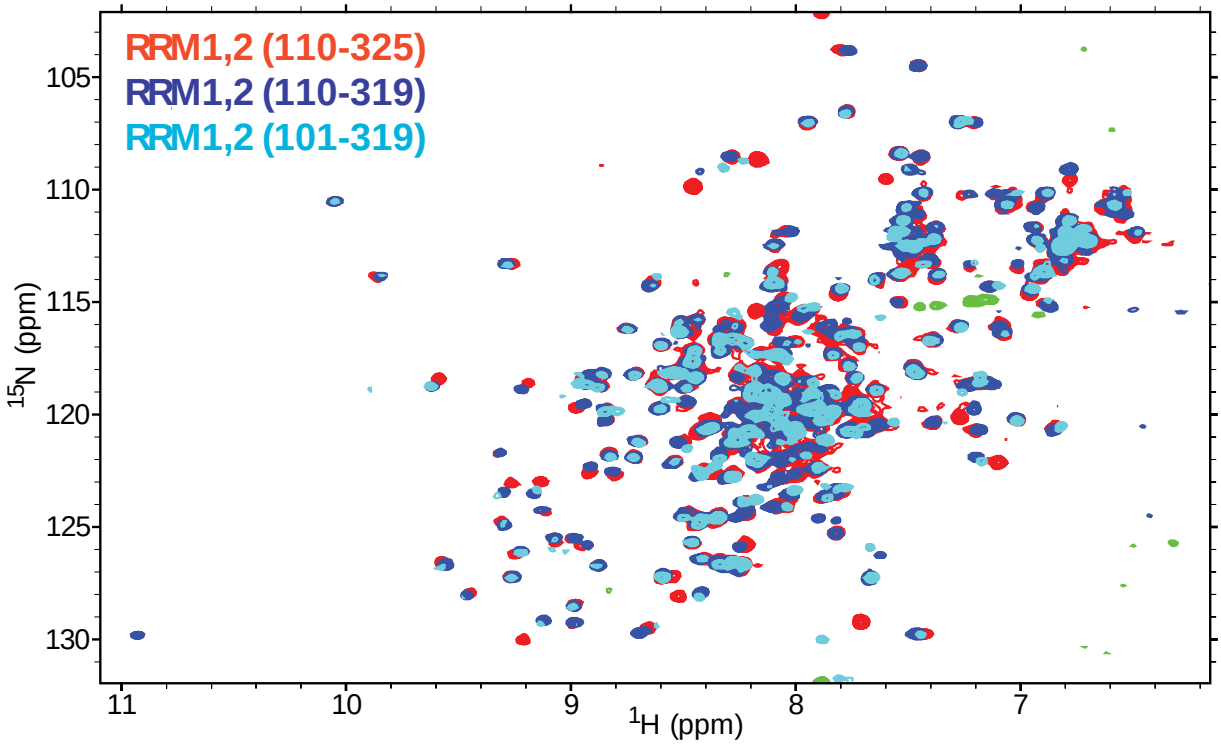



**Figure 3.4 Overlay of $^{1}H$-$^{15}N$ HSQC spectra of RRM1,2 with different domain boundaries.** Dispersed proton amide signals show that protein is well folded in all three constructs as well as have almost similar chemical shifts. Some chemical shifts changes were observed for the resonances of terminal residues and can be attributed to change in the domain boundaries.

In contrast, individual RRM domain constructs were found to be suitable for NMR data measurements. RRM2 was found soluble at higher concentration (>2.5 mM) even in the presence of low salt, whereas RRM1 showed aggregation-prone behavior, which could be overcome by using low protein concentration (≤250 µM) and high salt (≥300 mM). Subsequently, experiments for backbone, side chain resonance assignments as well as NOESY were recorded on single RRM constructs.

Meanwhile, linker-RRM2 and exRRM1 constructs were subcloned and tested to investigate the role of the linker. Linker-RRM2 construct also showed high solubility similar to the individual RRM2 construct. Hence, data for backbone, side chain assignments and NOESY was collected. On the other hand, exRRM1 showed aggregation-prone behavior as observed for RRM1,2 constructs.

### 3.1.1.3 Aggregation-prone behavior of LS2 RRM1,2

The aggregation of RRM1,2 found to be concentration, temperature as well as time dependent. At lower temperature and lower protein concentration, the UV-visible spectrum of LS2 RRM1,2 showed a peak at 280 nm, which is typically observed for most of the soluble proteins (containing aromatic residues). On the other hand, at higher protein concentration, UV-visible spectrum showed hyperbolic nature, which indicates the presence of soluble aggregates causing light scattering (Vincentelli, Canaan et al. 2004) (Figure 3.5). Thus, it shows that at higher concentration and lower temperature, protein has a tendency to exhibit soluble aggregates, which precipitates rapidly upon an increase in temperature.

After learning that protein shows better solubility at lower concentration and in presence of high NaCl concentration, temperature scan was performed to find out ideal temperature for NMR data measurements. Hence, series of $^{1}H$-$^{15}N$ HSQCs were recorded on RRM1,2 at temperatures ranging from 278 K to 298 K in the presence of 300 mM NaCl (Figure 3.6). Protein did not show visible precipitation up to 288K but started precipitating at 298 K. Better signal to noise was also observed at 288 K as compared to lower temperature. Hence, 288 K was chosen to record further experiments for backbone assignments along with high salt (300 mM NaCl) to keep protein stable enough to acquire NMR data. Backbone resonance assignments were challenging because of low spectral quality caused by protein aggregation. Nevertheless, backbone resonance assignments from single domains as well as linker-RRM2 were used to complete and confirm RRM1,2 backbone resonance assignments.

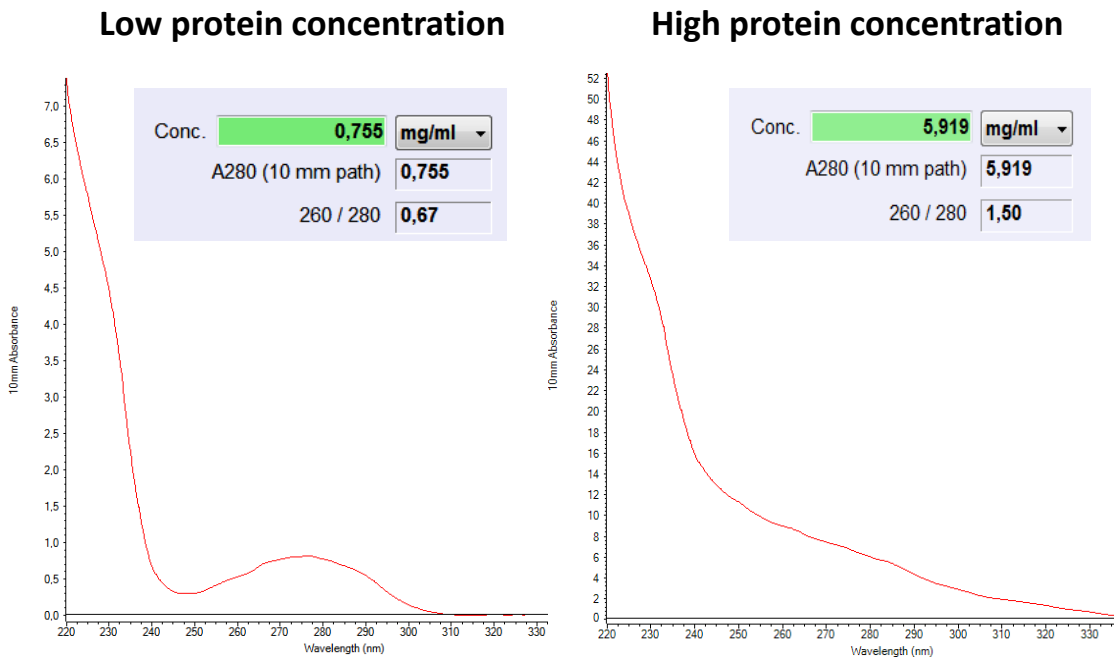



**Figure 3.5 Concentration-dependent aggregation of LS2 RRM1,2.** At higher concentration LS2 RRM1,2 forms soluble aggregates as seen by the hyperbolic curve in the UV-visible spectrum caused by light scattering by aggregated species.

The measured NMR data showed that LS2 RRM1,2 undergoes time-dependent changes at 288 K. The protein HSQC recorded after a 3D NMR experiments (3 days long) showed loss of signal intensity for some of the amino acids as well as some chemical shift changes in comparison to the HSQC recorded in the beginning of the experiment (Figure 3.7). These changes could be attributed to the high amount of soluble aggregates in the sample.

The ratio of signal intensities of each peak ($I_{after}/I_{before}$) was used to identify the residues undergoing maximum changes because of aggregation. This analysis shows that as a result fo aggregation, residues located on RRM1 and linker has maximum intensity loss. This indicates that RRM1 and linker are involved in aggregation (Figure 3.7).

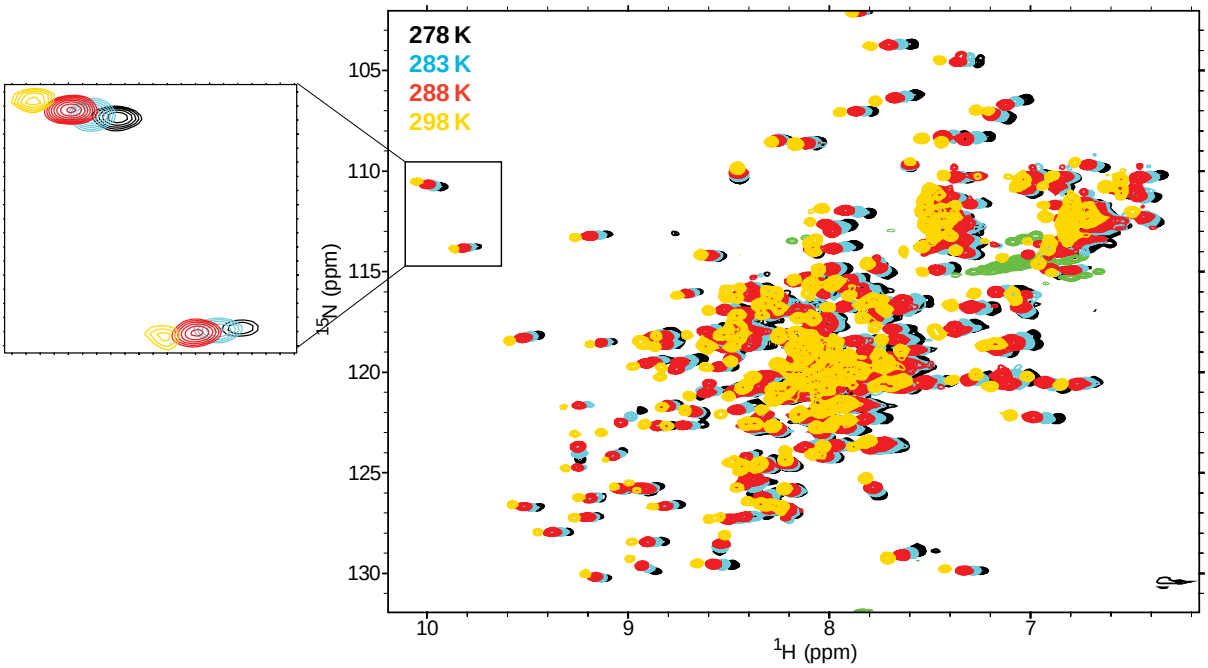



**Figure 3.6 Temperature studies of LS2 RRM1,2.** $^{1}$H-$^{15}$N HSQC spectra recorded at different temperatures show that at 288 K, RRM1,2 shows better signal to noise as compared to lower temperatures whereas have better solubility than at 298 K. The experiment was recorded on 150 µM protein at 900 MHz in presence of 20 mM potassium phosphate buffer pH 6.5, 300 mM NaCl and 5 mM DTT.

Aggrescan3d webserver (http://biocomp.chem.uw.edu.pl/A3D/)(Zambrano, Jamroz et al. 2015) was used to gain more insights about aggregation-prone behavior (Figure 3.8). It takes into account the 3D structure of the protein to predict aggregation hotspots present on the protein surface. LS2 RRM1,2 homology model was used as an input in this web server to identify the location of insoluble residues. The result shows that majority of insoluble residues are located on RRM1 as well as a linker. These residues cluster around one region, which may serve as a nucleation point during aggregation. On the other hand, no such aggregation-prone patch was observed on RRM2.

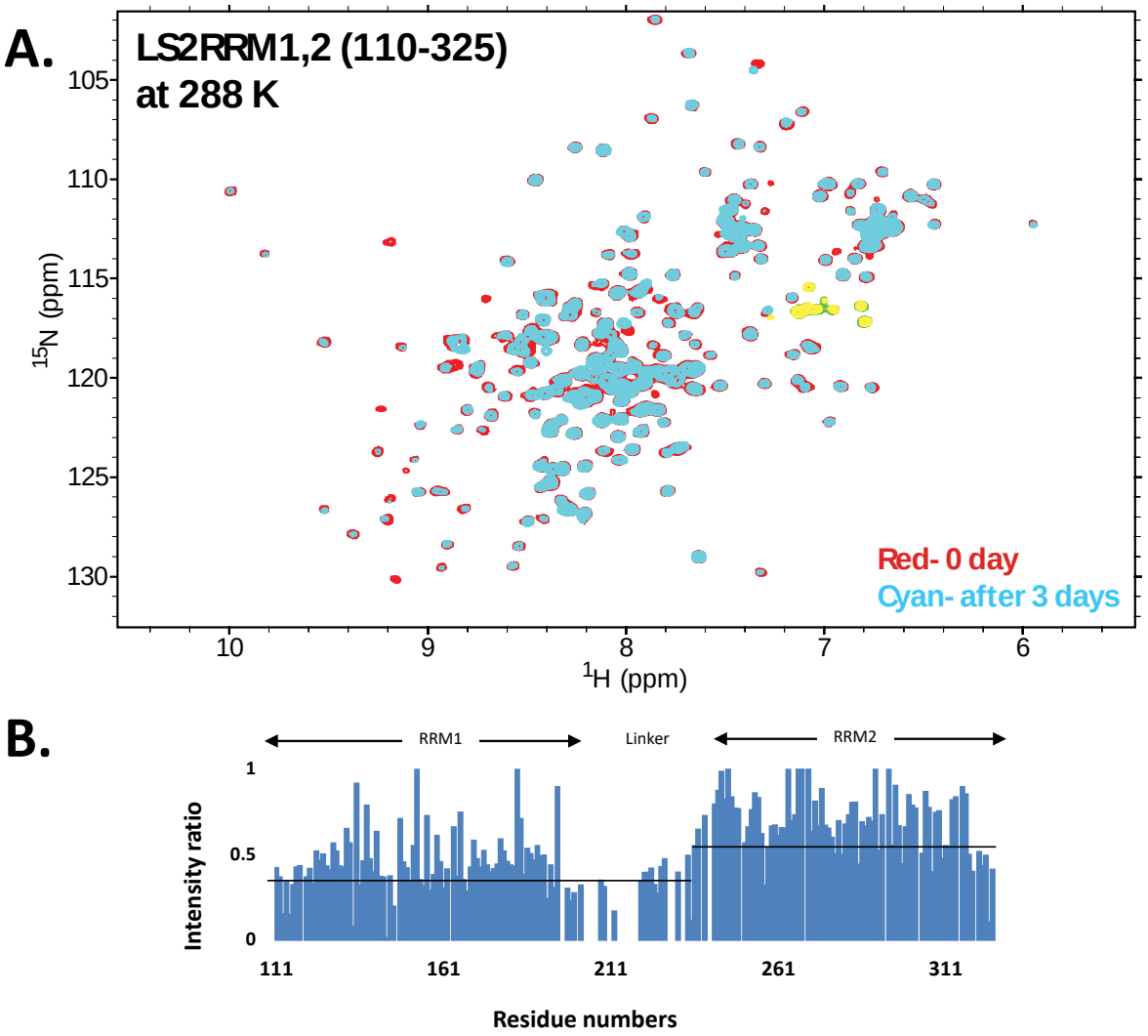



**Figure 3.7 LS2 shows time-dependent aggregation. A.** Comparison of $^{1}H$-$^{15}N$ HSQC of LS2 before and after 3 days shows changes in the intensity of certain residues. **B.** Residues located in RRM1 and linker, undergoes major intensity loss, indicating that these regions are more aggregation prone.

These findings are consistent with the aggregation-prone behavior of LS2 constructs where RRM2 is highly soluble, whereas RRM1 shows insolubility at higher concentration and/or in the presence of lower NaCl concentration. The combination of RRM1 followed by linker results in the highly aggregation-prone behavior of the protein, as seen for all RRM1,2 constructs as well as an exRRM1 construct. Surprisingly, higher solubility was achieved for linker-RRM2, despite containing aggregation-prone linker region.

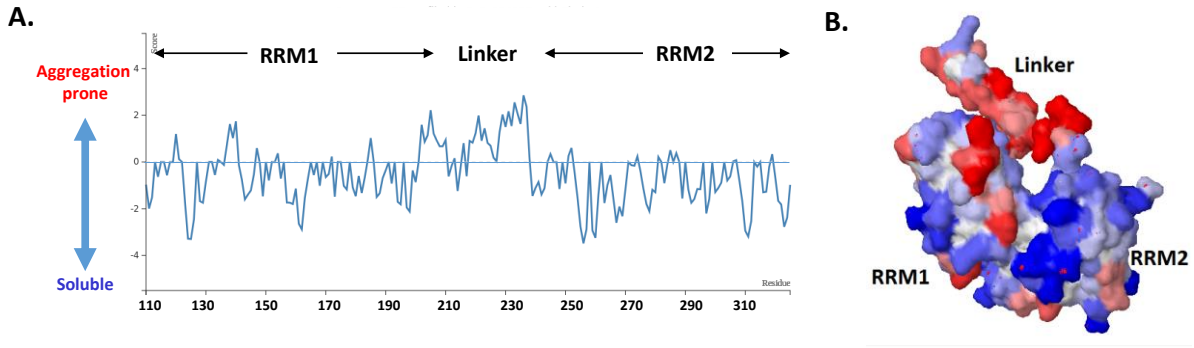



**Figure 3.8 Aggrescan3d prediction. A.** Aggrescan3d webserver prediction shows that LS2 RRM1 and linker has patches of aggregation-prone residues. **B.** The output structural model of aggrescan3d showing nucleation site for aggregation, in which aggregation-prone residues are shown in red, whereas soluble residues are colored blue.

### 3.1.1.4 NMR analysis and structures of LS2 individual RRM domains

$^{13}C_{\alpha}$-$^{13}C_{\beta}$ chemical shift analysis showed that LS2 RRM domains adopt canonical fold with βαββαβ topology. Backbone dynamics was assessed by measuring T1, T1rho as well as {$^{1}$H}-$^{15}$N heteronuclear NOE experiments. The average Correlation times ($\tau_c$) were calculated from T1 and T2 values and were found to be 8.69 ns and 7.98 ns for RRM1 and RRM2 respectively. These rotational correlation time values are in agreement with the size of the individual domains and indicate that both RRM domains are in monomeric form as well as they may tumble together in solution. {$^{1}$H}-$^{15}$N heteronuclear NOE data was also consistent with the T1 and T1rho, showing NOE value of around 0.75 for structured regions, whereas the lower value for flexible loops present in between (Figure 3.9).

$^{13}C_{\alpha}$-$^{13}C_{\beta}$ chemical shift derived secondary structure did not show distinct induced secondary structure for the additional residues present in between α1 helix and a β2 strand of LS2 RRM1. Backbone dynamics data such as {$^{1}$H}-$^{15}$N heteronuclear NOE, T1 and T1rho also shows that this region is highly flexible. Thus, NMR data indicate that this loop does not adopt any secondary structure as well as it does not interact with any structured region as a part of the RRM1 construct.

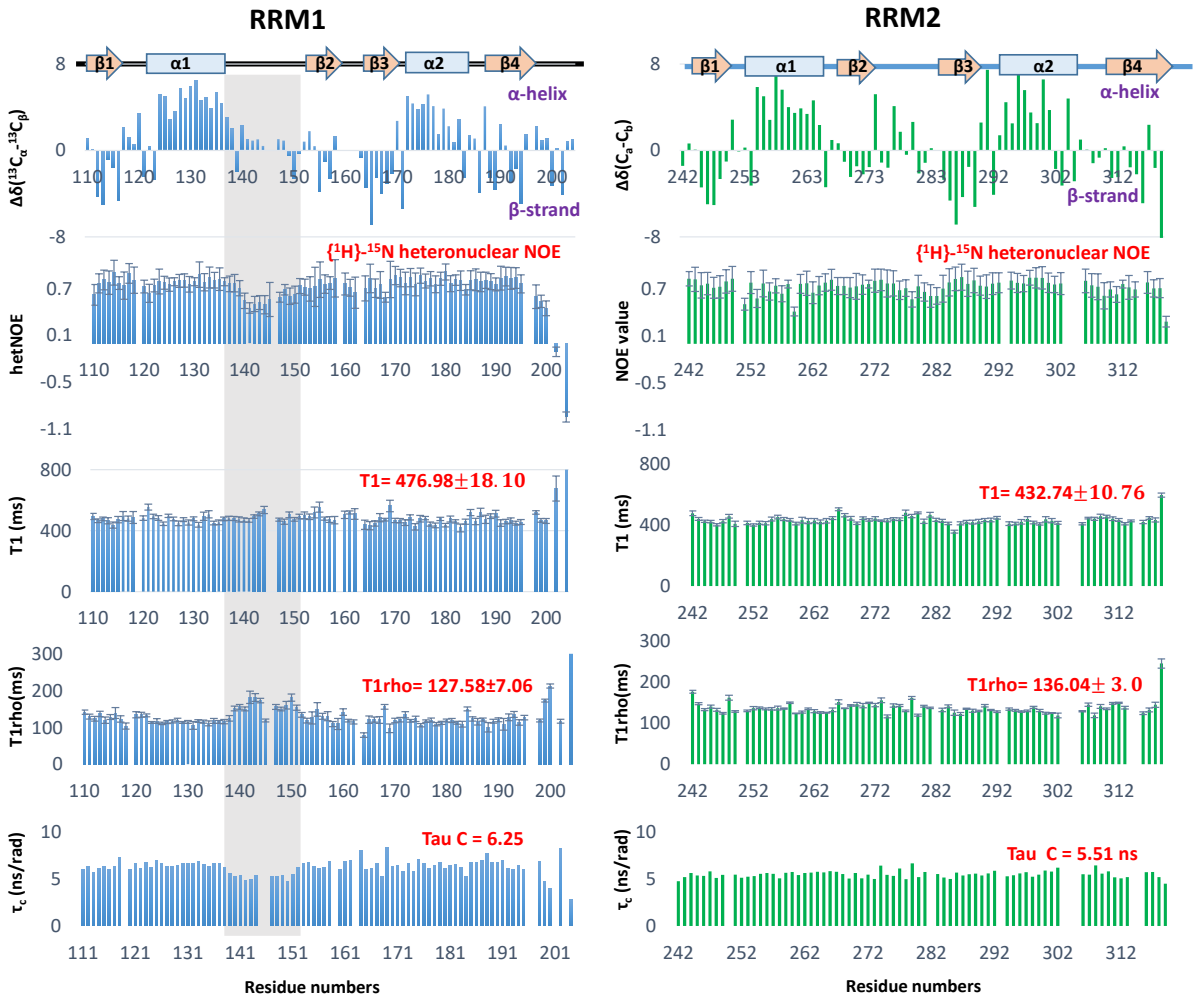



**Figure 3.9 Chemical shift derived secondary structure and backbone dynamics of LS2 RRM domains.** Chemical shift derived secondary structure prediction shows that both RRM domains of LS2 have βαββαβ topology (panel 1). Backbone dynamics data measured using {$^{1}$H}-$^{15}$N heteronuclear NOE, T1, and T1rho relaxation is consistent with the secondary structures of both RRM domains (panel 2,3,4 respectively). Correlation time ($\tau_c$) shows that both RRM domains exist in monomeric form (panel 5). The flexible loop formed by the additional residues of LS2RRM1 is shown in gray. Secondary structure elements are indicated on top for each RRM.

NOESY experiments were recorded to obtain interatomic distance information, aided by backbone and side chain resonance assignments of each domain. Talos+ was used to obtain dihedral angle restraints. NOESY spectra recorded on RRM1 were of poor quality which could be attributed to the aggregation-prone nature of protein sample. As a result, classical structure calculation approach could not be used. Hence, CS-Rosetta structure prediction protocol was employed (Shen, Lange et al. 2008), in which $^{13}C_\alpha$, $^{13}C_\beta$, $^{15}N$, $^{1}H_\alpha$, $^{1}HN$ resonance assignments were used to select fragments from a fragment library of known structures, comprising of fragments of which chemical shifts as well as torsion angles, are known. These selected fragments then used as a building block for generating RRM1 model (Shen, Lange et al. 2008, Shen, Vernon et al. 2009). In addition to this, few evident long-range intermolecular NOEs between β-strands of RRM1 were also taken into account for a model generation (Figure 3.10).

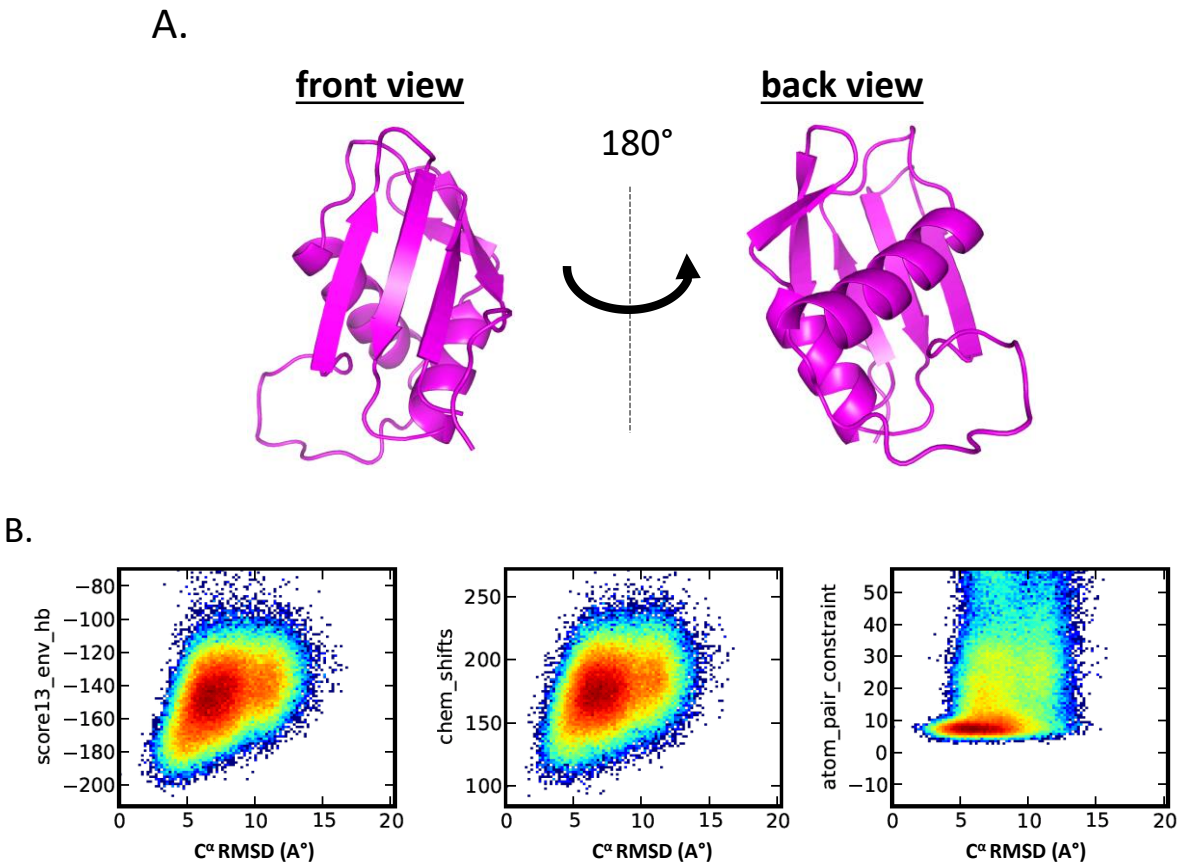



**Figure 3.10 CS-Rosetta model for LS2 RRM1. A.** Ribbon representation of RRM1 CS-Rosetta structure showing four β-strands of RRM1 are packed against two α-helices **B.** Plots of ROSETTA all atom energy score, chemical shift score and NOESY restraints versus $C^{\alpha}$ RMSD relative to the model with the best ROSETTA score, respectively.

On the other hand, the classical NOE-based approach was used to determine the solution structure of RRM2. Distance information was obtained from the NOESY spectra whereas dihedral angle restraints were obtained from chemical shifts and sequence-based prediction generated using Talos+. These restraints were incorporated by CYANA structure calculation protocol. Taking this information into account, 20 lowest energy structures were generated (Figure 3.11).

A.

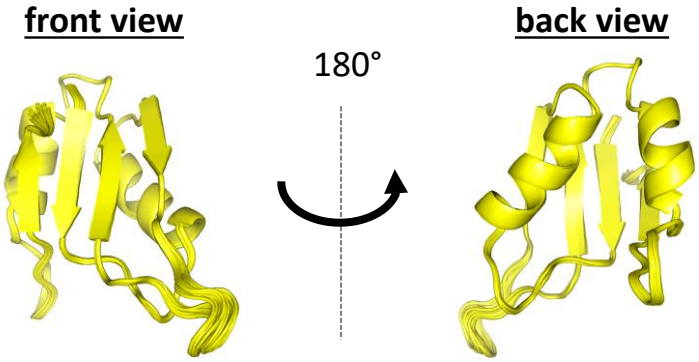



**Figure 3.11 Solution NMR structure of LS2 RRM2. A.** Ribbon representation of ensemble of 20 lowest energy structures calculated using solution NMR. Structure confirms βαββαβ topology of RRM2

**Table 3-1 Statistics of solution structure of LS2 RRM2**

| | |
|---|---|
| Completeness of $^{1}$H chemical shift assignments (%) | 96.2 |
| **NMR restraints** | |
| Distance restraints | 1686 |
| short-range, \|i-j\|<=1 | 766 |
| medium-range, 1<\|i-j\|<5 | 305 |
| long-range, \|i-j\|>=5 | 615 |
| | |
| **Ramachandran plot statistics (Å)** | |
| Residues in most allowed regions | 87.6 |
| Residues in additionally allowed regions | 12.4 |
| Residues in generously allowed regions | 0.1 |
| Residues in disallowed regions | 0 |
| | |
| **RMSD to mean structure statistics (Å)** | |
| backbone atoms | 0.21 ± 0.05 |
| heavy atoms | 0.59 ± 0.05 |
| | |
| **Cyana target function value** | |
| First cycle | 92.01 |
| Final | 7.10 |

*Statistics are reported for 20 lowest energy structure without explicit water refinement.

### 3.1.1.5 Interaction between LS2 Linker and RRM2

In order to check if the residues of individual RNA binding domains of LS2, i.e. RRM1 and RRM2 adopt the same conformation in bigger constructs, $^{1}$H-$^{15}$N HSQC of all the constructs were

compared. Comparison of $^{1}H$-$^{15}N$ HSQCs of various LS2 constructs (RRM1,2, RRM1,2 Δ linker, linker-RRM2, RRM1, RRM2) showed that though resonances from RRM1 superimpose quite nicely in all constructs, but the same is not true in the case of RRM2.

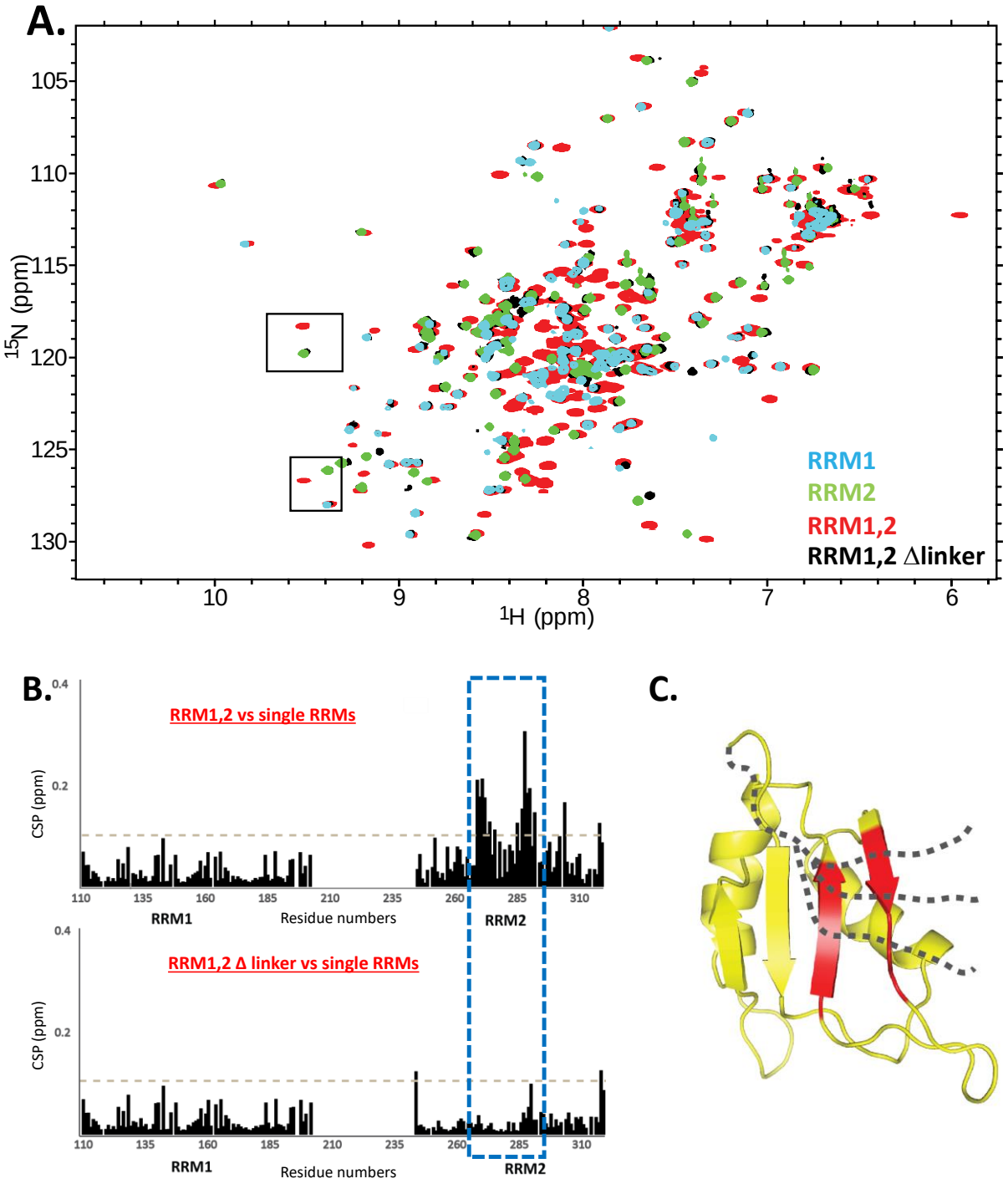



**Figure 3.12 LS2 RRM1,2 linker interacts with RRM2. A.** Overlay of $^{1}H$-$^{15}N$ HSQCs of various LS2 RRM constructs shows that some residues of RRM2 have different chemical shifts in the presence and the absence of linker, implying a possible interaction between linker and RRM2 residues. **B.** CSP analysis shows that linker interacting residues are located onto the β-strands of RRM2. **C.** Schematic representation of linker interacting residues of RRM2 (shown in red) mapped on the RRM2 structure.

In the presence of linker (i.e. in RRM1,2 and linker-RRM2 constructs), some residues of RRM2 show different chemical shifts in contrast to their chemical shifts in without linker constructs (RRM2; RRM1,2 Δ linker). This is indicative of an interaction between RRM2 and linker, which is disrupted in the absence of linker and thus, resulting in the different chemical shifts of RRM2 linker interacting residues. Chemical shift perturbation analysis shows that linker interacting residues are located in the β-strands of RRM2 (mainly β2 and β3), which also potential RNA binding sites.

### 3.1.1.6 Characterization of the LS2 RRM1,2 linker

LS2 RRM1,2 linker contains LS2-specific aromatic as well as positively charged residues which are located just after RRM1 and could be involved in RNA binding. Because of the aggregation-prone behavior of LS2 RRM1,2 construct, the study of these residues was not possible. Hence, in order to assign resonances of linker region as well as to study LS2-specific residues in more detail, the linker-RRM2 construct was subcloned. In contrast to RRM1,2 construct, linker-RRM2 was found stable at RT even in presence of low salt concentration. Backbone, as well as side chain resonance assignment experiments, were recorded on this construct and subsequently, assignments were performed. Backbone resonance assignments of the linker-RRM2 construct were used to complete the resonance assignments of the RRM1,2 construct.

Chemical shift-based secondary structure predicts these LS2-specific residues show helical propensity (Figure 3.13). More importantly, this helical propensity was also observed in the context of the RRM1,2 construct (Figure 3.19), indicating that this helical propensity is inherently present in LS2 linker not an artifact of protein truncation.

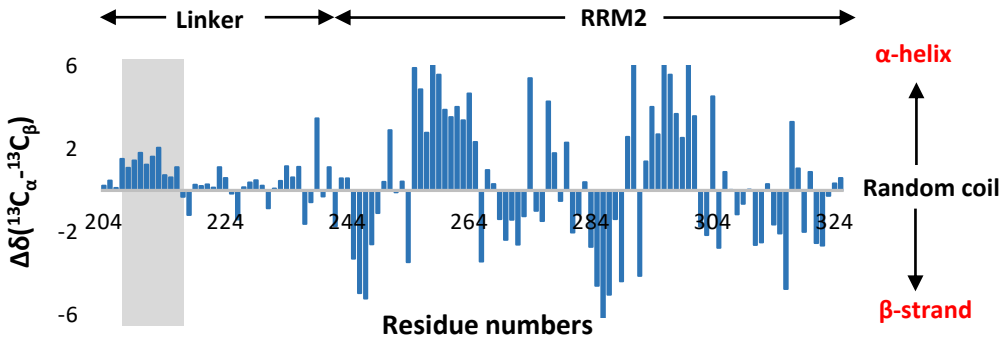



**Figure 3.13 Chemical-shift based secondary structure shows the presence of a region in a linker with helical propensity.** Chemical shift based secondary structure analysis shows that N-terminal region of the linker (which also comprises of potential RNA-binding residues) has a helical propensity. Grey box highlights residues of the linker, which show a helical tendency.

In order to gain more insights about linker, backbone dynamics experiments were recorded on the linker-RRM2 construct $\{^{1}H\}$-$^{15}N$ heteronuclear NOE data showed that the linker is not

completely flexible but rather have two patches of semi-rigidity. Interestingly, this includes an N-terminal region of the linker, which also shows helical propensity. This helix formation is transient, as backbone dynamics data values are lower than expected for a rigid secondary element. T1 and T1 rho measurement also confirmed that this stretch is not completely flexible. Thus, backbone dynamics data supported the transient helix formation of this unique stretch.

On the other hand, the second part of the linker exhibiting some degree of rigidity was observed in the C-terminal part of the linker, just preceding the RRM2 domain. This region is supposedly involved with the RRM2, and thus gains the rigidity.

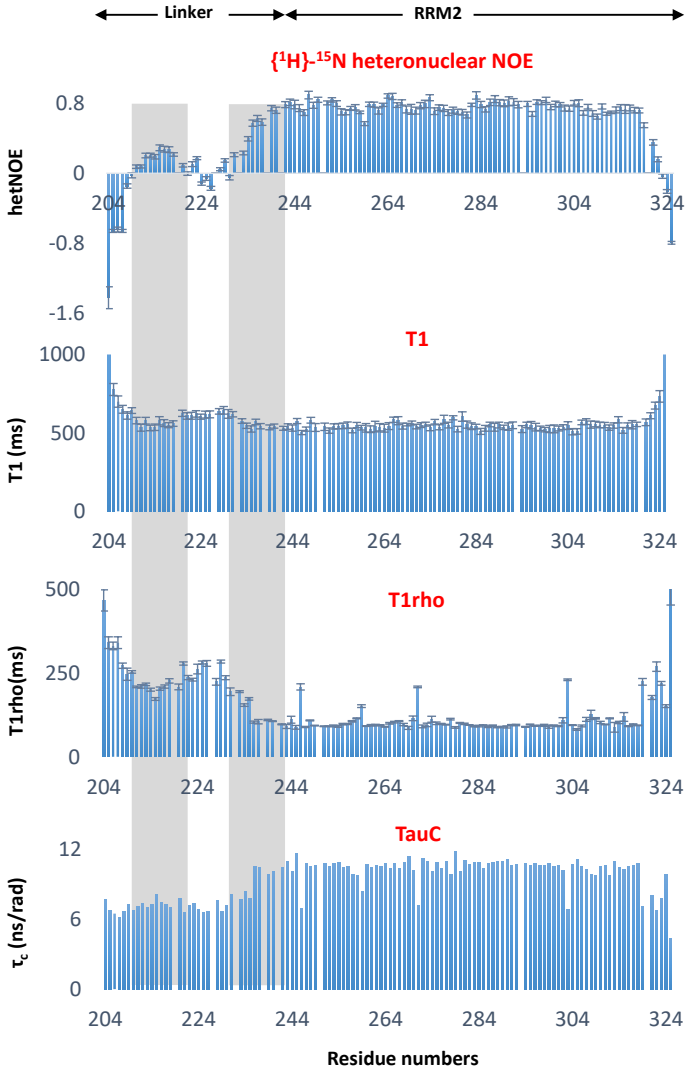



**Figure 3.14 NMR relaxation analysis of LS2 linker-RRM2.** Backbone dynamics data shows that linker has two semi-flexibility showing regions. The first region is located at the N-terminal of the region and rigidity gained by this region could be attributed to the helical propensity displayed by the residues. The other region which displays rigidity is located just preceding RRM2 which could be involved in interaction with RRM2 and thereby gains rigidity.

Multiple sequence alignment showed that this helix forming region is also well conserved among the LS2 homologs from other *Drosophila* species. Similarly, a region located at the C-terminal of the linker, which shows high rigidity in the backbone dynamics, was also found to have conservation among the Drosophila species (Figure 3.15). This region predominantly contains hydrophobic residues, which were later found out to interact with the β-strands of RRM2.

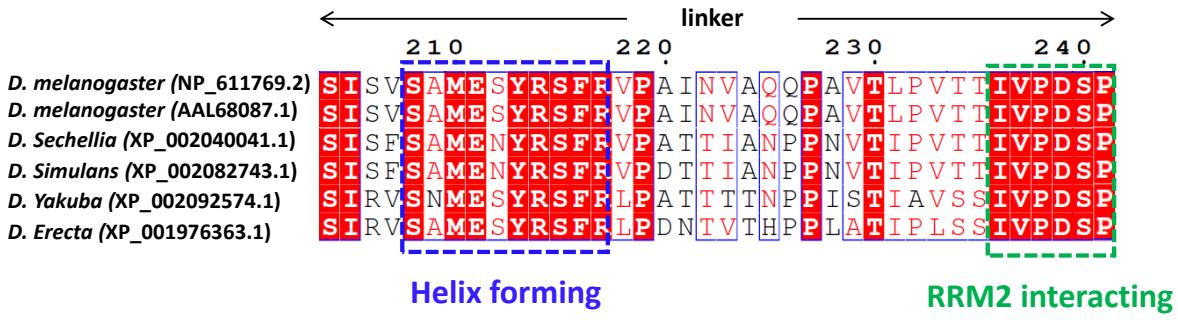



**Figure 3.15 Conservation of linker region among *Drosophila* species.** Multiple sequence alignment showing conservation of helix forming and RRM2 interacting regions among the LS2 proteins of *Drosophila* species.

Some RRM domains are known to have an additional third α-helix (helix C) in the C-terminal. This helix C has been shown to mediate hydrophobic interactions with RNP1 residues by positioning along the β-sheet surface. This hydrophobic interaction is crucial for the stability of the protein, and loss of this interaction has been shown to result in the aggregation-prone behavior of proteins (Clery, Blatter et al. 2008).

After discovering the presence of transient helix in the linker region, which is located just after RRM1, it could be speculated that it may be helix C of RRM1. To check whether inclusion of this helix improves solubility of RRM1, a new construct exRRM1 (101-221) was subcloned and tested.

Even after inclusion of this helix, the RRM1 was found to behave same as RRM1,2 by aggregating at RT even at low concentration in contrast to original RRM1 construct. Overlay of exRRM1 $^{1}H$-$^{15}N$ HSQC with RRM1,2 shows that chemical shifts of the amino acids comprising helix have different chemical shifts in both constructs. As chemical shifts of amino acids comprising helix in linker-RRM2 construct match quite reasonably with that of RRM1,2 chemical shifts, it could be speculated that this transient helix might be involved in interaction with RRM2.

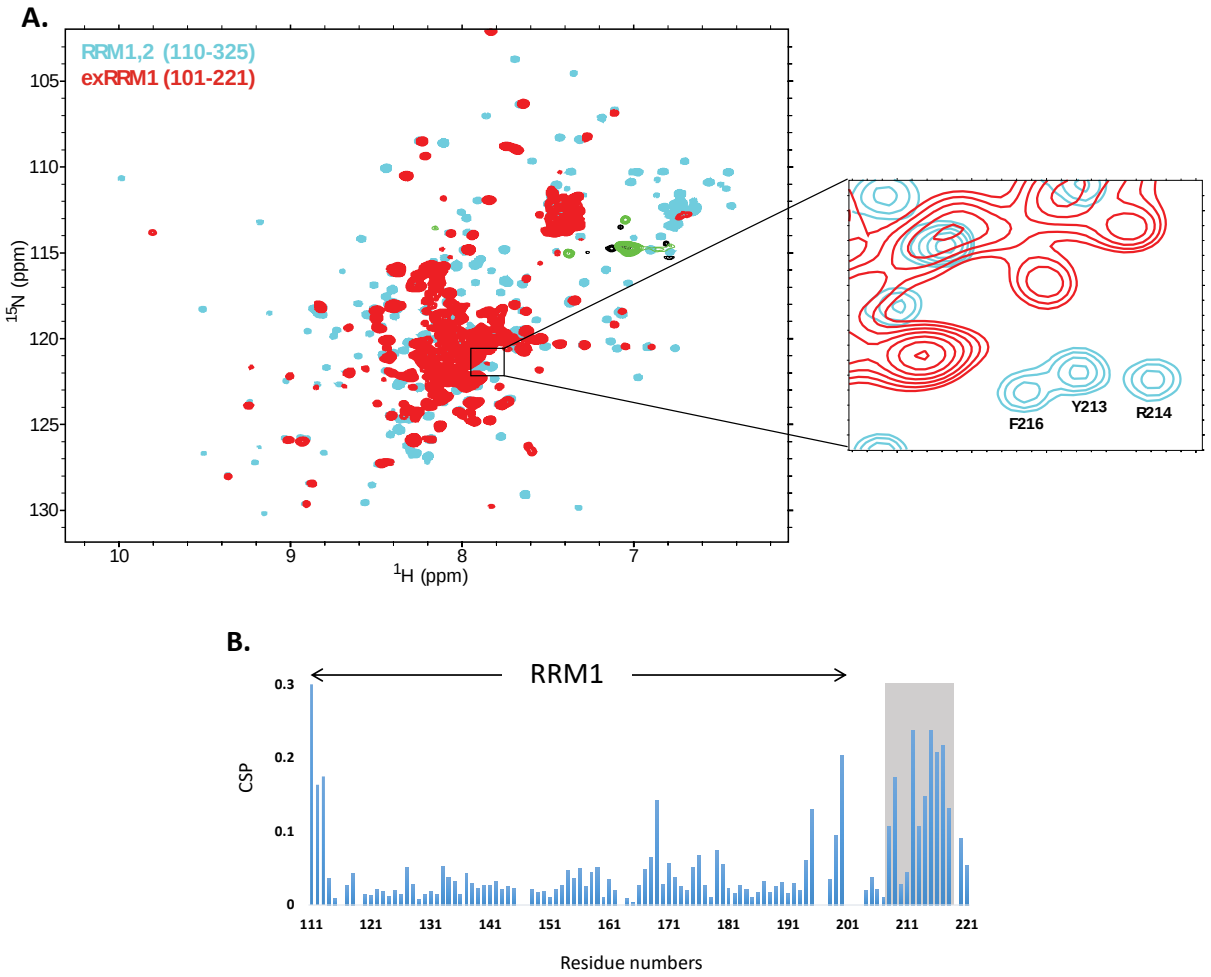



**Figure 3.16 Linker helix is not a part of RRM1. A. B.** $^{1}H$-$^{15}N$ HSQC overlay and CSP plot of RRM1,2 versus exRRM1 constructs showing that residues comprising of the transient helical region of linker have different chemical shifts in the presence and absence of RRM2.

One of the protein purification of linker-RRM2 yielded a truncated construct, lacking the N-terminal residues of the linker comprising helix. Comparison of the $^{1}H$-$^{15}N$ HSQC of this truncated construct with the intact linker-RRM2 construct showed chemical shift perturbations which are majorly located on the α1 helix of RRM2 (Figure 3.16). This indicates there is a potential interaction between linker helix as well as a helix of RRM2, which is ruled out in the absence of helix in truncated construct, thereby causing chemical shifts in the linker region as well as RRM2 (Figure 3.17).

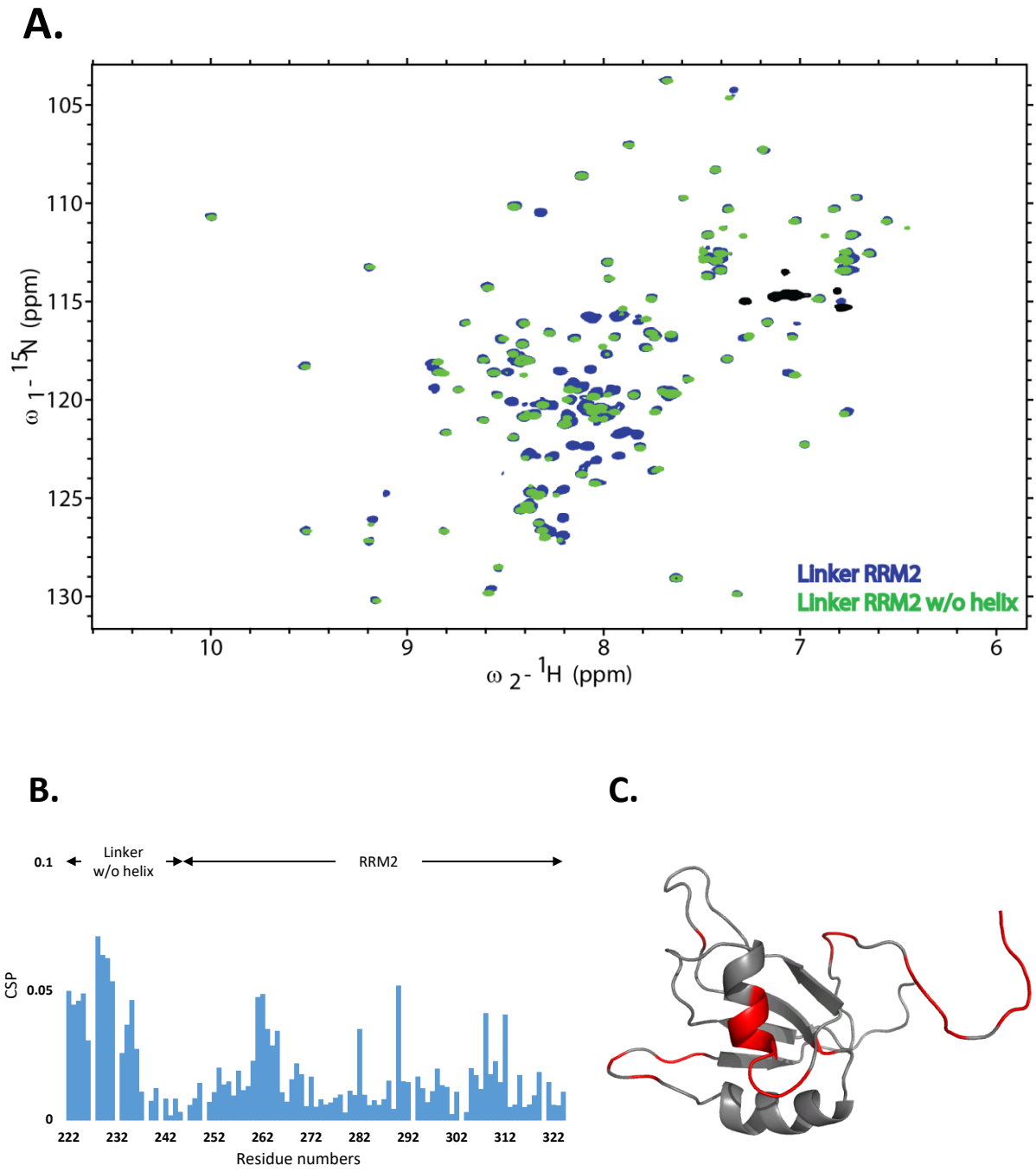



**Figure 3.17 Linker helix interacts with RRM2. A.** $^{1}H$-$^{15}N$ HSQC overlay of linker-RRM2 with and without linker helix constructs. **B. C.** upon helix deletion, major changes are observed residues located in rest of the flexible part of the linker as well as a α1 helix of RRM2, suggesting that linker helix interacts with α1 helix and nearby residues of RRM2.

### 3.1.1.7 Solution structure of linker-RRM2

In order to confirm linker and RRM2 interaction as well as helix formation by N-terminal of the linker solution NMR structure of linker-RRM2 was solved. Long range NOEs were observed

between hydrophobic residues of linker region with the β-strands of RRM2. The structure confirms the interaction between the C-terminal region of the linker, just preceding to the RRM2 linker β-strands of RRM2.

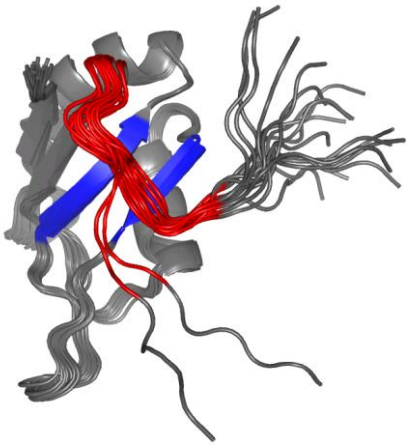


**Figure 3.18 solution structure of linker-RRM2.** Cartoon representation of ensemble of 20 lowest energy structures of linker-RRM2 showing linker interacting residues of RRM2 (colored blue) and RRM2 interacting linker residues (colored red).

**Table 3-3 Statistics of solution structure of LS2 linker-RRM2**

| Completeness of $^{1}$H chemical shift assignments (%) | **91.1** |
|---|---|
| **NMR restraints** | |
| Distance restraints | 1766 |
| short-range, \|i-j\|<=1 | 929 |
| medium-range, 1<\|i-j\|<5 | 295 |
| long-range, \|i-j\|>=5 | 542 |
| | |
| **Ramachandran plot statistics (Å)** | |
| Residues in most allowed regions | 70.7 |
| Residues in additionally allowed regions | 28.3 |
| Residues in generously allowed regions | 1.0 |
| Residues in disallowed regions | 0 |
| | |
| **RMSD to mean structure statistics (Å)** | |
| backbone atoms | 5.95 ± 2.41 |
| heavy atoms | 6.02 ± 2.30 |
| | |
| **Cyana target function value** | |
| First cycle | 676.98 |
| Final | 7.35 |

*Statistics are reported for 20 lowest energy structure without explicit water refinement.

In contrast, for most of the residues comprising linker helix $^{1}H$-$^{15}N$ NOESY, cross-peaks only for i±1 residues were observed in contrast to i±1, i±2, i±3, i±4 observed for a typical stable helix. For some residues, cross peak identification was challenging, as resonances overlapped heavily with diagonal peaks. Similarly, Talos+ did not identify the helical conformation from the chemical shifts and hence, structure calculation did not reveal the presence of helix in the linker.

In summary, the results show that the helical structure induced in the linker region is transient and it interacts with the α1 helix of RRM2.

### 3.1.1.8 Biophysical analysis of LS2 RRM1,2

Based on the $^{13}C_{\alpha}$-$^{13}C_{\beta}$ chemical shifts, the secondary structure of RRM1,2 was predicted. Chemical shift derived secondary structure analysis confirmed the βαββαβ topology of the RRM domains as well as the presence of transient α- helix in the linker region. (See section 3.1.1.6 Characterization of the LS2 RRM1,2 linker).

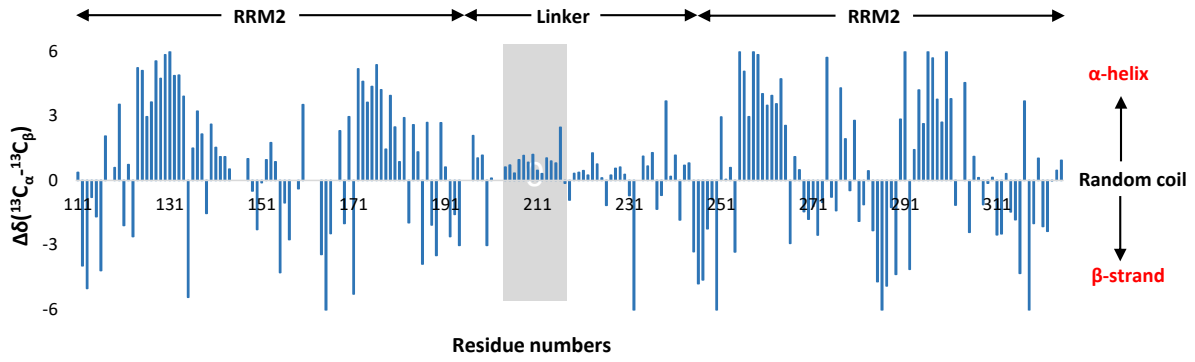



**Figure 3.19 RRM1,2 chemical shift derived secondary structure.** LS2 RRM1,2 chemical shift derived secondary structure also shows that LS2 RRM domains have canonical βαββαβ topology even in tandem form. It also confirms the presence of additional helix in the linker region identified using linker-RRM2 construct.

SAXS data was recorded to study the shapes of RRM1,2 and RRM1,2 Δ linker protein constructs in solution (Figure 3.20). The average dimensions (radius of gyration, $R_g$), as well as the maximum size ($D_{max}$) of the proteins, were calculated from the paired distance distribution function using GNOM. Data shows that the both RRM1,2 and RRM1,2 Δlinker protein constructs have dumbbell-like shape. The curve for RRM1,2 showed two peaks, indicating that the protein exists in the solution as an ensemble of conformation with compact and non-compact states. On the other hand, RRM1,2 Δ linker is more compact, as seen by the decrease in the $D_{max}$ value in comparison to RRM1,2. Together, this indicates that two RRM domains of LS2, do not tumble individually in the solution but rather either one domain interacts with the other and/or with the linker which is consistent with U2AF65 RRM domains.

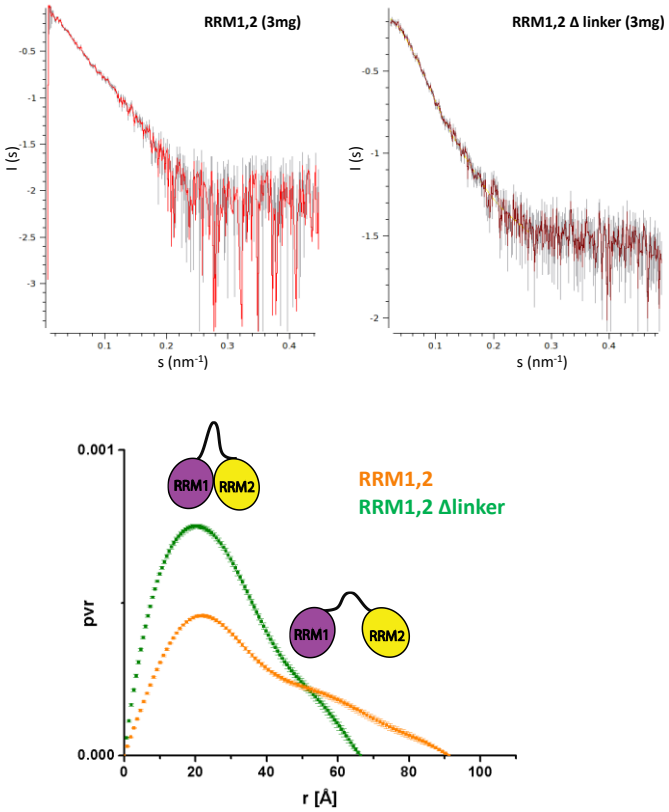



**Figure 3.20 SAXS analyses. A.** SAXS intensity plots of RRM1,2 and RRM1,2 Δ linker. **B.** Pair-distance distribution functions for RRM1,2 and RRM1,2 Δ linker. The data shows that both LS2 RRM1,2 and RRM1,2 Δ linker has dumbbell shapes in solution. RRM1,2 behaves as an ensemble of conformations with compact and non-compact states, seen by two lobes whereas RRM1,2 Δ linker is more compact.

### 3.1.1.9 Crystallization trials

Crystallization trials were performed on LS2 RRM1, RRM2, and RRM1,2 Δlinker construct. The linker region was thought to be flexible and thus expected to prevent the crystallization. Hence, protein constructs that lacked linker were of primary importance for crystallization trials.

A major focus was on LS2 RRM1,2 Δ linker construct, whose domain boundaries were delineated according to the crystallized U2AF65 RRM1,2 Δ linker construct. Because of aggregating nature of this construct, the lower temperature was maintained throughout the protein purification, concentration procedures as well as a crystal set up. 96 well protein crystallization conditions were tried by using screens such as Classics, Index, JCSG, Natrix, Nucleix as well as PACT. Initially, 22 mg/ml protein concentration was used which was varied later on. Crystal plates were incubated at 4 °C as well as RT, but in almost all of the conditions, precipitation was observed. Lowering the protein concentration also didn't promote crystallization. In three conditions crystals were observed, which upon diffraction at a synchrotron beamline, turned out to be salt crystals.

RRM1,2 Δ linker protein crystals were also set up in the presence of 21mer RNA (1:1), speculating that in the presence of RNA, the protein might show improved solubility and thus, could be more prone for the crystallization. But unfortunately, the presence of RNA also didn't improve the success of crystallization trials.

Similarly, crystallization trials were carried out on individual RRM domains with the above-mentioned screens. Several attempts, with varying protein concentration, did not yield protein crystals.

## 3.1.2 G-quadruplex formation by LS2 target RNA

### 3.1.2.1 21mer poly-G RNA forms G-quadruplex

LS2 is reported to bind guanosine-rich RNA targets. It was interesting to check whether LS2 target RNA can fold into G-quadruplex structures. Circular Dichroism is known to give characteristic spectrum in the presence of G-quadruplex species. Hence, to check G-quadruplex formation by LS2 target RNA (**21mer GGGGAGGAGGGGGGCGUAUGA**), CD spectroscopy was used (Figure 3.21). The CD spectrum of 21mer showed a positive peak at 265 nm and a negative peak at 240 nm at 25 °C in the presence of KCl or NaCl, which indicates the formation of parallel stranded G-quadruplex structure (Vorlickova, Kejnovska et al. 2012). In addition to this, in the presence of KCl, CD spectrum also showed absorption around 305 nm, which indicates the formation of hexad in the structure. Interestingly RNA also showed G-quadruplex-specific spectrum in the absence of any salt albeit with lower intensity, indicating that G-quadruplex is not very stable in the absence of salt and along with G-quadruplex RNA also exhibits either duplex or single-stranded form.

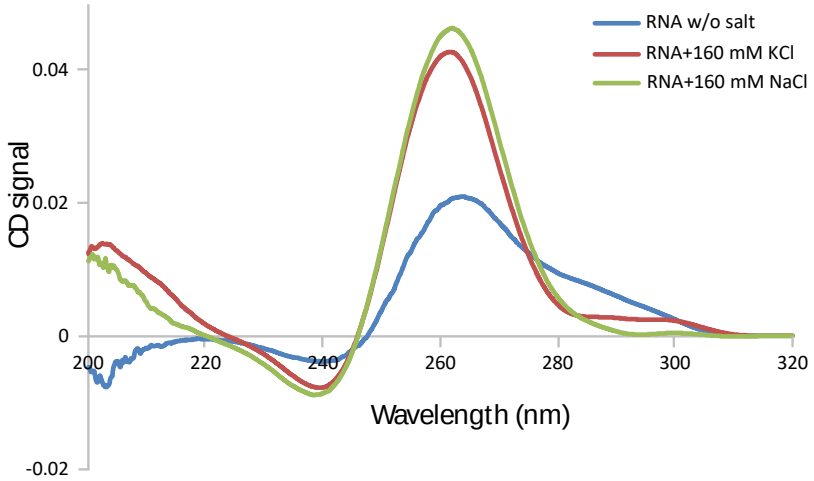



**Figure 3.21 G-quadruplex formation by CD spectroscopy.** The negative peak at 240 nm and a positive peak at 265 nm indicate that 21mer RNA forms parallel G-quadruplex, which is more stable in the presence of salt.

### 3.1.2.2 NMR shows that G-quadruplex has multiple conformations

To confirm G-quadruplex formation by LS2 target RNA sequence $^{1}$H 1D NMR spectroscopy was employed (Figure 3.22). In $H_2O$, 21mer RNA showed imino resonances specific for G-quadruplex (around 10-12 ppm as a result of wobble base pairing pattern) as well as for

duplex (12-14 ppm because of Watson and Crick base pairing), indicating the presence of duplex and quadruplex mixture.

On the other hand, in the presence of 150 mM KCl, the imino resonances characteristic of duplex structure disappeared and single broad envelope of imino protons between 10-11 ppm along with minor peak around 13 ppm appeared. It indicated that KCl changed the duplex-quadruplex equilibrium in favor of G-quadruplex. Single broad peak instead of isolated peaks of G-quadruplex imino protons represented the presence of multiple conformations instead of single homogeneous conformation.

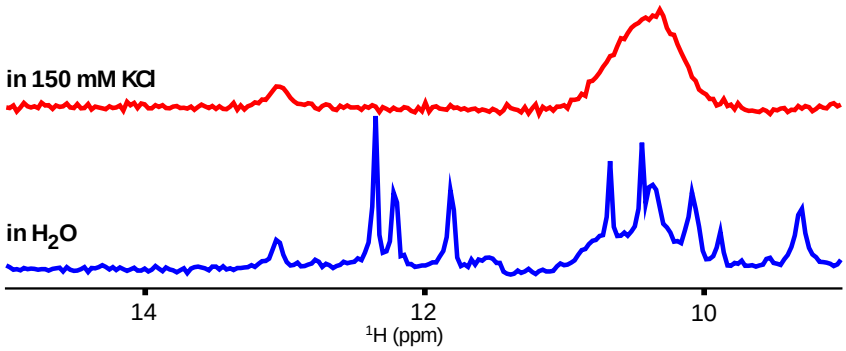



**Figure 3.22 NMR spectra for 21mer RNA shows multiple conformations.** In $H_2O$, 21mer RNA shows a mixture of the duplex and G-quadruplex species as seen by two sets of signals (12-14 ppm and 10-12 ppm). On the other hand, in presence of 150 mM KCl, majorly one broad envelope was observed (10-12 ppm), indicating the formation of G-quadruplex with multiple conformations. The experiment was recorded on 100 µM RNA at 278 K and 500 MHz.

These multiple conformations could be attributed to the presence of many guanine nucleotides (14 out of 21) in 21mer with each capable of involvement in G-quadruplex formation. As predicted by QGRS mapper (Kikin, D'Antonio et al. 2006), a server designed to predict the sequence forming G-quadruplex, there are numerous possibilities of G-quadruplex formation depending upon which guanine residues are involved, with all of them are having quite similar scores. (Figure 3.23). Given the fact that QGRS mapper only takes into account intramolecular G-quadruplex formation, intriguing possibilities of G-quadruplex structures with intermolecular topologies increases the complexity of possible conformations.

| Position | Length | QGRS | G-Score |
|---|---|---|---|
| 1 | 10 | GGGGAGGAGG | 20 |
| 1 | 11 | GGGGAGGAGGG | 19 |
| 1 | 12 | GGGGAGGAGGGG | 18 |
| 1 | 12 | GGGGAGGAGGGG | 18 |
| 1 | 13 | GGGGAGGAGGGGG | 17 |
| 1 | 13 | GGGGAGGAGGGGG | 17 |
| 1 | 13 | GGGGAGGAGGGGG | 19 |
| 1 | 13 | GGGGAGGAGGGGG | 18 |
| 1 | 14 | GGGGAGGAGGGGGG | 16 |
| 1 | 14 | GGGGAGGAGGGGGG | 17 |
| 1 | 14 | GGGGAGGAGGGGGG | 16 |
| 1 | 14 | GGGGAGGAGGGGGG | 19 |
| 1 | 14 | GGGGAGGAGGGGGG | 19 |
| 1 | 14 | GGGGAGGAGGGGGG | 18 |
| 2 | 11 | GGGAGGAGGGG | 19 |
| 2 | 12 | GGGAGGAGGGGG | 20 |
| 2 | 12 | GGGAGGAGGGGG | 19 |
| 2 | 13 | GGGAGGAGGGGGG | 20 |
| 2 | 13 | GGGAGGAGGGGGG | 20 |
| 2 | 13 | GGGAGGAGGGGGG | 18 |
| 3 | 10 | GGAGGAGGGG | 20 |
| 3 | 11 | **GGAGGAGGGGG** | 21 |
| 3 | 11 | GGAGGAGGGGG | 19 |
| 3 | 12 | GGAGGAGGGGGG | 20 |
| 3 | 12 | GGAGGAGGGGGG | 20 |
| 3 | 12 | GGAGGAGGGGGG | 18 |

**Figure 3.23 QGRS Mapper prediction.** QGRS web server, which predicts the possible combinations of guanine residues to adopt G-quadruplex conformation, predicts that there are many possibilities of G-quadruplex conformations with the almost similar score.

### 3.1.2.3 Designing Shorter oligonucleotides on the basis of SELEX

In order to check if a change in RNA sequence improves the structural homogeneity of G-quadruplex, shorter oligonucleotides were designed according to SELEX data obtained for predicting LS2 target RNA(Taliaferro, Alvarez et al. 2011). Shorter oligonucleotides containing GGX motif as a basic unit, such as 14mer, 8mer and 5mer were designed (Figure 3.24).

21mer GGGGAGGAGGGGGGCGUAUGA

14mer GGGUGGUGGAGGGG

8mer GGUGGUGG

5mer GGUGG

**Figure 3.24 Overview of shorter poly-G oligonucleotides.** Shorter poly-G oligonucleotides that were designed according to SELEX identified 'GGX' motif.

These sequences (except for 5mer) were checked by NMR for their capacity to form G-quadruplex structure. But in presence of 50 mM KCl, both oligonucleotides also showed the presence of multiple conformations.

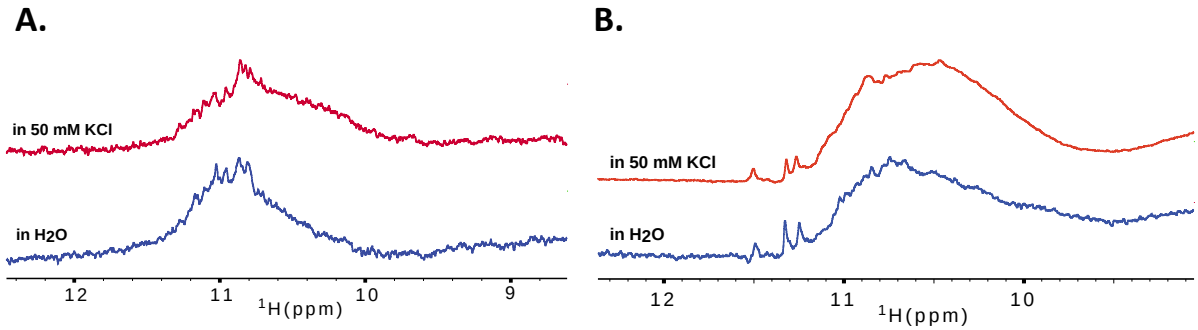



**Figure 3.25 Multiple conformations adopted by 14mer and 8mer. A.** 14mer exists in multiple conformations in the presence of $H_2O$, as well as 50 mM KCl as seen by $^1H$ NMR **B.** $^1H$ NMR showing that 8mer also adopts multiple conformations in the absence or even in the presence of 50 mM KCl.

### 3.1.2.4 Low KCl concentration induces uniform conformation for 21mer and 8mer

G-quadruplex conformations also depend on the type as well as the concentration of the monovalent cation used. To check whether G-quadruplex adopts a uniform conformation in the presence of NaCl, 1D NMR spectrum of 21mer was recorded in the presence of 10 mM NaCl at 298 K. But the spectrum showed the presence of duplex-quadruplex mixtures as observed in the absence of any salt (Figure 3.26).

On the other hand, upon titration with KCl at 278 K, 21mer showed several well-resolved resonances in the range of 10-12 ppm, indicating the presence of G-quadruplex with

homogeneous conformation (Figure 3.26). Upon increasing KCl concentration, broad peak around 11 ppm also showed up. This could be the result of either oligomerization of G-quadruplex species or involvement of remaining guanines in wobble base pairing in presence of increased availability of potassium ions.

$^{1}$H NMR spectrum of 21mer recorded in the presence of the 10 mM KCl and 25 mM NaCl showed the uniform conformation adopted by 21mer in the presence of KCl, indicating that G-quadruplex adopted in the presence of KCl is major conformation.

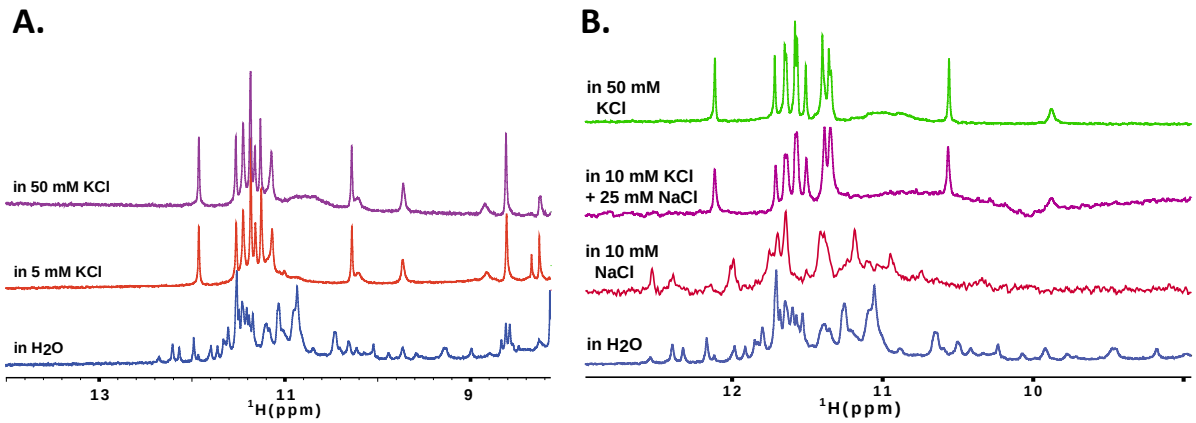



**Figure 3.26 21mer adopts uniform conformation in the presence of low KCl concentration. A.** Titration of 21mer showing that 21mer RNA adopts G-quadruplex conformation at lower KCl concentration **B.** In the presence of NaCl, the 21mer shows mixture of G-quadruplex-duplex species, whereas, in the presence of mixture of KCl and NaCl, it prefers uniform conformation along with some multiple conformations as observed in the presence of KCl. The experiments were recorded on 65 µM RNA at 800 MHz at 278 (A.) and 298 K (B.).

Further characterization shows that multiple conformational topologies adopted by 21mer RNA are not reversible to uniform conformation by a reduction in salt concentration. Extensive dialysis with $H_2O$ or buffer with low KCl concentration did not stability of multiple conformations. On the other hand, it was learned that heating at 95 °C and further slow cooling at RT leads to unfolding of multiple conformations and depending upon the final KCl concentration, it adopts majorly either uniform conformation or multiple conformations (Figure 3.27). Whereas, after snap cooling different imino signal pattern was observed, suggesting that G-quadruplex adopts totally different topology after snap cooling.

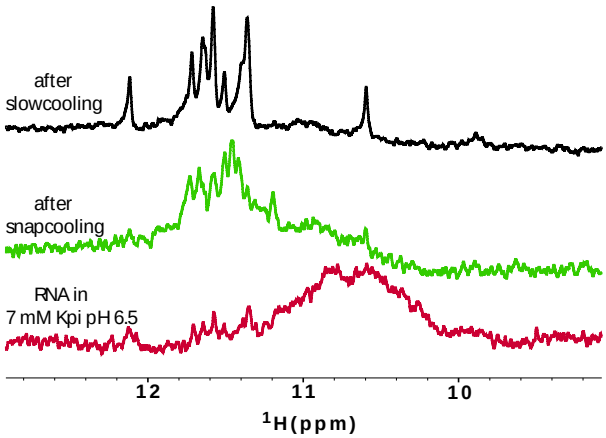



**Figure 3.27 Slow cooling induces conversion of multiple conformations to uniform conformation.** Comparison of $^{1}$H NMR spectra of the RNA after snap cooling and slow cooling shows that imino proton spectrum after slow cooling matches with the imino spectrum obtained during KCl titration. The experiment was recorded on 30 µM RNA at 278 K and 500 MHz.

8mer in low KCl concentration also showed well-resolved imino resonances in the range of 10-12 ppm, indicating that it also adopts G-quadruplex topology with a homogeneous population. But at the same time, signals corresponding to multiple conformations were also observed. On the other hand, 14mer failed to give well-resolved peaks even after titration with KCl or NaCl, indicating that it G-quadruplex structure formed by it, exists only in multiple conformations (Figure 3.28).

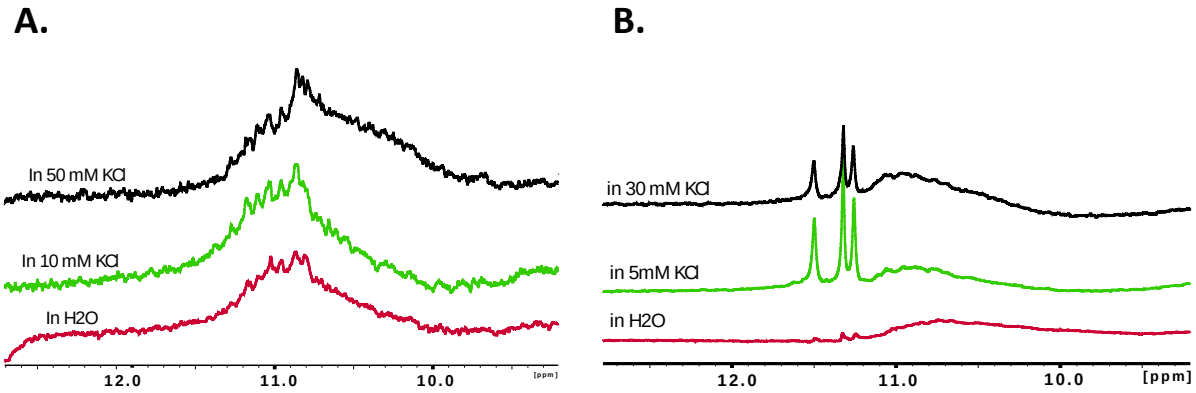
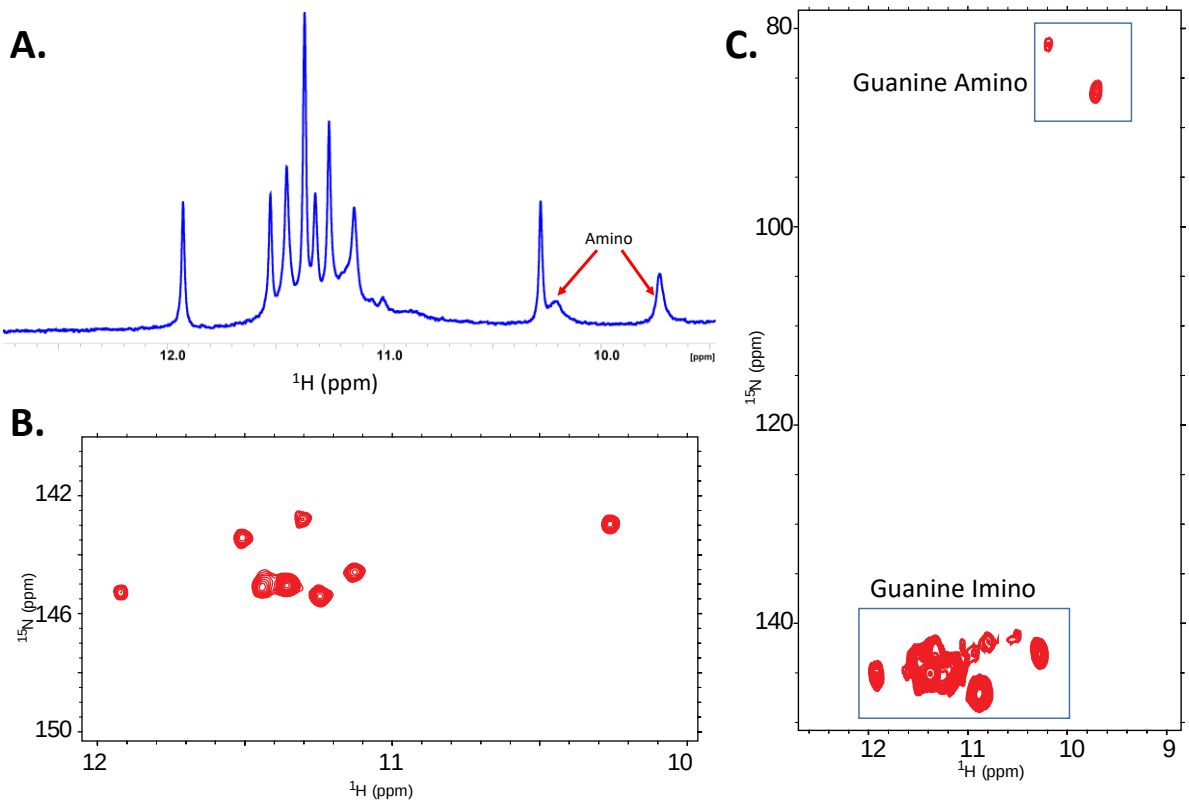



**Figure 3.28 KCl titration of 14mer and 8mer. A.** KCl titration of 14mer showing that it does not adopt a uniform conformation in the presence of KCl, rather it exists only in multiple conformations. **B.** Similar to 21mer, 8mer adopts a uniform conformation in the presence of low KCl concentration and upon an increase in the KCl concentration, multiple conformations also show around 11 ppm.

### 3.1.2.5 Biophysical characterization of 21mer RNA G-quadruplex

$^{1}$H NMR spectrum of 21mer RNA showed at least eight distinct resonances in the imino region. Natural abundance $^{1}$H-$^{15}$N HSQC showed that all these resonances belong to guanine imino protons, because of their characteristic nitrogen chemical shift. Given the fact that intensity of all these resonances is not same, some additional overlapping resonances were suspected. Hence, the adopted G-quadruplex topology can be speculated to be either intramolecular, two planar (eight iminos) or intermolecular, three planar (twelve iminos) (Figure 3.29).

Additionally, two slightly broader resonances were detected downfield to imino resonances, around 10 ppm (Figure 3.29). These two amino resonances were attributed to purine amino protons, because of their characteristic chemical shift in natural abundance $^{1}$H-$^{15}$N HSQC. Similar type of resonances were also observed in other G-quadruplex forming RNA sequences, which predominantly contain GGA motif, such as $(AGG)_2A$, $A(GG)_4A$ (Malgowska, Gudanis et al. 2014), $(UGGAGGU)_4$, $(GGA)_4$, (GGAGGUUUUGGAGG) as well as DNA sequences such as (GGAGGAG), (GCGGAGGAT) (Lipay and Mihailescu 2009). These amino resonances were shown to reflect the formation of A:(G:G:G:G):A hexads, with two hexad subunit stacked on each other (Lipay and Mihailescu 2009).



**D.**

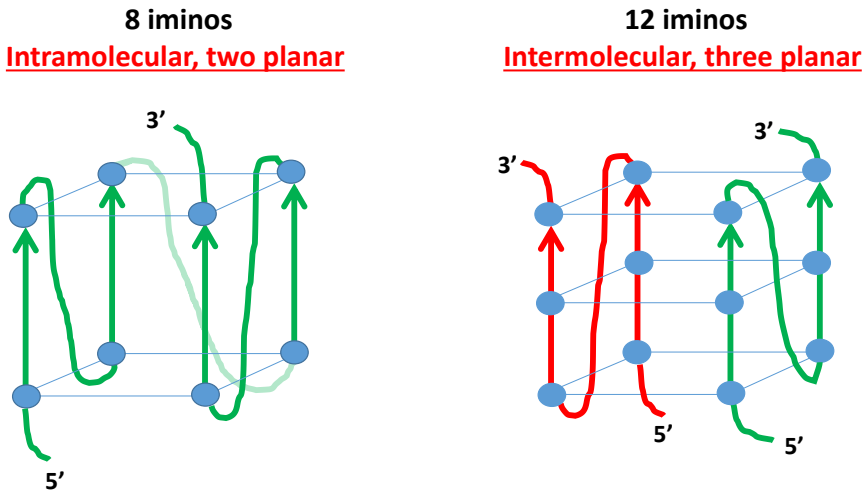



**Figure 3.29 Identification of imino and amino protons of 21mer G-quadruplex. A.** Uniform conformation of 21mer G-quadruple shows the presence of 10 protons at the exchangeable region in $^{1}H$ NMR spectra. **B.C**. $^{1}H$-$^{15}N$ HSQC shows that eight are guanine imino protons whereas two are purine amino protons which are observed in the downfield of the spectrum. **D.** Two possible models for 21mer G-quadruplex uniform topology predicted depending upon a number of imino signals.

$^{1}H$ 1D NMR spectrum revealed that upon increasing the temperature, some imino signals showed splitting (Figure 3.30). At 318 K temperature, ten peaks were observed which were well resolved as well as have almost similar intensity except two, which appeared to be a doublet. Thus, temperature studies show the presence of twelve iminos can be observed which suggests that the G-quadruplex topology might be three planar. Upon increasing the temperature above 318 K, imino signals started losing the intensity indicating that G-quadruplex structure starts melting above 318 K.

$D_2O$ exchange studies showed that six iminos were quite resistant to solvent exchange with $D_2O$. Out of them, four remained distinct even after two weeks of $D_2O$ exchange (Figure 3.30). In addition, two amino resonances were also found to be resistant to exchange with $D_2O$. This data hints at the three planar model of G-quadruplex with hexad formation, as the central plane sandwiched in between two other G-quartet, it is more resistant to exchange. Similarly, it is reported that guanine imino protons that are involved in hexad formation are exchanged more slowly because of neighboring adenosine nucleotides which protect them from the solvent. This may explain the partial resistance of other two imino protons.

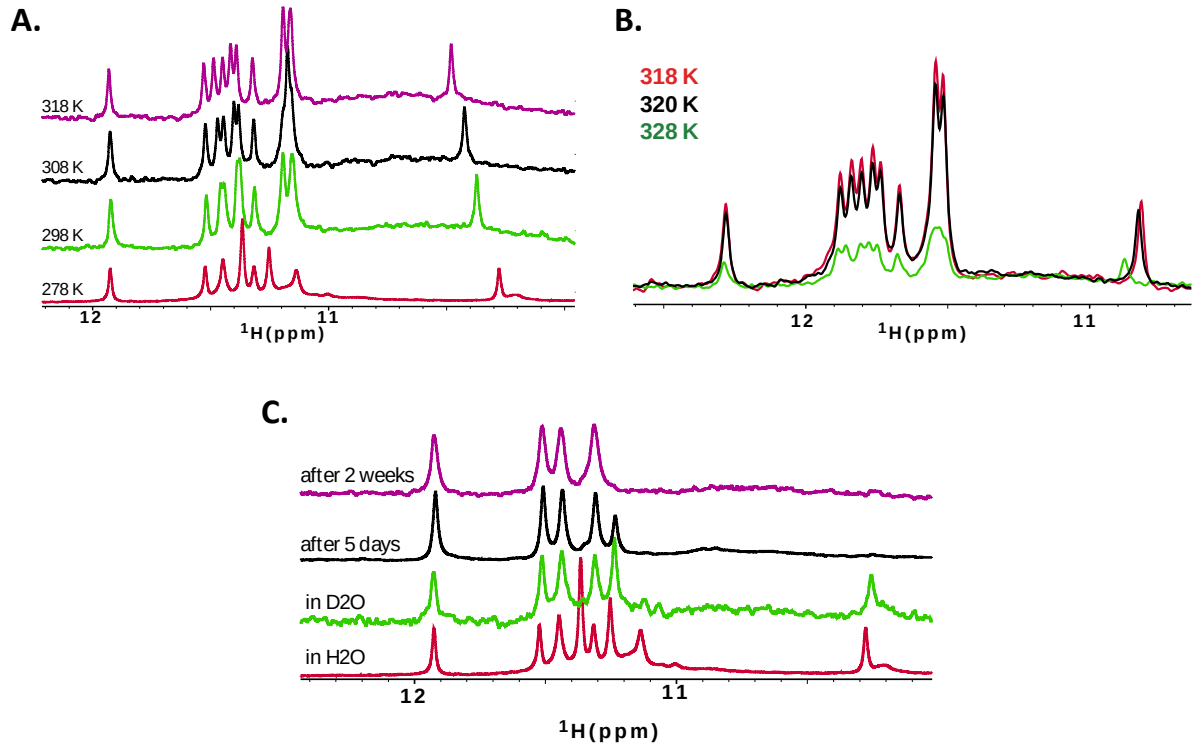
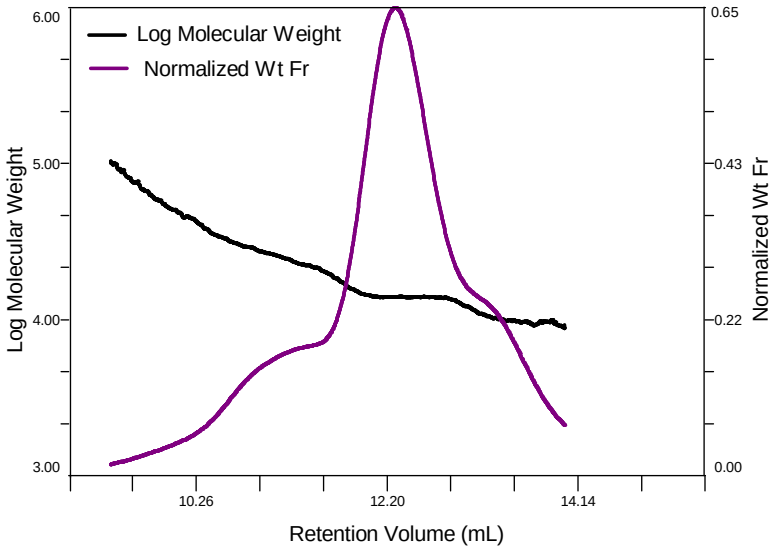



**Figure 3.30 Temperature and $D_2O$ exchange studies of 21mer. A.** The increase in temperature leads to splitting in overlapping imino resonances of 21mer G-quadruplex in $^1H$ 1D spectra. At 318K, the $^1H$ spectrum shows ten resonances whereas, two signals appear to be still duplets, indicating that there are in total twelve imino resonances **B.** G-quadruplex signals start melting above 318 K temperature, indicated by a reduction in the intensity of imino signals. **C.** $D_2O$ exchange shows that there are four imino resonances which are resistant to exchange even after two weeks, indicating that these residues are present in the core region, which keeps them protected from solvent. Overall, these results, suggests that 21mer adopts three planar topology.

Given the distribution of guanosine nucleotides in the 21mer RNA sequence, it is highly unlikely that it can adopt intramolecular, parallel, three planar G-quadruplex topology. Hence, the expected G-quadruplex has to be intermolecular with either dimeric or tetrameric topology. Also, the observed preference for slow cooling rather than snap cooling in order to refold into desired uniform conformation, suggests that 21mer adopts intermolecular G-quadruplex topology. Hence, to get more insights about the topology of 21mer G-quadruplex in solution, static light scattering was performed (Figure 3.31). The measurement showed one major peak flanked by two shoulder peaks. Importantly, the molecular weight obtained for the major peak corresponds to the size of 13.58 kDa, which is almost twice of the theoretical molecular weight of 21mer (7.23 kDa). This indicates that 21mer behaves as a dimer in solution and supports intermolecular three planar topology model of G-quadruplex.




| Theoretical mass (kDa) | Measured mass (kDa) |
|---|---|
| 7.23 | 13.58 |

**Figure 3.31 SLS analysis of 21mer RNA.** SLS data showing that major species of the 21mer RNA shows a mass of 13.58 kDa in solution and exists as a dimer. The data was recorded in 20 mM potassium phosphate buffer pH=6.5, 50 mM NaCl, 5 mM BME at 4 °C.

To carry out the $^{1}H$ resonance assignments of 21mer RNA in G-quadruplex form, imino NOESY in $H_2O$, $^{1}H$-$^{1}H$ TOCSY, and natural abundance $^{1}H$-$^{13}C$ HSQC experiments were recorded in presence of 5 mM KCl. The assignments were challenging because of lack of sufficient resolution of $^{1}H$-$^{1}H$ correlation imposed by too many guanine iminos. Imino NOESY spectrum in $D_2O$ aided the assignments of the core region of the G-quadruplex, as well as two more imino resonances protected from the solvent exchange. Remaining iminos were correlated with the help of imino NOESY spectrum recorded in $H_2O$. This correlation again supports 3-planar topology and orientation of each imino resonance could be identified relative other iminos but sequence specific could not be performed. 21mer RNA sequence shows the presence of two hotspots, which could be involved in the formation of discussed topology (**GGGG**AGGA**GGGGGG**CGUAUGA). But to correlate this information with the RNA sequence such as to identify which three guanines are involved in the G-quadruplex formation from each hotspot is a very challenging task. It requires identification of H8-H8 and H1-H1 NOE cross peaks from successive guanosines in the $^{1}H$-$^{1}H$ NOESY spectrum, which is further complicated by the syn or anti conformation of each nucleotide. Hence to assist resonance assignments, systematic replacement of certain guanosines by inosines, which maintains the same G-quadruplex fold but renders disappearance of associated guanosine peak. Identification of such resonances could be helpful in correlating the $^{1}H$-$^{1}H$ NOESY spectrum with RNA sequence. Or else, site-specific labeling strategy could be an effective but expensive alternative to achieve the similar results.

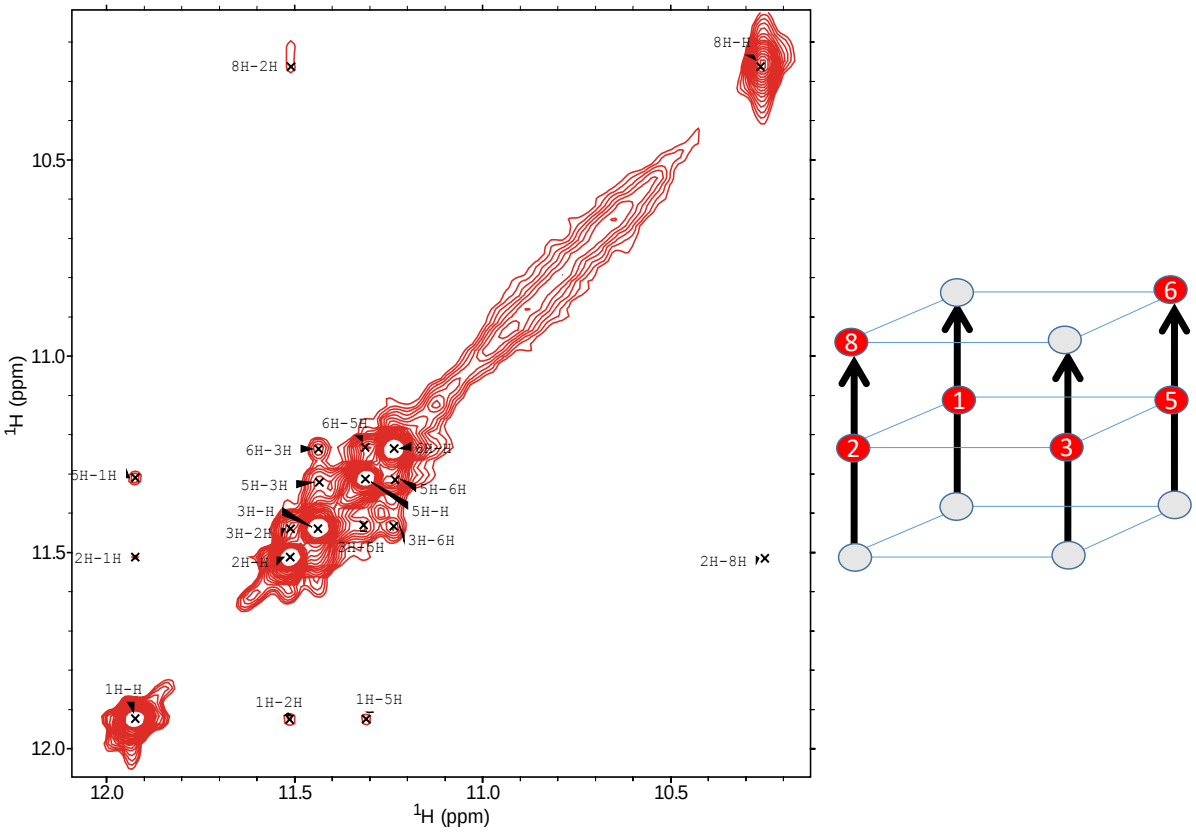



**Figure 3.32 $D_2O$ NOESY of 21mer RNA G-quadruplex.** D2O NOESY spectrum showing cross-peaks observed for solvent protected 21mer G-quadruplex resonances and tentative resonance assignments. The data was recorded on 500 µM RNA in 5 mM KCl at 278 K, 600 MHz spectrometer with mixing time of 300 ms.

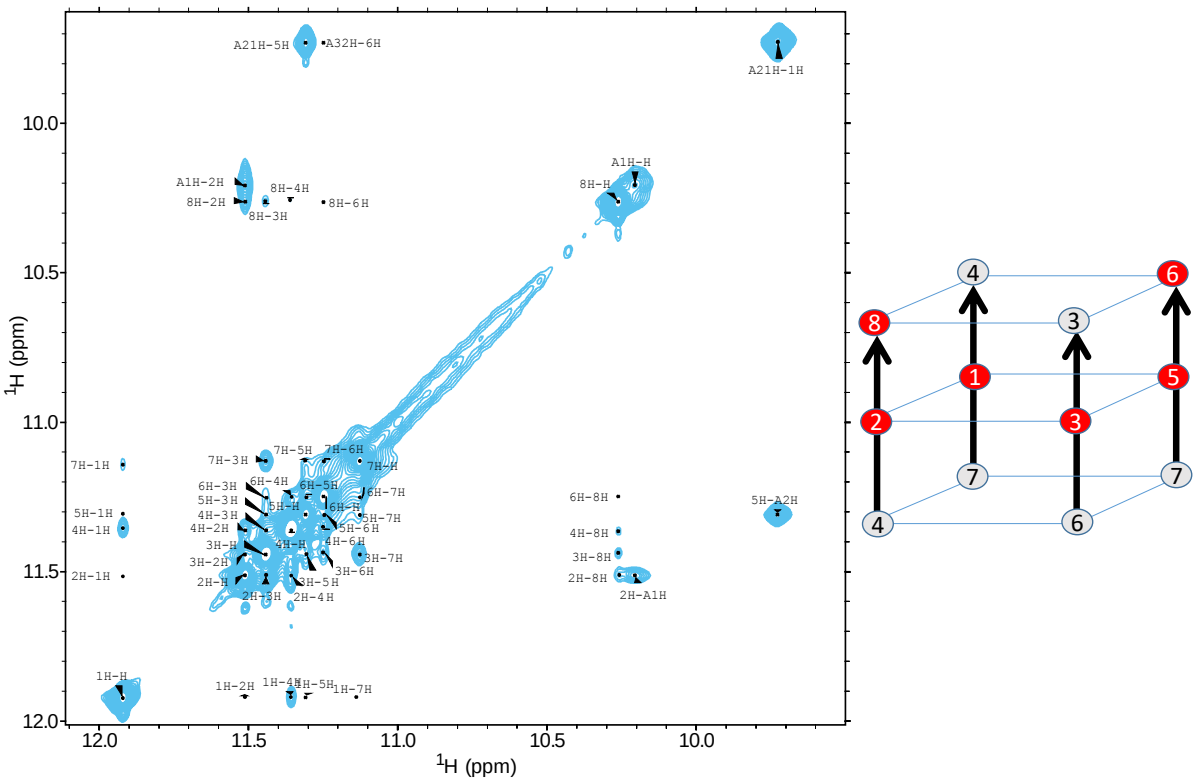



**Figure 3.33 $H_2O$ NOESY of 21mer RNA G-quadruplex.** NOESY spectrum showing cross-peaks of 21mer RNA G-quadruplex resonances and tentative resonance assignments. The data was recorded on 500 µM RNA in 5 mM KCl at 278 K, 600 MHz spectrometer with mixing time of 300 ms.

In summary, the biophysical characterization suggests that 21mer forms parallel, dimeric G-quadruplex with three planar topology. It also strongly suggests that two planes of the G-quadruplex are also involved in hexad formation with adenine nucleotide (Figure 3.34).

5'- GGGGAGGAGGGGGGCGUAUGA **-3'**

5'- GGGGAGGAGGGGGGCGUAUGA **-3'**

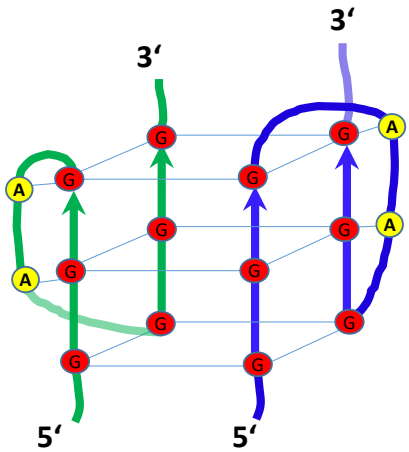



**Figure 3.34 21mer G-quadruplex model.** Schematic representation of topology adopted by 21mer G-quadruplex in the presence of low concentration of KCl.

## 3.1.3 Characterization of LS2 RRM domains and poly G RNA interaction

### 3.1.3.1 Interaction of LS2 RRM1,2 with 21mer poly G RNA by NMR

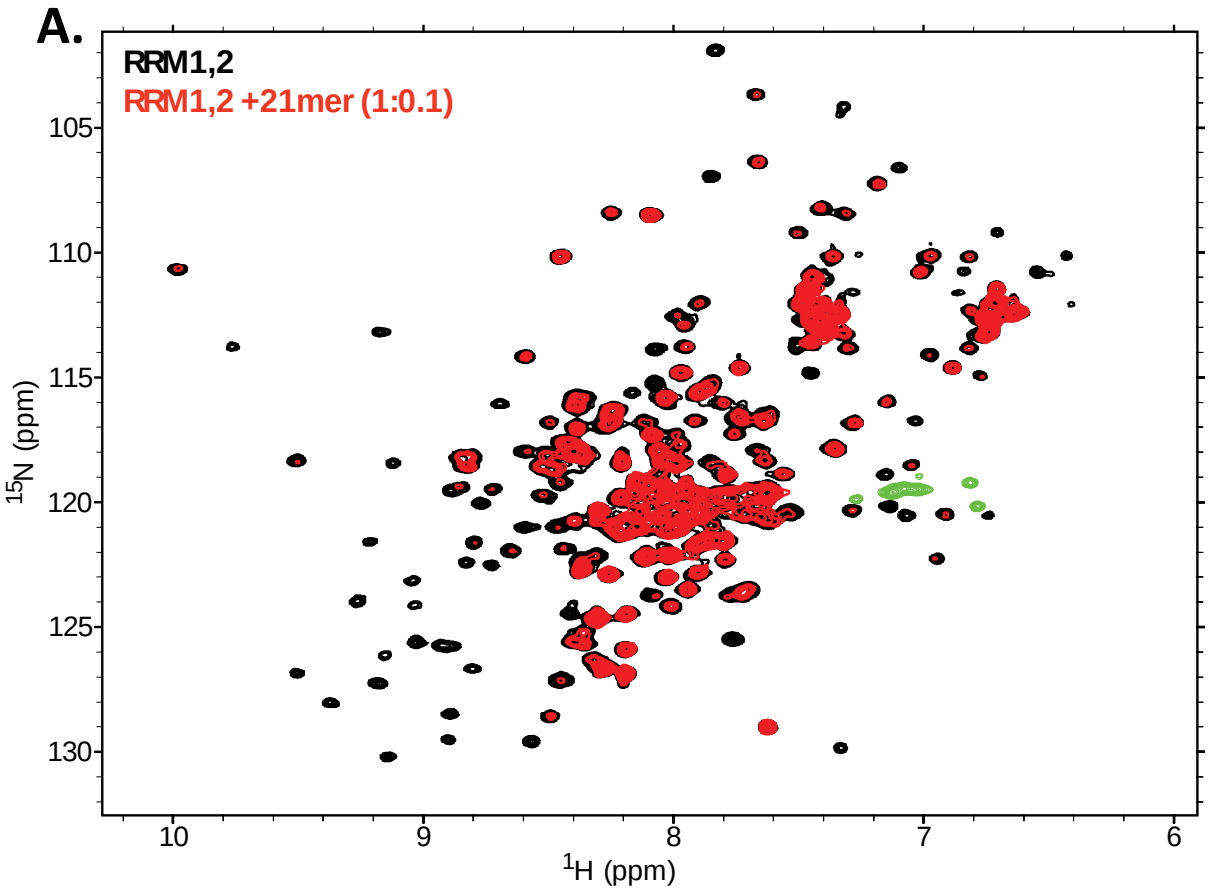
To characterize the interaction between LS2 RRM1,2 and 21mer poly G RNA NMR titration was performed. For titration lower temperature (283 K) was chosen to keep protein soluble. Changes in protein and RNA resonances were monitored by $^{1}H$-$^{15}N$ HSQC and $^{1}H$ 1D NMR experiments, respectively (Figure 3.35).

Upon addition of 0.1 molar excess of RNA, RRM1,2 protein signals showed line broadening along with small chemical shift changes, indicating protein-RNA interaction. Upon successive addition of RNA, protein signals disappeared even more and did not reappear even after the addition of 1.5 molar excess of RNA. Chemical shift changes, as well as intensity loss, was observed in both RRM domains, indicating that both RRM domains are involved in RNA binding. LS2 specific linker residues also showed chemical shift perturbations as well as loss in intensity. But, as the changes were observed all over the protein, it cannot be concluded whether the interaction mediated by these residues was specific or not. Comparison of the intensity ratio of the HSQCs in the absence and the presence of 0.2 molar excess RNA, indicated that signal intensity of the all the residues was reduced by more than half (except some residues located in the linker region).

Loss of signal intensity upon RNA addition could be the result of intermediate exchange regime of the protein-RNA interaction which is characterized by line broadening of the NMR signals. Protein can also bind to RNA in multiple registers, causing it to slide over the RNA, and thereby resulting into line broadening of the signals. Or else, molecular size of the protein-RNA complex gets bigger, causing it to tumble slower and hence results in line broadening of the signals. On the other hand, loss of signal intensity could also mean that upon interaction with RNA, protein is no more soluble in the solution and forms a thick aggregate which is not detectable by NMR spectroscopy.

Indeed, the protein sample after RNA titration showed the presence of aggregation but not in the form of thick precipitate as observed for Apo form of the protein, but rather as a gel-like form. This indicates that loss signal intensity upon RNA titration was a result of a protein precipitation, though intermediate exchange regime, as well as binding in multiple registers, cannot be ruled out.

On the other hand, upon successive addition of RNA, $^{1}H$ 1D resonances specific to G-quadruplex, both in uniform conformation as well as multiple conformations became stronger. These resonances could be attributed to the free form of RNA, as most of the protein is precipitated during titration. Given the tendency of 21mer to adopt G-quadruplex conformation and also the presence of G-quadruplex-specific signals during titration indicates that protein may interact with a G-quadruplex form of RNA.


A.
RRM1,2
RRM1,2 +21mer (1:0.1)
105
110
115
120
125
130
15N (ppm)
10
9
8
7
6
1H (ppm)

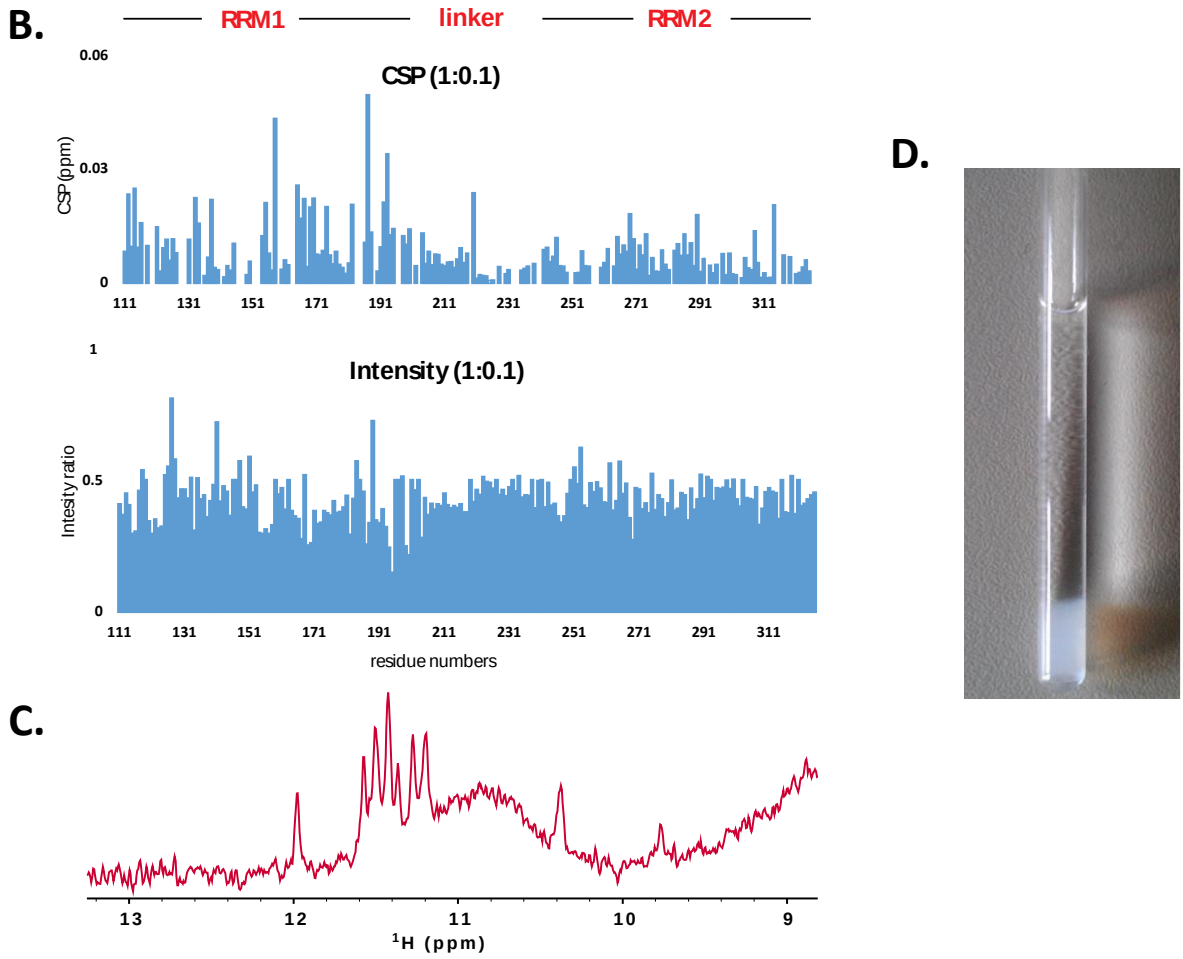


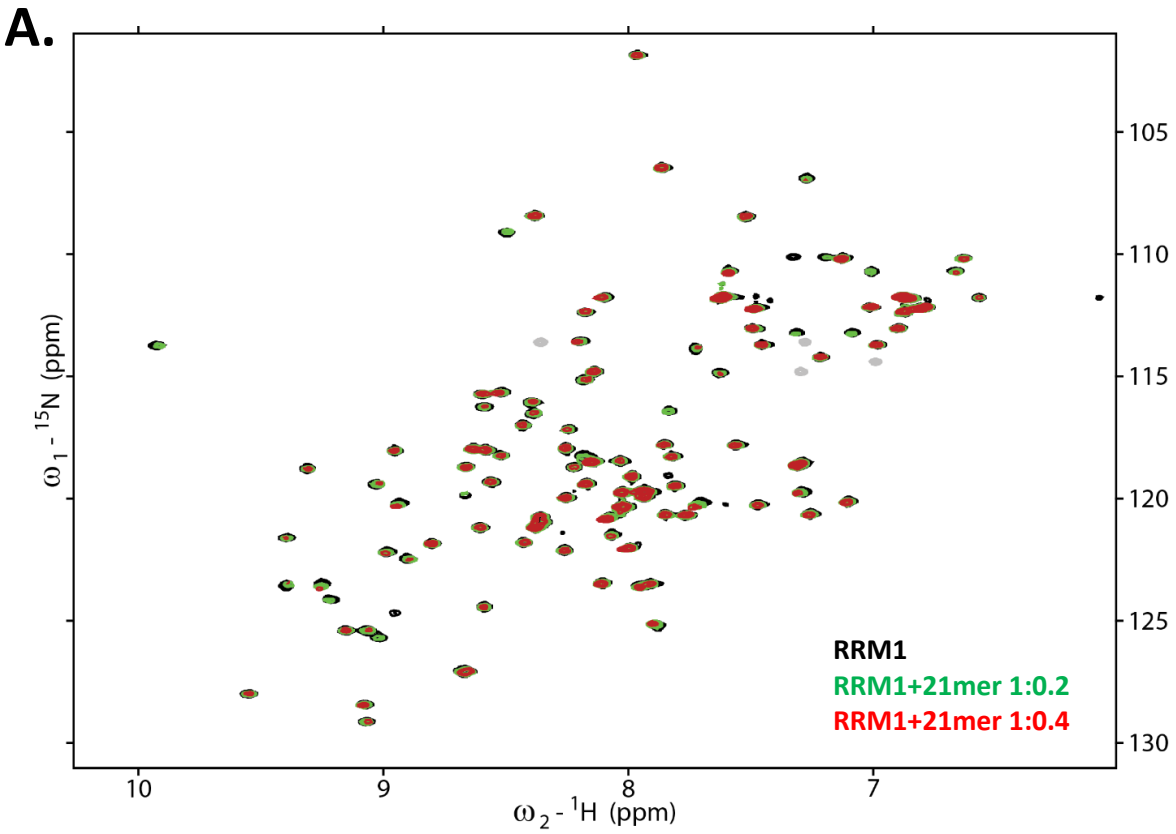

**Figure 3.35 NMR titration of RRM1,2 with 21mer. A.** The overlay of $^{1}H$-$^{15}N$ spectra of RRM1,2 in the absence of RNA (black) and in the presence 0.1 molar excess of 21mer RNA. It shows that upon binding with RNA, RRM1,2 resonances undergo line broadening. The experiment was performed at 500 MHz and 288 K in the presence of 20 mM potassium phosphate pH 6.5, 50 mM NaCl and 5 mM DTT. **B.** Plots of chemical shift perturbations and intensity ratio with respect to residue numbers show that RRM1 shows major chemical shift perturbations while intensity loss is observed throughout protein sequence. **C.** $^{1}H$ NMR spectrum shows the presence of G-quadruplex signals in a sample in 1.5 molar excess of RNA, which indicate that RNA adopts G-quadruplex conformation during titration. **D.** Gel-like agglomeration observed during protein-RNA titration.

### 3.1.3.2 Interaction of individual RRM domains with 21mer by NMR

The interaction of individual RRM domains with 21mer RNA was characterized by NMR titrations. Because of aggregation-prone nature of RRM1,2 and 21mer, important information about RNA-binding residues, $K_d$ of the interaction as well binding stoichiometry was missing. It was speculated that titration with individual RRM domains will answer these questions.

Unfortunately, titration of RRM1 with 21mer resulted in aggregation similar to what has been observed for RRM1,2 (Figure 3.36). Upon successive addition of RNA, RRM1 started forming a gel-like aggregation. Nevertheless, chemical shift changes could be traced. The plot of

chemical shift perturbation with respect to residues number shows that with the exception of α1 helix, the chemical shift changes were observed all over the protein.



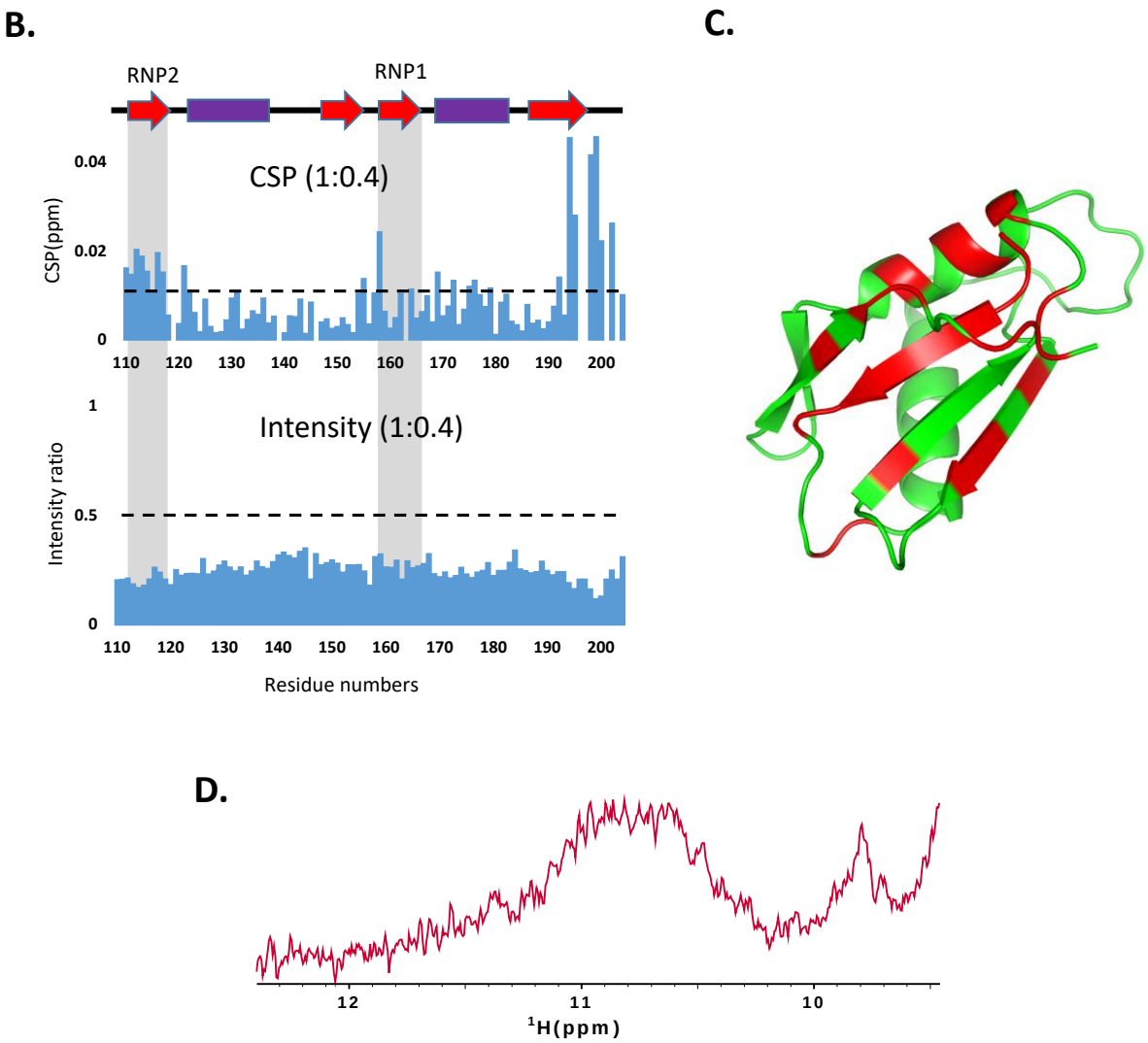


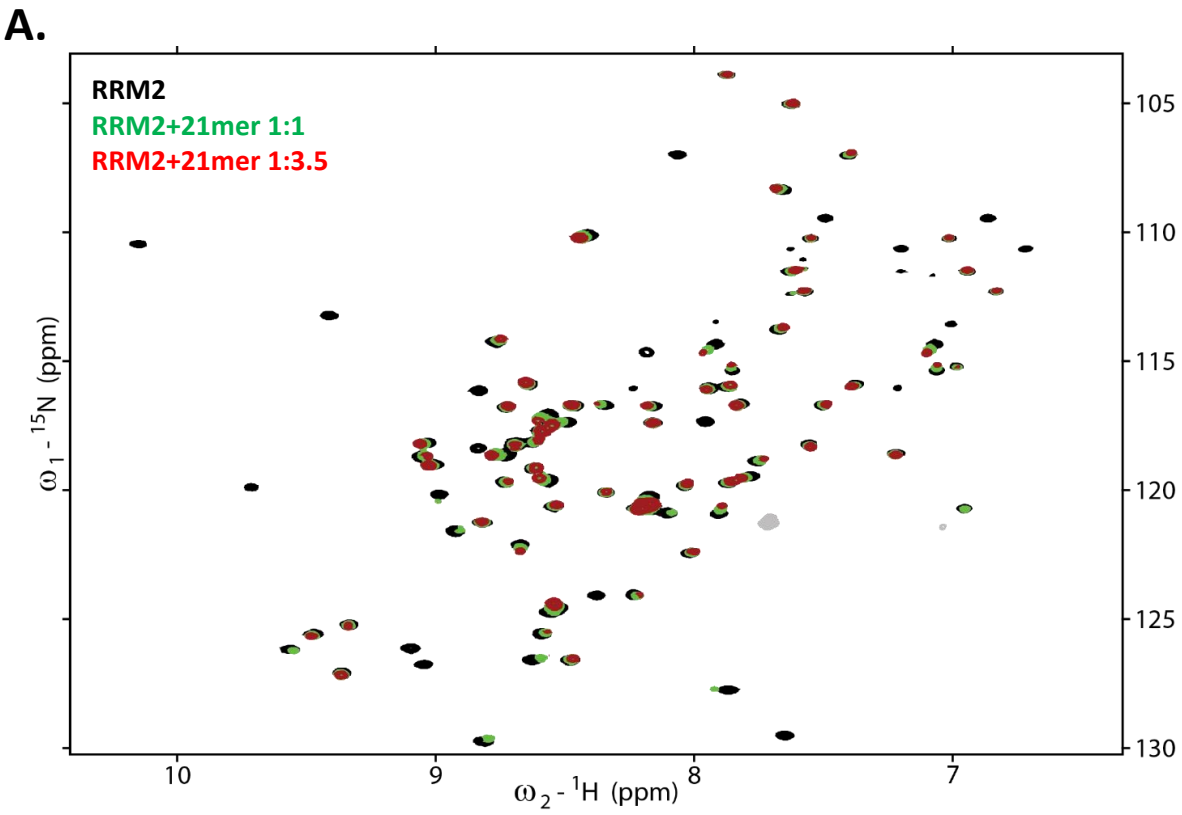

**Figure 3.36 Titration of RRM1 with 21mer. A.** The overlay of $^{1}H$-$^{15}N$ spectra of RRM1 in the absence of RNA (black) and in the presence 0.2 molar excess (green) and 0.4 molar excess (red) of 21mer RNA. It shows that in the presence of RNA, RRM1 undergoes line broadening as well as chemical shift changes. The experiment was performed at 600 MHz and 298 K in the presence of 20 mM potassium phosphate pH 6.5, 35 mM NaCl and 5mM DTT. **B.** Plots of chemical shift perturbation (CSP) with respect to residue numbers show that major chemical shift perturbations are located on β1 and β4 strand followed by C-terminal flexible residues. Surprisingly, no changes were observed on RNP1 residues which are located on β3 strand. On the other hand, intensity loss was observed all over protein, which is shown by intensity plot. **C.** Residues undergoing **c**hemical shift perturbations mapped in red on the ribbon structure of RRM1. **D.** $^{1}H$ NMR spectrum showing the presence of G-quadruplex signals in a sample in 0.4 molar excess of RNA, which indicates that RNA adopts G-quadruplex conformation during titration.

Surprisingly, no significant chemical shift perturbations were observed for the RNP1 residues of the RRM1, which are usually involved in crucial base stacking for canonical RRM domains. Similarly, the additional loop from RRM1 did not show any significant chemical shift changes, indicating that it does not play any role in RNA binding.

In contrast to RRM1,2 and RRM1, upon titration of RRM2 with 21mer, specific changes were observed (Figure 3.37). Successive RNA addition resulted into major line broadening of certain

resonances, whereas minor chemical shift perturbations were observed. This line broadening of the resonances was very specific to the residues located on the β-strands (involving both RNP sites), as opposed to non-specific aggregation of RRM1,2 or RRM1.

Again, $^{1}$H 1D spectrum showed that RNA is in G-quadruplex form, as observed for RRM1,2 and RRM1.



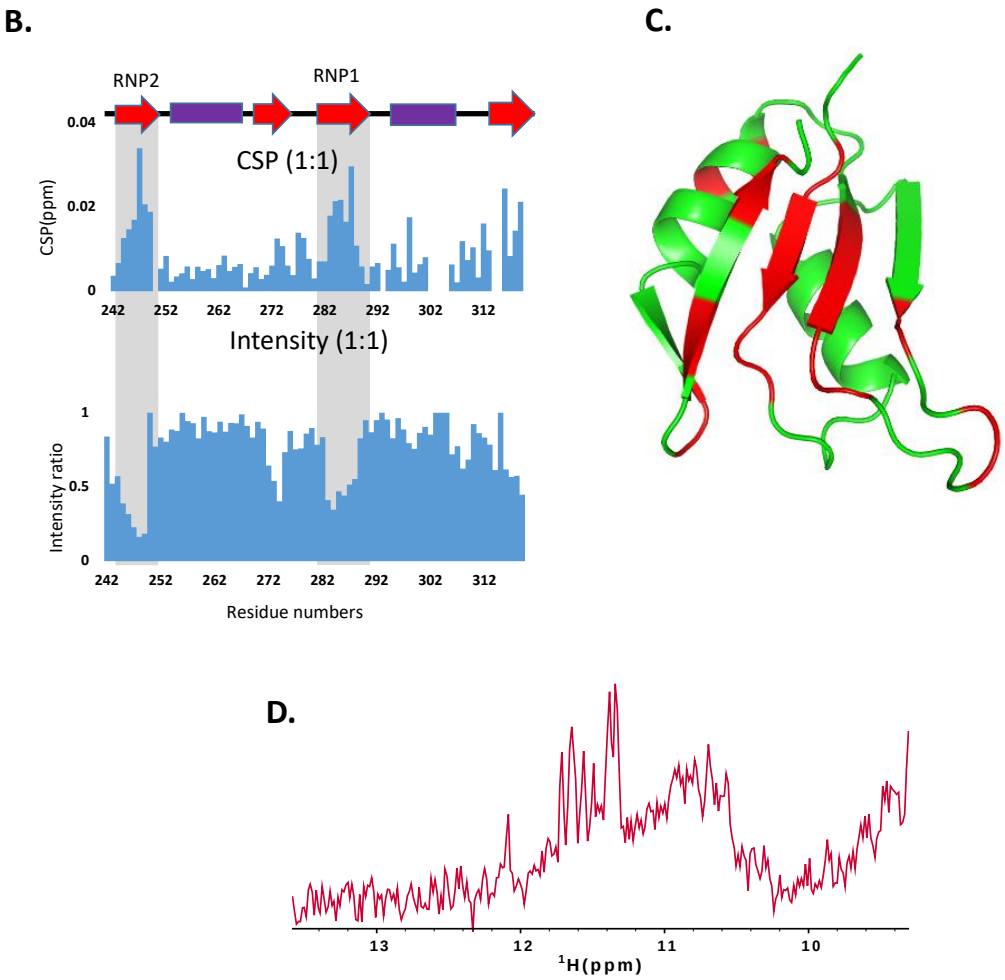



**Figure 3.37 Titration of RRM2 with 21mer. A.** $^{1}H$-$^{15}N$ spectra of RRM2 in the absence of RNA (black) and in the presence 1 molar excess (green) and 3.5 molar excess (red) of 21mer RNA. The overlay shows that in the presence of RNA, RRM2 undergoes line broadening as well as chemical shift changes. The experiment was performed at 600 MHz and 298 K in the presence of 20 mM potassium phosphate pH 6.5, 50 mM NaCl and 5mM DTT. **B.** Plots of chemical shift perturbation and intensity with respect to residue numbers show that mainly residues located on β-strands show effect upon RNA binding, with β1 (RNP2 site) and β3 (RNP1 site) showing maximum effect. **C.** Residues undergoing chemical shift perturbations mapped in red on the ribbon structure of RRM2. **D.** $^{1}H$ NMR spectrum shows the presence of G-quadruplex signals in a sample in 3.5 molar excess of RNA indicating that RNA adopts G-quadruplex conformation during titration.

### 3.1.3.3 Interaction of single RRM domains with shorter poly G oligonucleotides

LS2 RRM1,2 domain construct as well as individual RRM domains interaction with 21mer RNA resulted in the line broadening of the resonances, which is not suitable for studying protein-RNA complex using NMR. Typically, single RRM is known to bind four RNA nucleotides, hence we hypothesize that may 21mer is too long for LS2 RRM domains and this may be causing a problem such as multiple register binding. Also, change in RNA length could also potentially affect the interaction regime and might shift it towards fast or slow exchange regime. Shorter nucleotides such as 14mer, 8mer, 5mer were designed according to SELEX experiments

carried out on LS2 and tested for their ability to adopt G-quadruplex structure (see 3.1.2.3 Designing Shorter oligonucleotides on the basis of SELEX).

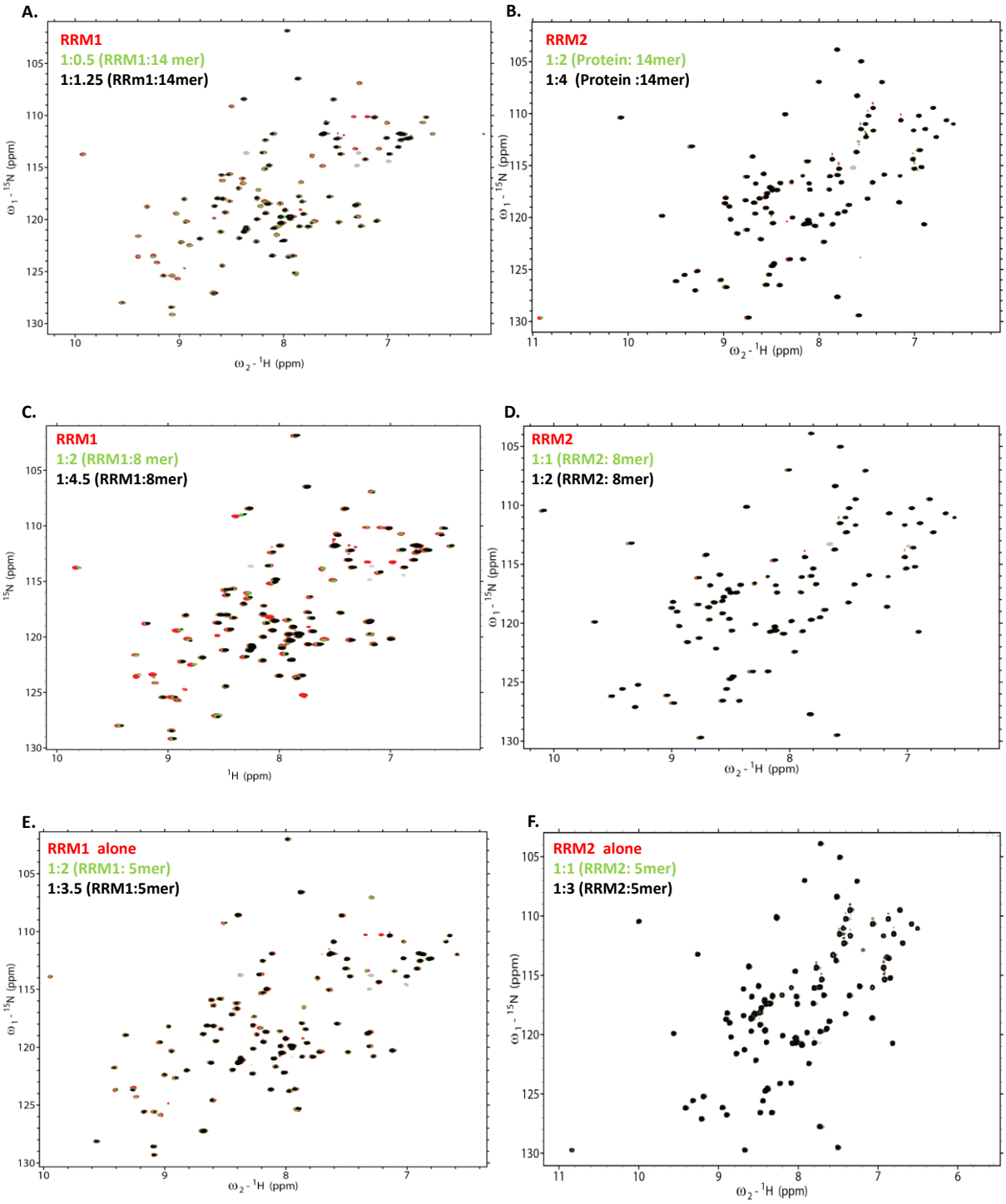



**Figure 3.38 Summary of NMR titration of single RRM domains with shorter oligonucleotides.** The overlay of $^{1}H$-$^{15}N$ HSQC spectra for titration of RRM1 with **(A.)** 14mer **(C.)** 8mer **(E.)** 5mer shows that RRM1 interacts with all 3 oligonucleotides**.** On the other hand, titration of RRM2 with **(B.)** 14mer **(D.)** 8mer **(F.)** 5mer shows that RRM2 doesn't interact with any of these oligonucleotides. All the titration experiments were recorded in 20 mM potassium phosphate buffer pH 6.5, 50 mM NaCl, 5 mM DTT at 298 K.

Titration of RRM1 with 14mer, 8mer and 5mer showed chemical shift perturbations with all the oligonucleotides (for 5mer at higher concentration) (Figure 3.38, A, C, E). As like interaction with 21mer, the chemical shift changes upon RNA titration were located all over the protein except α1 helix. It is important to note that 14mer, 8mer exist in multiple conformations in the buffer used for the protein-RNA titration (20 mM potassium phosphate, 50 mM NaCl, 5 mM DTT). RNA 1D spectrum recorded during the titrations confirmed that all the RNA exists in the G-quadruplex form albeit in multiple conformations during titration. Thus, this data indicates that RRM1 does not show any preference for G-quadruplex structure or rather interacts with the G-quadruplex RNA non-specifically.

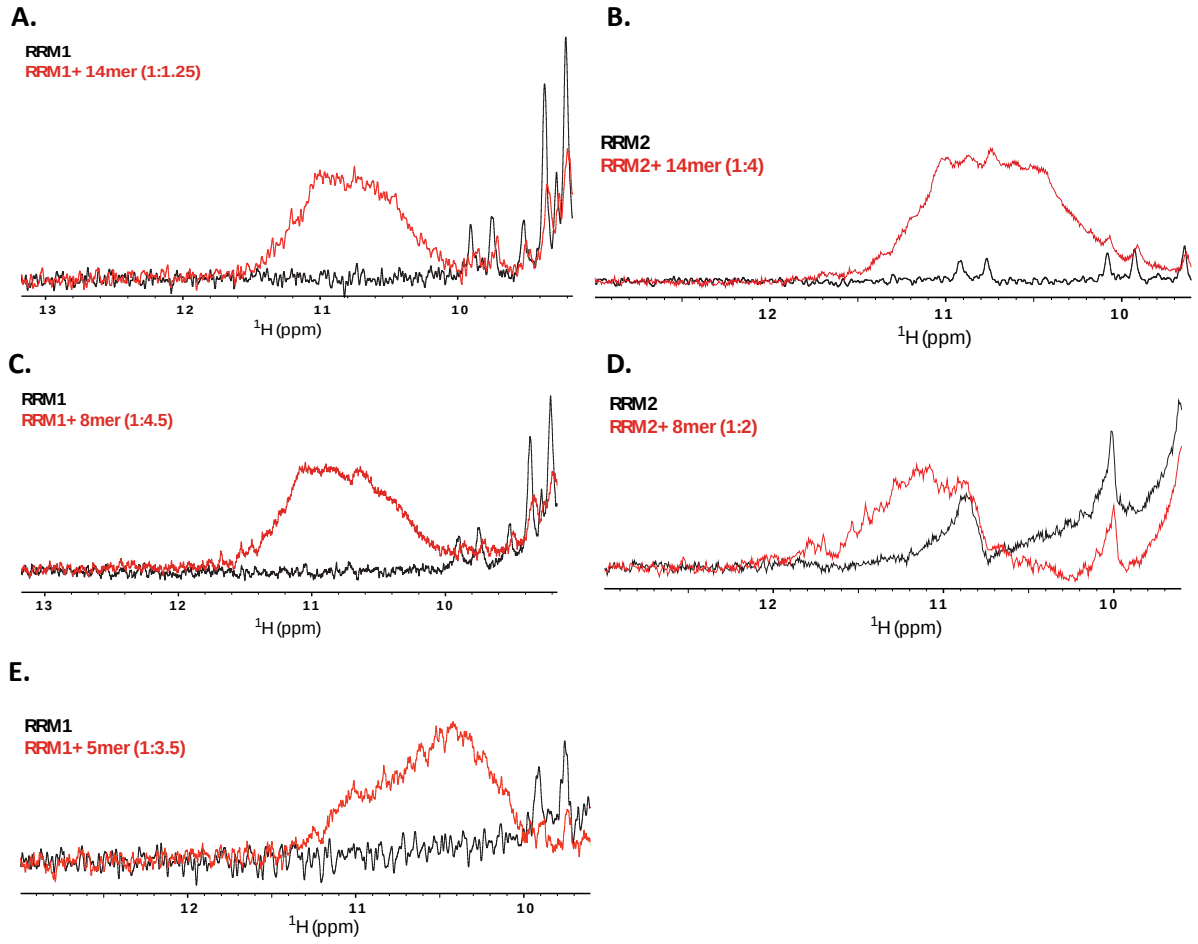



**Figure 3.39 G-quadruplex resonances during single RRM domains with shorter oligonucleotides titration.** Overlay of $^1$H 1D NMR spectra shows that during each titration G-quadruplex signals (albeit in multiple conformations) were observed.

On the other hand, titration of RRM2 with 14mer, 8mer and 5mer showed no or little binding (Figure 3.38, B, D, F). RNA 1D showed all the RNA formed G-quadruplex but with multiple conformations. 8mer still showed some binding, as compared to 14mer and 5mer, which

could be speculated to be because of residual structured G-quadruplex species present in the sample, which 8mer could form. This result indicated that RRM2 is specific for 21mer RNA.

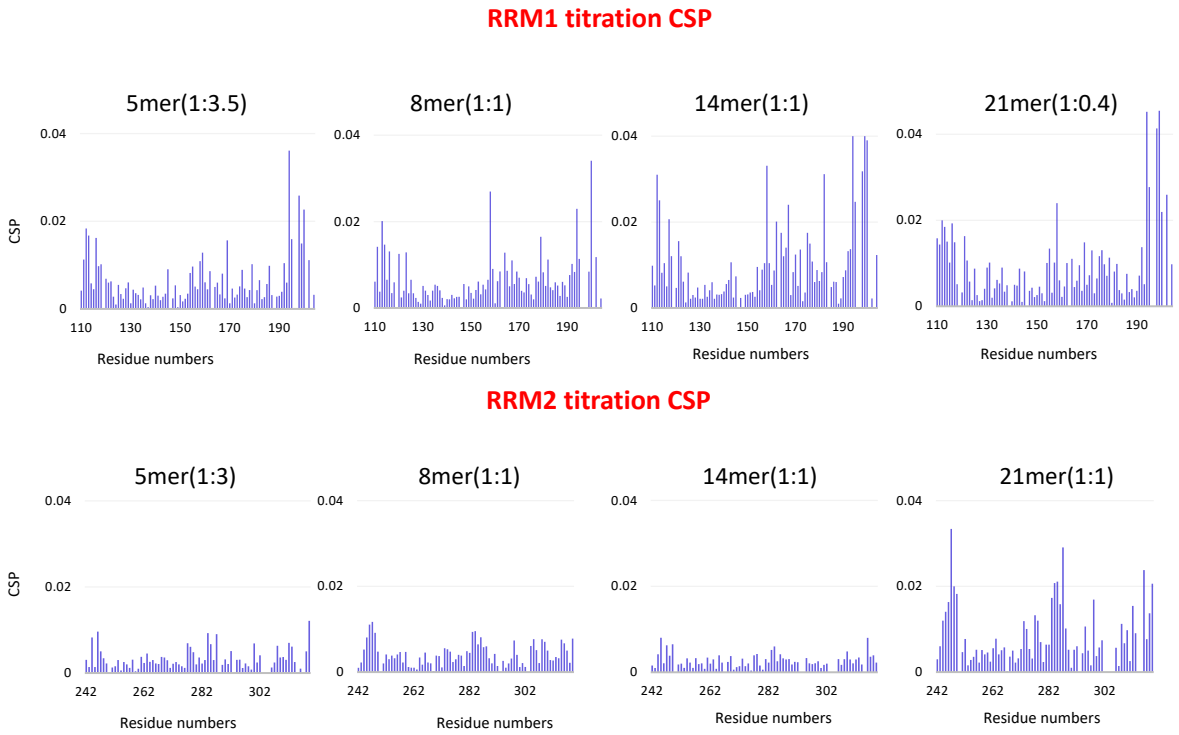



**Figure 3.40 Comparison of CSPs of single RRM domains during NMR titration.** RRM1 interacts with all the oligonucleotides, the CSP values gets smaller with a decrease in RNA length. Also, 5mer oligonucleotide is needed in a higher amount for the interaction. On the other hand, RRM2 shows specific interaction with 21mer.

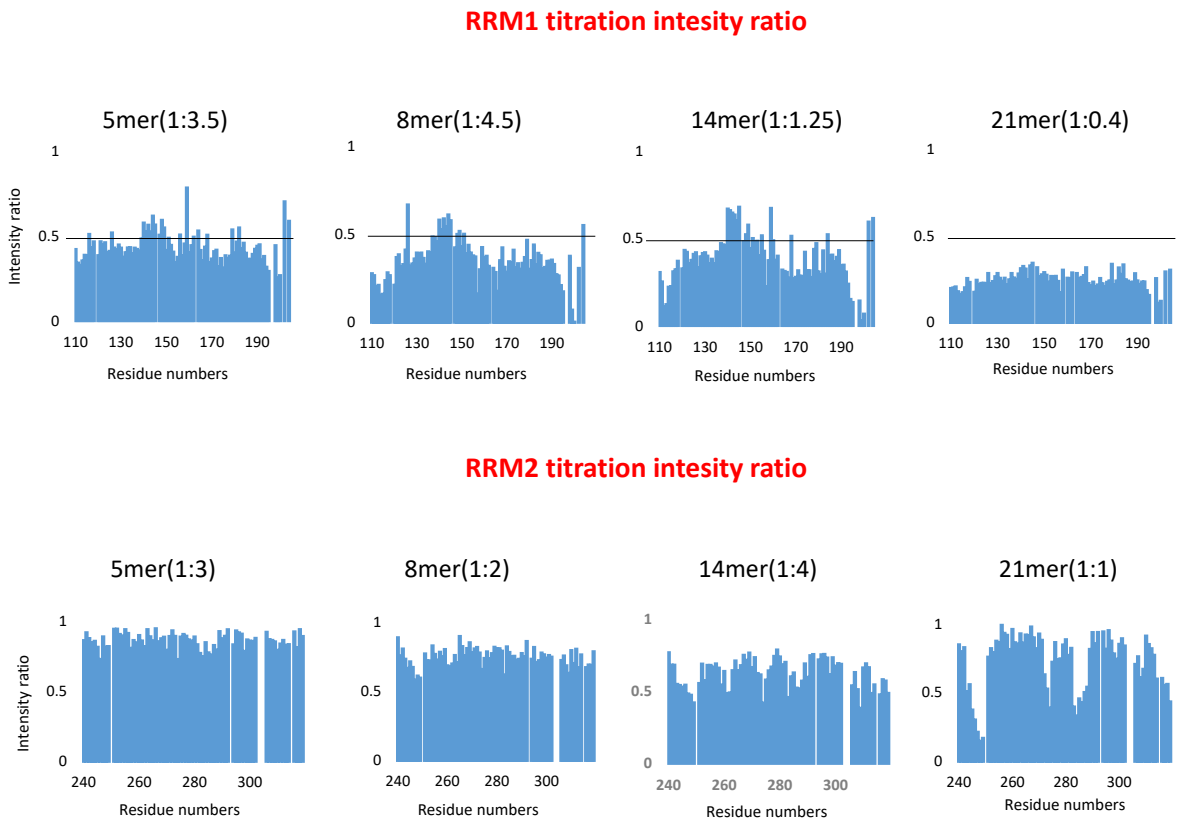



**Figure 3.41 Comparison of intensity change of single RRM domains during NMR titration.** In the case of RRM1 intensity loss is observed all over the protein characterized by precipitation in presence of all the RNA oligonucleotides. On the other hand, in the case of RRM2, specific residues show line broadening in the presence of RNA.

### 3.1.3.4 RRM2 specifically interacts with 21mer G-quadruplex structure

The cause of interaction of RRM2 with 21mer could be either sequence specific or structure specific. i.e. specificity is determined by the 7 oligonucleotides present at the 3'end of the 21mer sequences which are not part of 14mer or 8mer, or RRM2 rather specifically interacts with the G-quadruplex structure adopted by 21mer RNA, which 14mer and 8mer fail to adopt.

To check whether RRM2 interacts with 7mer non-G sequence present in 21mer, NMR titration of respective protein-RNA was performed. The addition of 7mer didn't result in any significant chemical shift perturbations or line broadening of RRM2 resonances in $^{1}H$-$^{15}N$ HSQC, indicating that RRM2 does not interact with this sequence.

Next, to check whether RRM2 interacts with free GTP, which doesn't adopt G-quadruplex structure, titration of GTP with RRM2 was performed. The addition of GTP also didn't have significant any effect on the RRM2 resonances monitored by $^{1}H$-$^{15}N$ HSQC, except chemical shift perturbations of Histidine residues, caused by a change in pH upon addition of GTP. This

shows that RRM2 does not interact with free GTP nucleotides either and thus, might be specific for G-quadruplex structure adopted by 21mer.

To confirm that RRM2 indeed interacts with 21mer specific G-quadruplex structure, reverse NMR titration was performed in which 21mer was in G-quadruplex form and RRM2 was titrated into it. The changes in RNA resonances were monitored by $^{1}$H 1D NMR. Upon successive addition of RRM2, G-quadruplex signals showed line broadening along with minor chemical shift perturbations, same as observed for forward titration performed with RRM2 and 21mer. This confirms that RRM2 specifically interacts with the G-quadruplex structure adopted by 21mer RNA.

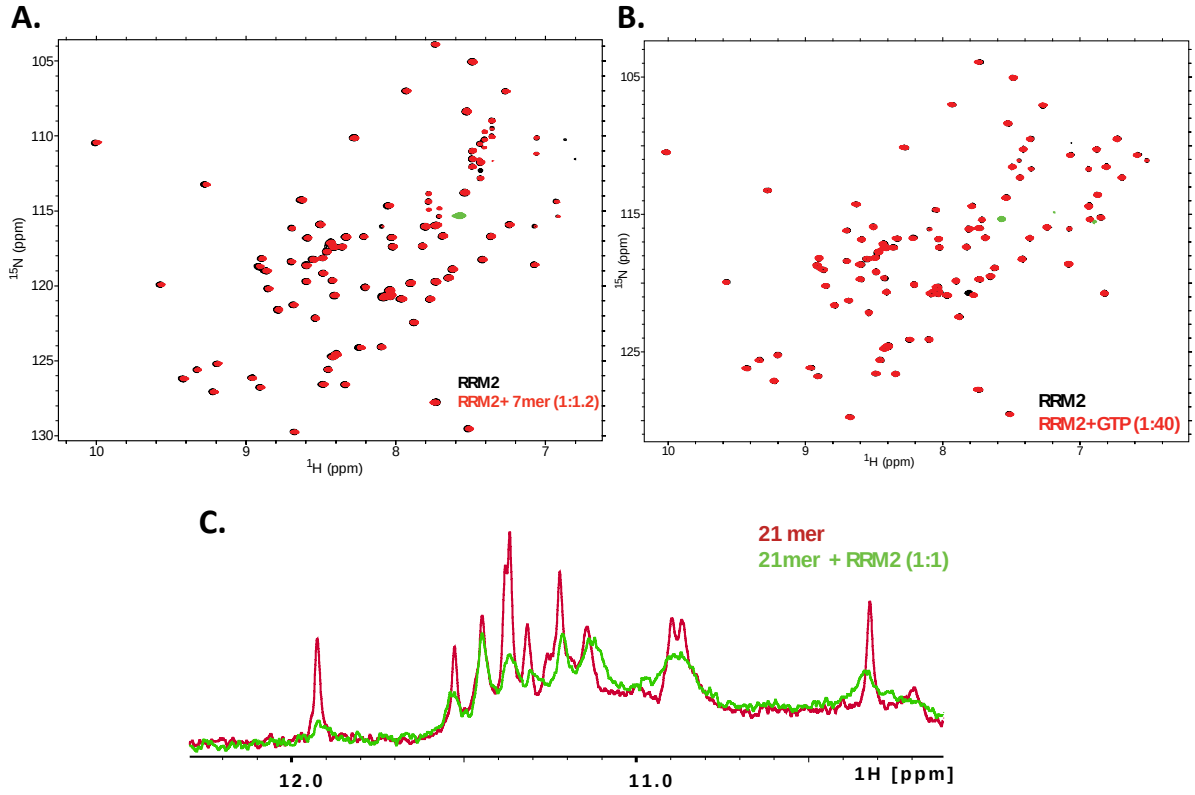
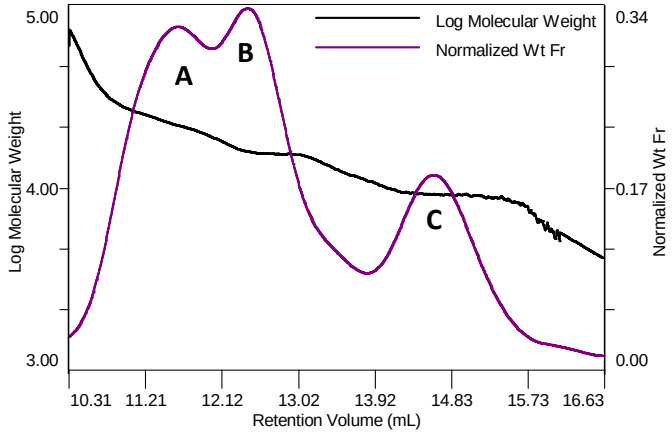



**Figure 3.42 RRM2 shows specificity for 21mer G-quadruplex structure. A.** $^{1}$H-$^{15}$N spectra of RRM2 in the absence (black) and in the presence 1.2 molar excess (red) of 7mer (oligonucleotides at 3'end of 21mer) shows that RRM2 does not interact with 7mer. The experiment was performed at 800 MHz and 298 K in the presence of 20 mM potassium phosphate pH 6.5, 50 mM NaCl and 5mM DTT. **B.** $^{1}$H-$^{15}$N spectra of RRM2 in the absence (black) and in the presence 40 molar excess (red) of GTP showing that RRM2 does not interact with GTP. The experiment was performed at 500 MHz and 298 K in the presence of 16 mM MES pH 6.5, 40 mM KCl and 5mM DTT. **C.** $^{1}$H NMR spectra showing line broadening of 21mer G-quadruplex resonances in presence of RRM2 (green), confirming the interaction of RRM2 with 21mer G-quadruplex. The experiment was performed at 900 MHz and 278 K in the presence of 5 mM potassium phosphate pH 6.5.

### 3.1.3.5 SLS analysis of RRM2-21mer complex

SLS studies were performed on RRM2-21mer complex to get more insights about the molecular mass of the complex as well as binding stoichiometry (Figure 3.43). The data shows that RRM2-21mer complex doesn't pass through the column as a single species. This could be seen by the three peaks during measurement with molecular masses similar to the free form of RNA (13.58 kDa) and protein (9.2 kDa). The poor data quality of the third peak, which may be comprised of protein-RNA complex (1:1), made the data analysis not reliable.




| peak | Measured mass (kDa) | Expected component |
|---|---|---|
| A | 25.6 kDa | Complex (1:1) |
| B | 14.5 kDa | 21mer (13.58 kDa) |
| C | 8.9 kDa | RRM2 (9.2 kDa) |

**Figure 3.43 SLS analysis of RRM2-21mer complex.** SLS data shows that RRM2-21mer complex shows 3 peaks. The two peaks show the molecular mass of 14.5 kDa and 8.9 kDa and could be attributed to the free form of RNA and protein, respectively. The third peak shows measured mass of around 25.6 kDa, but the data is not reliable because of lack of monodispersity of the sample component.

### 3.1.3.6 Linker-RRM2 interaction with 21mer

N-terminal region of LS2 linker has an additional patch of charged, aromatic residues which adopts helical conformation and could potentially interact with RNA. At the same time, a C-terminal region of the linker interacts with the RNA binding β-strands of the RRM2. Hence, it would be interesting to check whether in the presence of linker-RRM2 still binds to 21mer RNA and if it does then is there any change in the interaction regime. It would be also interesting to check whether in the presence of linker the exchange regime of the interaction or the binding in multiple registers.

In order to check this out, titration of linker-RRM2 with 21mer was performed. Upon addition of RNA, protein resonances again started disappearing, indicating that in the presence linker-RRM2 also RRM2 interacts with 21mer RNA. As observed for individual RRM2, residues located on β-strands of RRM2 showed line broadening after a successive addition of RNA. At the same time, the N-terminal residues of linker, which show helical propensity also showed line broadening along with minor chemical shift perturbations indicating that they also interact with the RNA.

21mer G-quadruplex interaction with linker-RRM2 also resulted in line broadening of G-quadruplex signals indicating interaction.

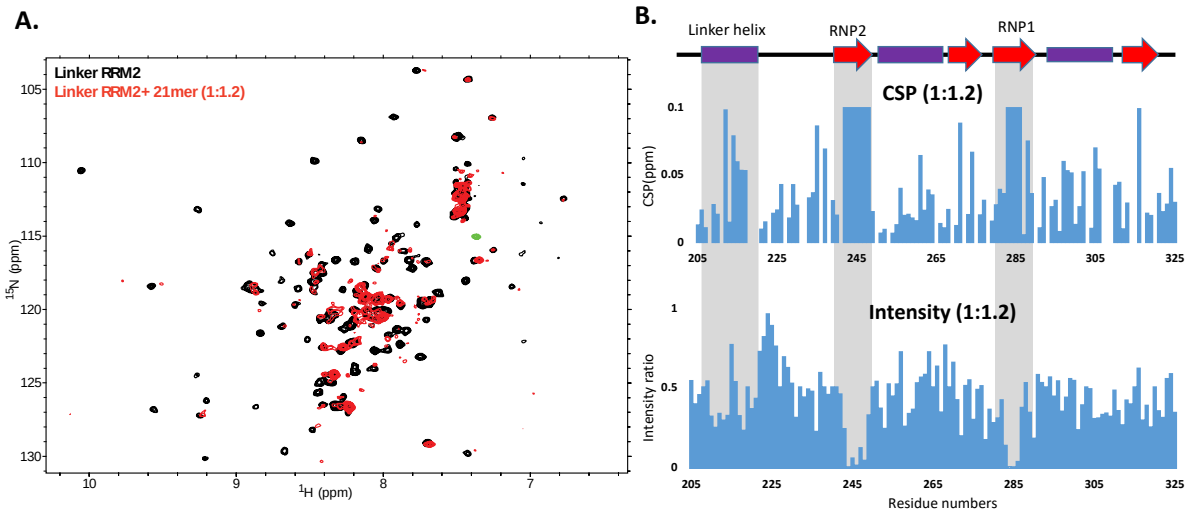



**Figure 3.44 Titration of Linker-RRM2 with 21mer. A.** $^{1}$H-$^{15}$N spectra of RRM2 in the absence of RNA (black) and in the presence 1.2 molar excess (red) of 21mer RNA recorded at 500 MHz and 298 K in the presence of 5 mM potassium phosphate pH 6.5 and 2mM DTT. The overlay shows that upon addition of RNA, the majority of RRM2 signals show line broadening. **B.** Plots of CSP and intensity with respect to residue numbers show that along with the RNP sites of RRM2 (β1 and β3), LS2 linker specific residues, which form a helix, are also involved in RNA binding and thus shows CSPs as well as intensity loss.

## 3.1.3.7 LS2 interacts with the uniform conformation adopted by 21mer

$^{1}$H 1D NMR spectra of 21mer G-quadruplex shows two sets of resonances, one specific for uniform conformation, which appears to be arising from dimer and another broad peak around 11 ppm, characterized by multiple conformations which may be arises from an oligomeric form of G-quadruplex.

$^{1}$H RNA 1D NMR spectrum showed that 21mer RNA samples used for the titration also contained these two forms. To check whether LS2 interacts with 21mer in multiple conformational forms, titration of linker-RRM2 with multiple conformations was performed. Upon successive addition of this RNA, no significant changes were observed. The observed line broadening was very small, which could be attributed to the presence of minor

population, which is uniform G-quadruplex. Thus, indicating that linker-RRM2 specifically interacts with uniform G-quadruplex conformation and not with oligomeric multiple conformational forms.

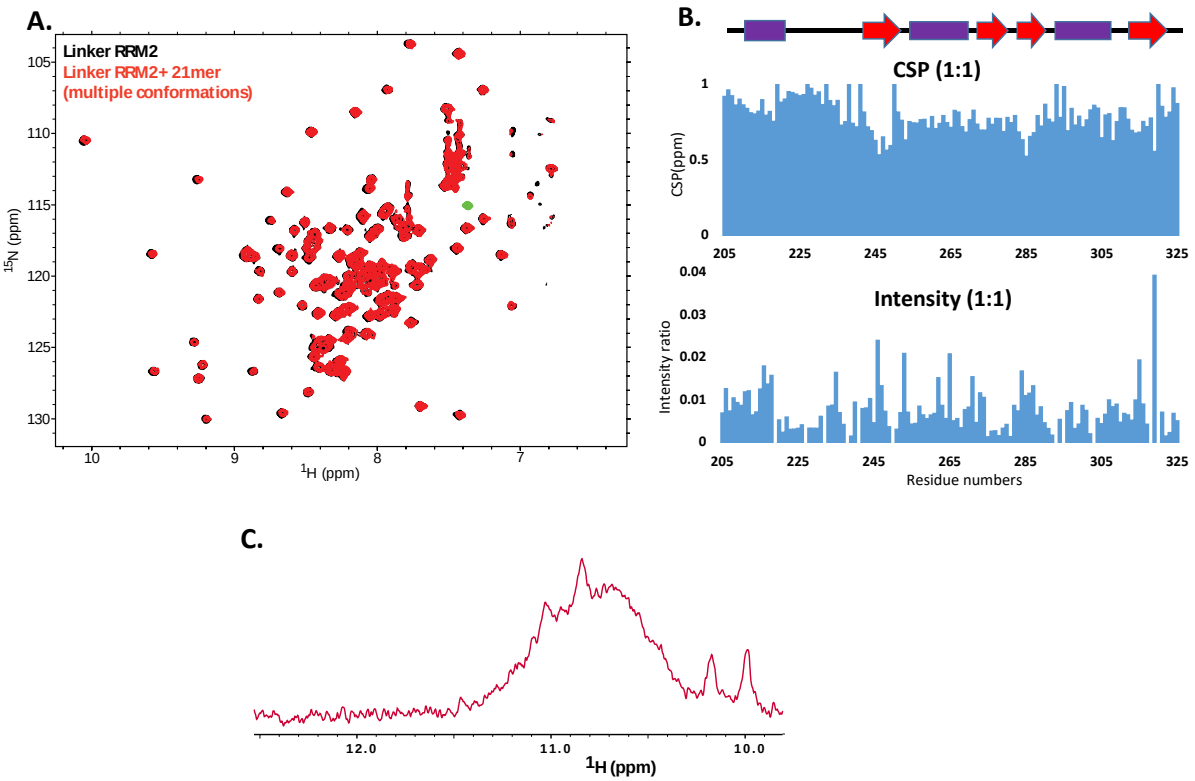



**Figure 3.45 Linker-RRM2 does not bind 21mer with multiple conformations. A.** $^{1}H$-$^{15}N$ spectra of RRM2 in the absence of RNA (black) and in the presence 2 molar excess (red) of 21mer RNA in multiple conformations recorded at 500 MHz and 298 K in the presence of 20 mM potassium phosphate pH 6.5, 50 mM NaCl and 5 mM DTT. The overlay shows that upon addition of 21mer with multiple conformations, RRM2 signals does not show significant line broadening. **B.** Plots of CSP and intensity with respect to residue numbers shows that only minor CSPs and intensity loss of RRM2 RNA-binding residues were observed, which could be attributed to the presence of minor conformation of 21mer uniform conformation, thus indicating that linker-RRM2 rather prefers uniform conformation over multiple conformations of 21mer RNA. **C.** $^{1}H$ NMR spectrum shows that 21mer RNA used for titration has majorly multiple conformations.

To have more direct evidence that protein indeed interacts with the uniform conformation of 21mer, the resulting mixture or protein-RNA (RNA in multiple conformations) in 7 mM Potassium phosphate pH 6.5 was heated to 95°C and slow cooled at RT, allowing RNA to unfold first and then fold into uniform conformation. The recorded HSQC showed the line broadening of most of the peaks, confirming the protein-RNA interaction. Thus, this proves that protein interacts with uniform G-quadruplex form and not with the multiple conformations (Figure 3.46).

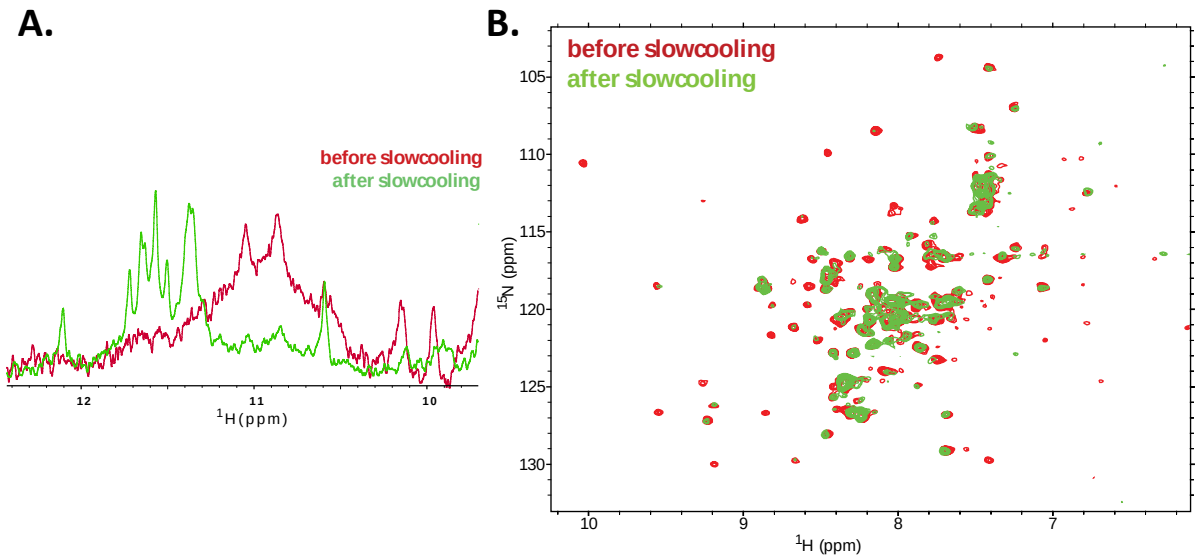



**Figure 3.46 Linker-RRM2 shows specificity for 21mer in uniform conformation. A.** $^{1}$H NMR spectra showing 21mer RNA with multiple conformations before slow cooling (red) and uniform conformation after slow cooling (green). **B.** $^{1}$H-$^{15}$N spectra of linker-RRM2 in the presence of 21mer with multiple conformations i.e. before slow cooling (red) and in the presence of uniform conformation i.e. after slow cooling (green). Line broadening of linker-RRM2 signals in presence of uniform conformation shows that linker-RRM2 specifically interacts with 21mer uniform conformation.

### 3.1.3.8 G-quadruplex-specific inhibitor disrupts linker-RRM2-21mer complex

Interaction of linker-RRM2 with G-quadruplex often resulted in the line broadening of the signals eventually disappearing into noise level. It is hard to distinguish whether protein recognizes the G-quadruplex structure and forms a stable complex or rather it prefers to single-stranded poly G nucleotides and hence binds G-quadruplex, eventually unfolding it to single stranded form and forms stable complex with it.

To get more insights into this interaction, we employed G-quadruplex inhibitor, called as TMPyP4. It is a cationic porphyrin ring, known to stabilize G-quadruplex DNA structure, at the same time to inhibit the RNA G-quadruplex formation. TMPy4 is known to be specific towards G-quadruplex structures and does not interact with duplex or single-stranded RNA. It could be speculated that if protein unfolds G-quadruplex RNA and forms a stable complex with single-stranded RNA, the addition of TMPyP4 should not have any effect on the complex. But, on the other hand, if protein forms complex with G-quadruplex RNA, TMPyP4 should be able to disrupt the G-quadruplex structure and thereby disrupting the complex.

To test the effect of TMPyP4, the 1D $^{1}$H NMR spectra was recorded on 21mer G-quadruplex in the absence and the presence of TMPyP4. In the presence of TMPyP4, the G-quadruplex-specific imino resonances, disappeared confirming that TMPyP4 inhibits G-quadruplex formation.

To check the effect of TMPyP4 on linker-RRM2-21mer complex (RNA in excess), first HSQC of protein-complex was recorded, characterized by line broadening of most of the signals.

Similarly, $^{1}$H 1D NMR spectrum was recorded. Upon successive addition of TMPyP4, 1D $^{1}$H NMR as well as $^{1}$H-$^{15}$N HSQC spectra were recorded to monitor the effect of TMPyP4.

$^{1}$H 1D spectra confirmed the disruption of G-quadruplex structure in the presence of 9x TMPyP4. Similarly, the resonances specific for free linker-RRM2 started appearing upon successive addition of TMPyP4. To allow complete disruption of the complex by TMPyP4, the sample was incubated overnight at RT. The subsequent $^{1}$H-$^{15}$N HSQC showed sharp linker-RRM2 signals, specific for free linker-RRM2, indicating disruption of the complex. This confirms that linker-RRM2 binds specifically to the G-quadruplex structure as opposed to single-stranded RNA.

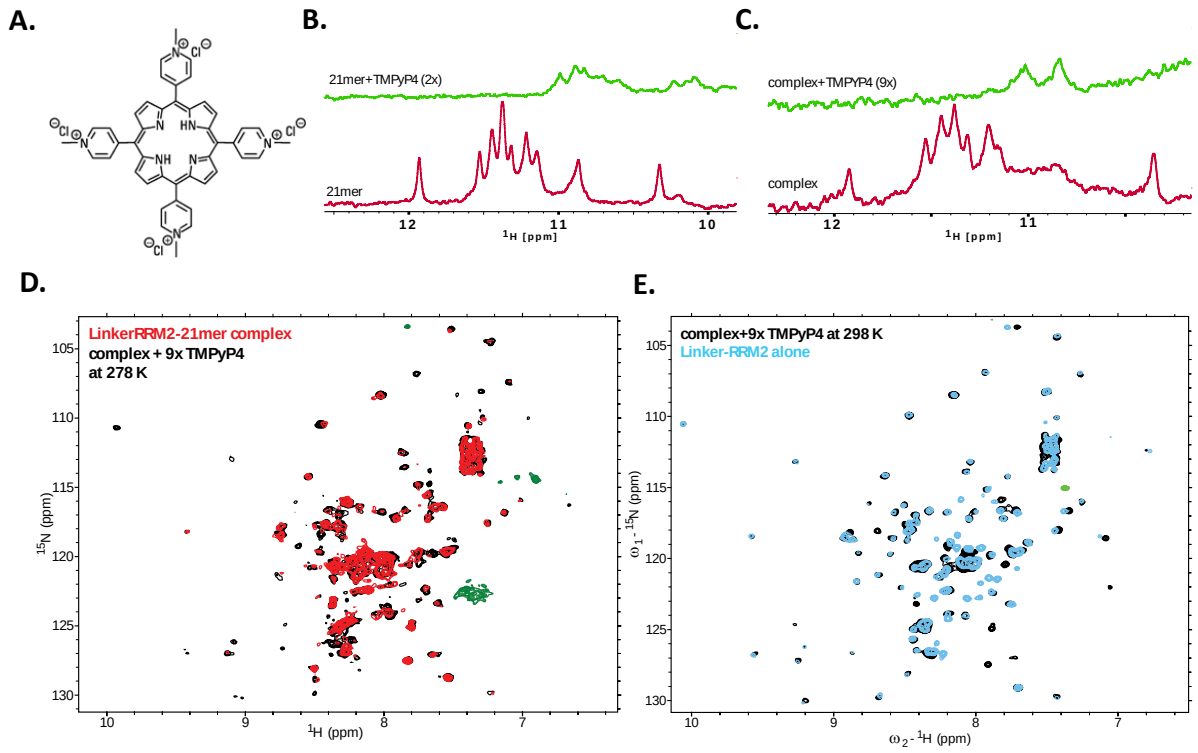



**Figure 3.47 G-quadruplex-specific inhibitor disrupts linker-RRM2-21mer complex. A.** Structure of G-quadruplex-specific inhibitor, TMPyP4**, [**adapted from (Yaku, Murashima et al. 2013)] **B.** $^{1}$H NMR spectra showing 21mer G-quadruplex imino signals are gone in presence of 2 molar excess of TMPyP4 **C.** $^{1}$H NMR spectra showing 21mer G-quadruplex imino signals from linker-RRM2-21mer complex in the absence of TMPyP4 (red) and in the presence of 9 molar excess of TMPyP4 (green), showing that in the complex G-quadruplex structure is disrupted in presence of TMPyP4. **D.** $^{1}$H-$^{15}$N spectra of linker-RRM2 in complex with 21mer G-quadruplex in the absence of TMPyP4 (red) and in the presence of 9 molar excess of TMPyP4 (black). Upon addition of TMPyP4, line broadened protein resonances become sharper and **E.** Overlay of $^{1}$H-$^{15}$N HSQC spectra of Linker-RRM2-TMPyP4, after overnight incubation at RT, with free Linker-RRM2 showing that TMPyP4, disrupts protein-RNA complex.

## 3.1.3.9 Mutational analysis to abolish RNA binding contribution

To check in vivo that RRM2 is indeed the crucial in determining specificity in identifying LS2 target RNA, mutations were carried out on each domain respectively, to abolish the RNA

binding activity of each domain. In vivo splicing assays will be performed with full-length LS2 proteins carrying mutations in each domain.

To abolish the RNA binding by RRM1, three double mutants were generated namely, RRM1,2$^{F164D,F166D}$; RRM1,2$^{R112D,Y114D}$; RRM1,2$^{K192E\ R194E}$, which involved residues located in RNP1, RNP2 and a β-5 strand of RRM1 respectively. β-5 strand and RNP2 sites residues were chosen because they show changes during NMR titration whereas, although RNP1 site doesn't show significant changes during RNA titration by NMR, it was chosen to mutant, to see if it is really dispensable.

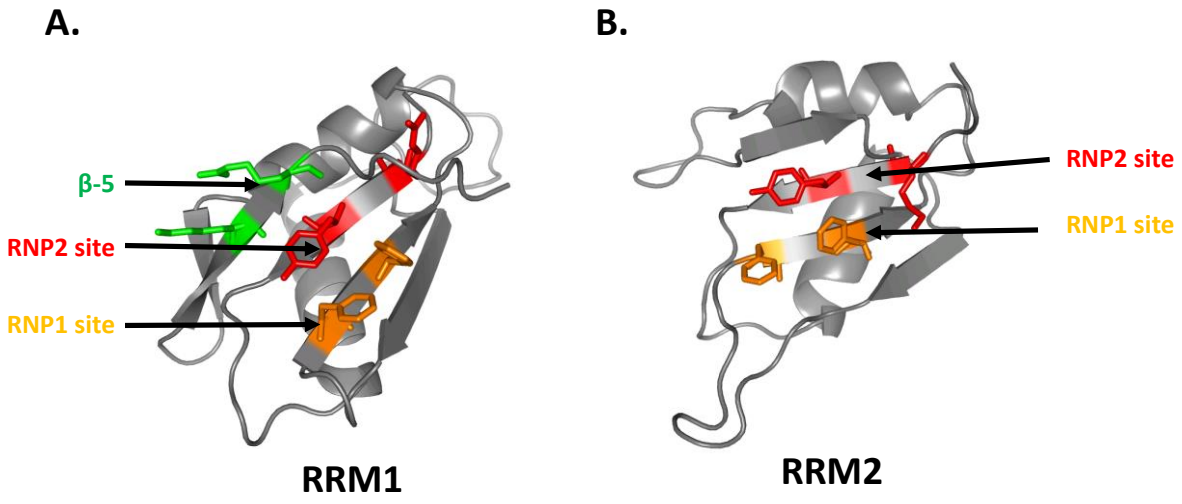



**Figure 3.48 Designed LS2 mutants for abolishing RNA binding.** Residues selected for mutagenesis of **A.** RRM1 and **B.** RRM2, based on NMR RNA titrations.

Similarly, to abolish RNA binding by RRM2, double mutants such as RRM1,2$^{F285D,F287D}$; RRM1,2$^{K243E,Y245D}$ , were generated. These residues were chosen because of their location in the RNP1 and RNP2 sites of RRM2 respectively and show changes upon RNA titration.

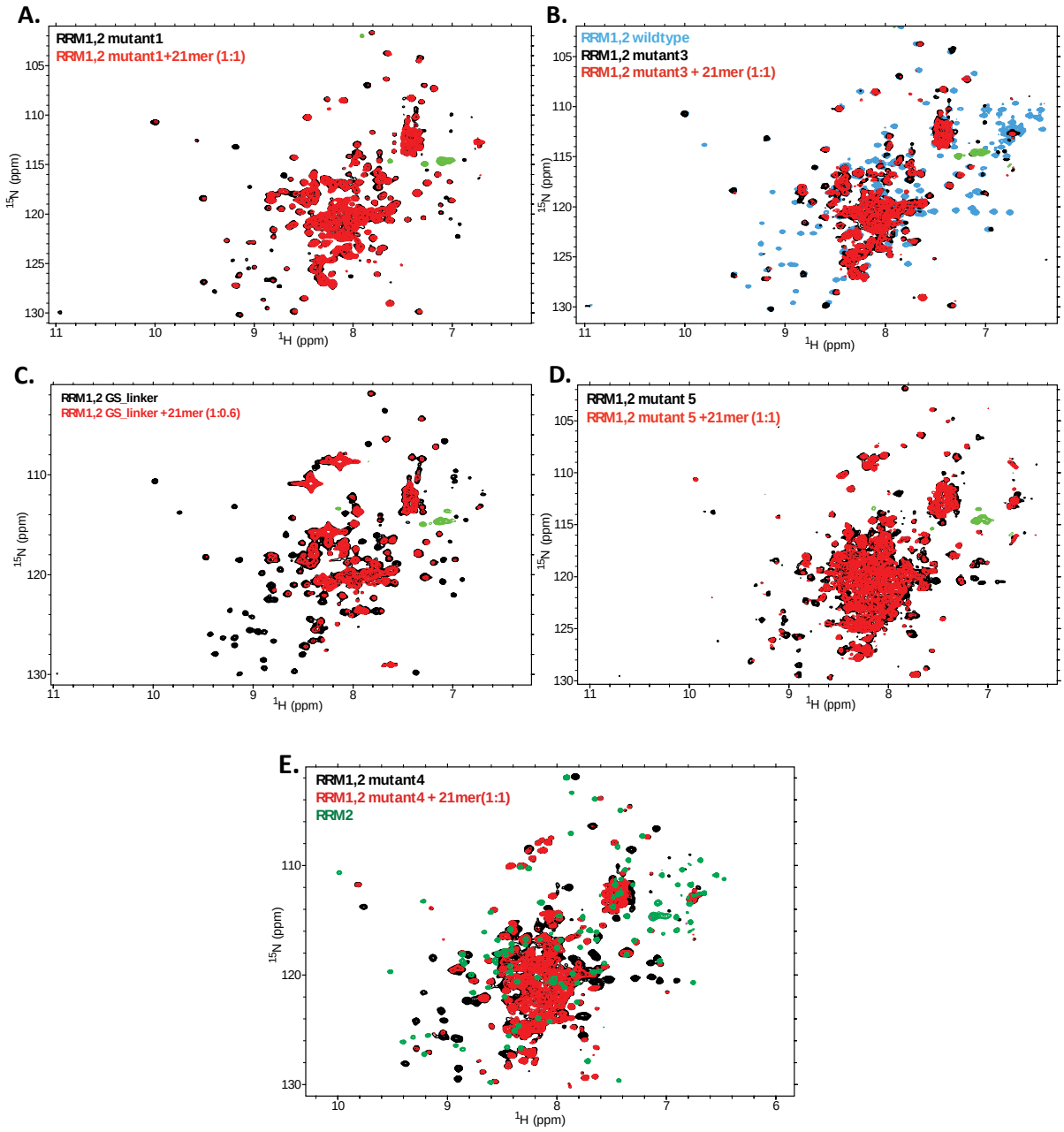



**Figure 3.49 Titration of LS2 mutants with 21mer RNA.** $^{1}H$-$^{15}N$ HSQC overlay of LS2 RRM1,2 mutants in absence and presence of 21mer RNA. All the experiments were performed in 10 mM potassium phosphate pH 6.5, 25 mM NaCl, 5 mM DTT at 500 MHz spectrometer (except for **C.** at 800 MHz).

On the other hand, to abolish the RNA binding contribution from the transiently forming helix of the linker, H197-P233 residues of the linker were substituted with GGS residues which are known to form a flexible loop, thus RRM1,2 GGS linker construct was generated.

All these mutants showed aggregation-prone behavior as observed for RRM1,2. To find out the effect of each mutation in terms of RNA binding, NMR titrations were carried out. The summary of the findings is listed in Table 3-2.

**Table 3-2 Summary of LS2 mutants titration with 21mer.**

| Name | Mutations | Reason for mutation (with respect to wildtype RNA titration) | NMR RNA titration |
|---|---|---|---|
| **Mutant 1** (RRM1,2) | F164D, F166D | RNP1, RRM1 (RNP site, but no effect during RNA titration) | RNA binding but no precipitation |
| **Mutant 2** (RRM1,2) | R112D, Y114D | RNP2, RRM1 (RNP site, changes during RNA titration) | Not tested |
| **Mutant 3** (RRM1,2) | K192E, R194E | Beta-5 strand, RRM1 (strong changes during RNA titration) | No RNA binding from RRM1, as it is unfolded |
| **Mutant 4** (RRM1,2) | F285D, F287D | RNP1, RRM2 (RNP site, changes during RNA titration) | No RNA binding from RRM2 |
| **Mutant 5** (RRM1,2) | K243E, Y245D | RNP2, RRM2 (RNP site, changes during RNA titration) | RNA binding with precipitation (same as wild type) |
| **GS linker_mutant** (RRM1,2) | H197-P233 replaced by GS linker | Helical propensity residues from linker (changes during RNA titration) | Protein is folded, RNA binding by both RRMs |

In the case of RRM1, mutant 1 and mutant 3 were purified successfully and checked for RNA binding. Mutant 1, which was targeted to the canonical RNP1 site, does not affect RNA binding contribution from RRM1 (Figure 3.49, A). This is consistent with the wild-type protein, in which RNP1 site does not show any effect during RNA titration. Whereas, mutant 3, which was target towards β-5 resulted in the unfolding of RRM1 domain (Figure 3.49, B). This mutation still maintains RRM2 in folded form and thereby, still shows RNA binding contribution from RRM2 while ruling out the RRM1 contribution and thus, could be used for checking the significance of RRM1 during in vivo splicing assays.

On the other hand, mutant 4, which targeted RNP1 site of RRM2 ruled out RNA binding activity of RRM2 while maintaining RNA binding contribution from RRM1, as seen by NMR RNA titration. This indicates that mutant 4 could be used for in vivo splicing assays (Figure 3.49, E). Whereas, mutant 5, which targeted RNP2 sites of RRM2, didn't show reduced RNA binding activity (Figure 3.49, D).

Similarly, RRM1,2 GS linker construct was purified and tested. The $^{1}H$-$^{15}N$ HSQC spectrum shows that both RRM domains exist in folded form. Both RRM domains could also bind RNA, as seen by RNA NMR titration. Thus, indicating that this mutant could be used for in vivo splicing assays to check the significance of linker RNA binding contribution (Figure 3.49, C).

# 3.2 Analysis of dU2AF50 RRM domains and their interaction with poly-U RNA

## 3.2.1 NMR analysis dU2AF50 RRM1,2

Multiple sequence alignment was used to delineate the domain boundaries of dU2AF50. Two constructs including both RNA binding domains, namely RRM1,2 (92-286) and exRRM1,2 (85-290) were designed, with boundaries similar to the constructs used for NMR and crystallography studies of U2AF65 RNA binding domains respectively. These constructs were subcloned and protein purification was performed as described in section 2.4.

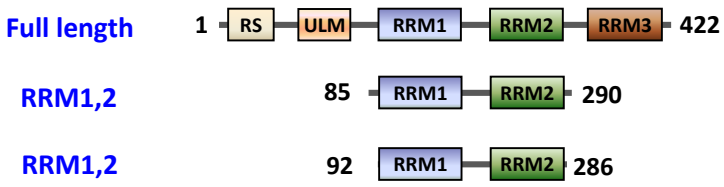



**Figure 3.50 Summary of dU2AF50 constructs used in this study.**

In contrast to LS2 RRM1,2; dU2AF50 RRM1,2 did not show aggregation-prone behavior. The protein was found to be soluble up to 800 µM concentration even in the presence of lower NaCl concentration (50 mM). Further, NMR experiments for performing backbone assignments were recorded on RRM1,2 (92-286) (Table 2-11) and subsequent assignments were performed.

Based on $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ chemical shifts, the secondary structure of RRM1,2 was derived (Figure 3.51). As like LS2, both RRM domains of dU2AF50 adopt common βαββαβ topology which is also consistent with the topology of U2AF65 RRM domains. Similarly, residues comprising interdomain linker of dU2AF50 did not show any significant secondary structure, which is again consistent with linker region of U2AF65. To understand the backbone dynamics of dU2AF50 RRM1,2 relaxation experiments such as $\{^{1}H\}$-$^{15}N$ heteronuclear NOE, T1 and T1rho were recorded. The $^{15}N$ relaxation analysis agrees well with the secondary structure of dU2AF50 RRM domains with the first half of the linker showing flexibility. The other half of the linker (just preceding RRM2) showed a considerable amount of rigidity. The residues of this region show high conservation with LS2 and U2AF65 and the rigidity displayed by them might be the outcome of linker-RRM2 interaction as observed in the case of LS2 (section 3.1.1.6) as well as U2AF65 (unpublished data).

Rotational correlation time values were calculated for dU2AF50 RRM1,2 based on T1 and T1rho relaxation values. The calculated rotational correlation time values were found to be ~11.14 ns which agree well with the theoretical value expected for a monomeric protein with given number of residues. Both RRM1 and RRM2 showed similar correlation times, which

indicates that both domains tumble together as one unit, which is consistent with U2AF65 RRM1,2.

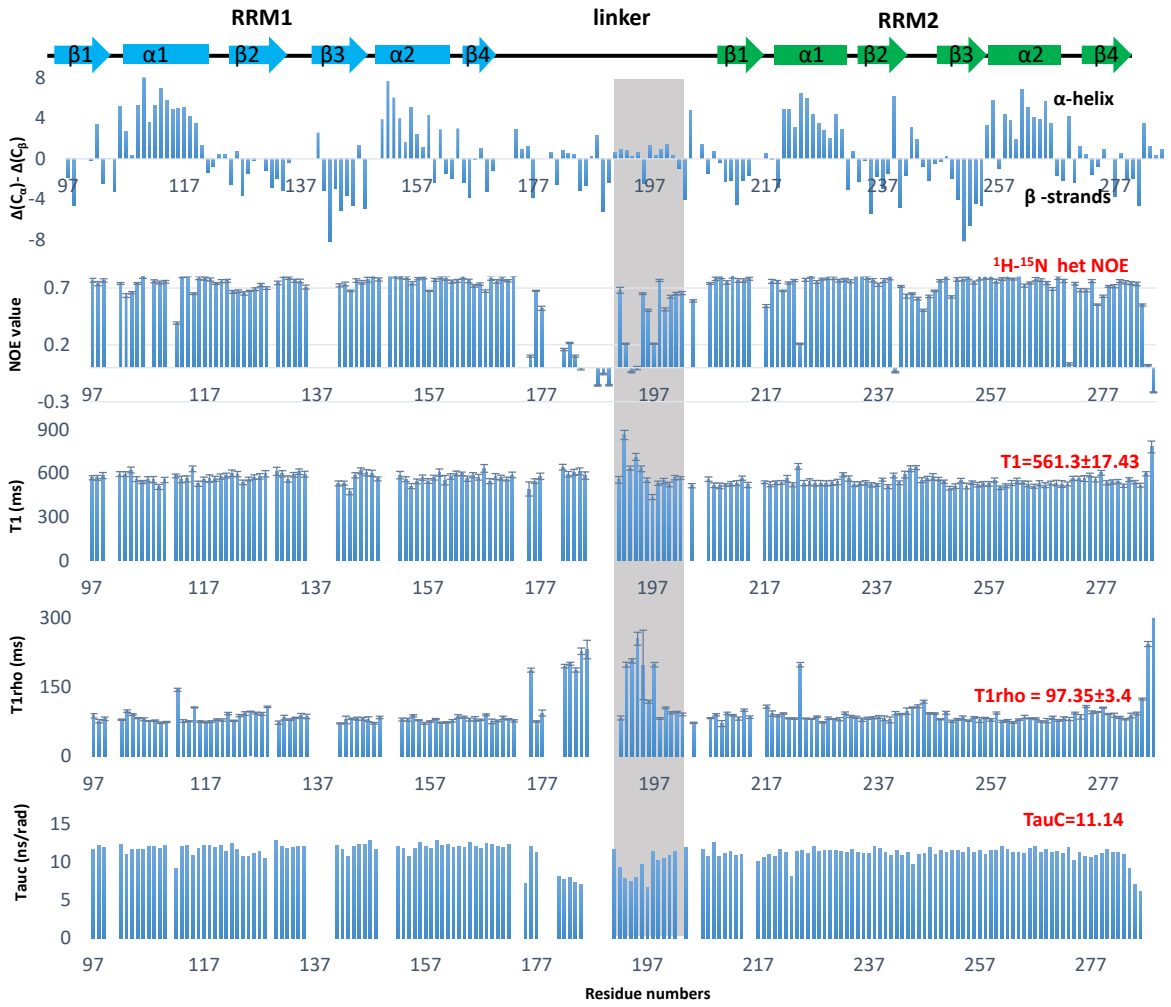



**Figure 3.51 Chemical shift derived secondary structure and backbone dynamics of dU2AF50 RRM1,2.** Chemical shift derived secondary structure prediction shows that both RRM domains of dU2AF50 have canonical βαββαβ topology (panel 1). Backbone dynamics data measured using {$^{1}$H}-$^{15}$N heteronuclear NOE, T1, and T1rho relaxation is consistent with the secondary structures of both RRM domains (panel 2,3,4). Correlation time (τc) shows that RRM1,2 exist in monomeric form and both RRM domains tumble together (panel 5). The linker residues showing some degree of rigidity are highlighted by a gray box. Secondary structure elements are indicated on top for each RRM.

## 3.2.2 NMR titration of dU2AF50 RRM1,2 with U9

To study the interaction of dU2AF50 with Py tract RNA, NMR titration of RRM1,2 (92-286) with U9 RNA was performed. Increasing concentration of RNA was added into $^{15}N$ labeled protein and series of $^{1}H$-$^{15}N$ HSQCs were recorded to monitor changes upon RNA addition.

As shown in Figure 3.52, in presence of RNA, protein resonances show severe line broadening as well as chemical shift perturbations (CSPs), indicating protein-RNA interaction which is in intermediate-fast exchange regime.

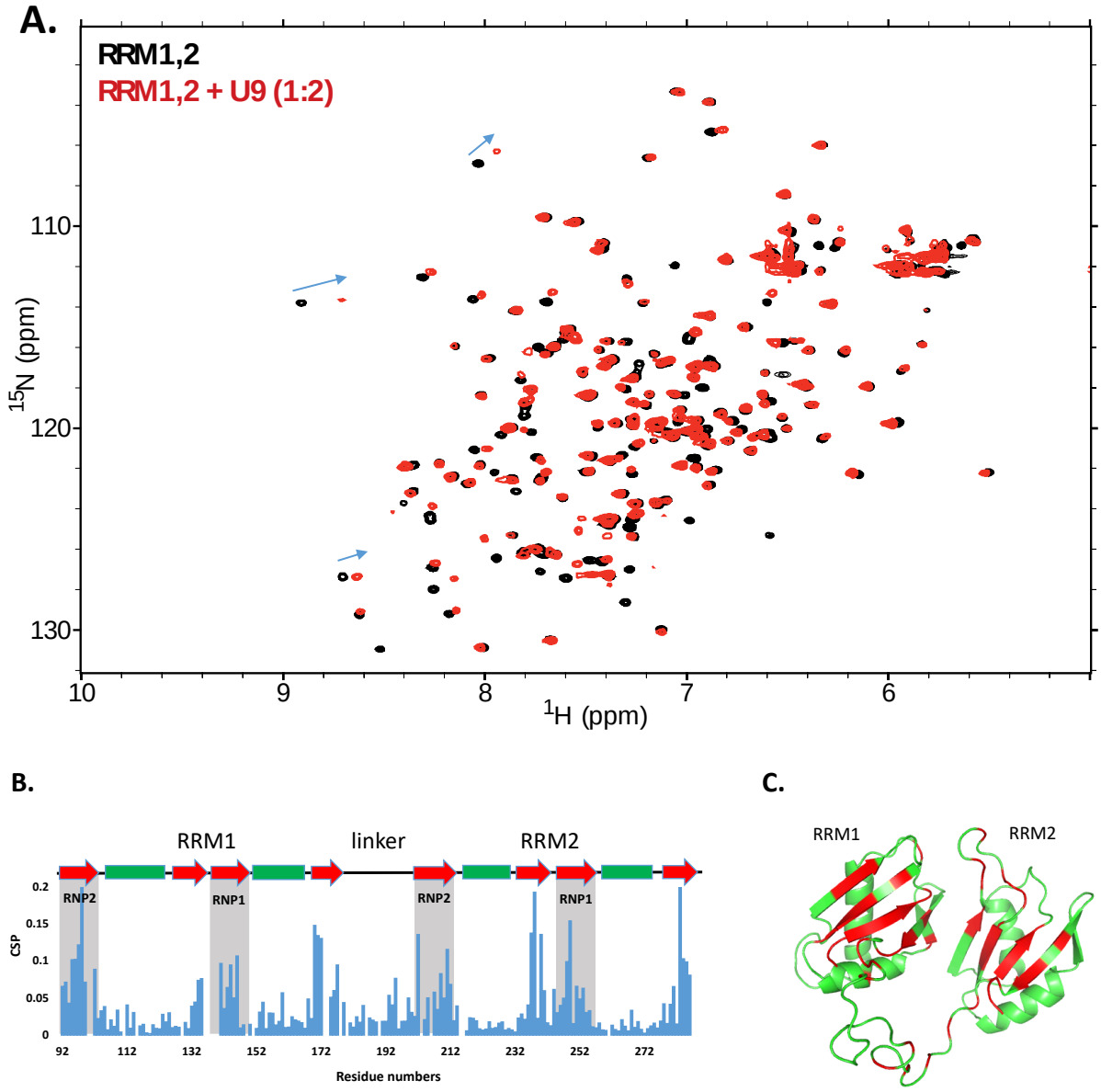



**Figure 3.52 NMR titration of RRM1,2 with U9 RNA. A.** $^{1}H$-$^{15}N$ spectra of dU2AF50 RRM1,2 in the absence of RNA (black) and in the presence 2 molar excess of 21mer RNA. The overlay shows that in presence of RNA, dU2AF50 RRM1,2 resonances undergo line broadening and chemical shift changes. The experiment was performed at 800 MHz and 298 K in the presence of 20 mM potassium phosphate pH 6.5, 50 mM NaCl and 5 mM DTT. **B.** The plot

of chemical shift perturbations with respect to residue numbers show that β-strands of both RRM domains are involved in RNA binding. **C.** Residues undergoing **c**hemical shift perturbations are mapped in red on the homology model of dU2AF50 RRM1,2 (generated using SWISS-MODELLER using U2AF65 RRM1,2 RNA bound structure as a template).

The plot of chemical shift perturbation with respect to residue numbers shows that both RRM domains are involved in RNA binding. Chemical shift changes are majorly located on β-strands of both RRM domains, which also includes RNP2 and RNP1 sites (β1 and β3 strands respectively). This indicates that canonical mode of RNA binding is adopted by both RRM domains of dU2AF50. The saturation of protein-RNA interaction is reached in presence of ~1.2-fold excess of RNA concentration, indicating that each molecule of dU2AF50 RRM1,2 binds one molecule of U9 RNA.

The overall pattern of chemical shift perturbation indicates that dU2AF50 RRM1,2 interaction with U9 is reminiscent of U2AF65 RRM1,2 and U9 interaction (Mackereth, Madl et al. 2011).

## 3.2.3 ITC study of dU2AF50 RRM1,2 and U9 interaction

Isothermal titration calorimetry (ITC) experiments were performed to get insights about thermodynamic parameters of protein-RNA interaction (Figure 3.53).

Titration of dU2AF50 exRRM1,2 with U9 gave a dissociation constant ($K_d$) of 0.59 ± 0.08 µM. This $K_d$ value is in agreement with the NMR titration results showing intermediate-fast exchange regime. The N value for the interaction was obtained to be around 1.3, which indicates that possibly the binding stoichiometry is 1:1. This binding stoichiometry is also consistent with the NMR titration data in which saturation of protein binding sites is achieved with ~ 1.2-fold excess of RNA molecules. Thus, thermodynamic parameters obtained from ITC analysis of dU2AF50 exRRM1,2 with U9 are in agreement with that of U2AF65 exRRM1,2 and U9 interaction (unpublished data).

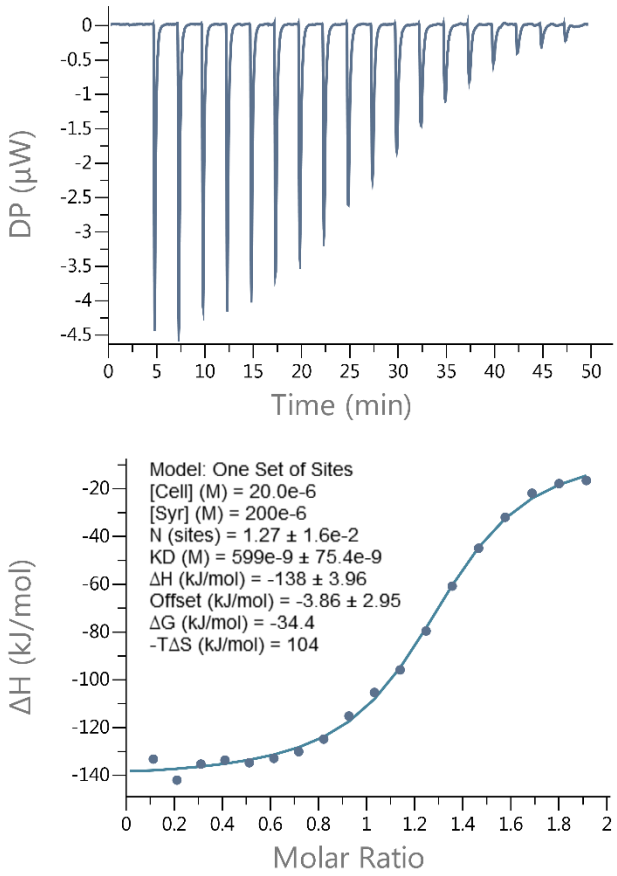



**Figure 3.53 ITC analysis of RRM1,2 and U9 RNA interaction.** The measurement was carried out at 25 °C in 20 mM potassium phosphate buffer pH 6.5, 50 mM NaCl, 5 mM DTT. For titration, RNA was taken in cell whereas protein was in the syringe. Upper panel indicates raw data, which was fitted to one set of site binding model. Thermodynamic parameters obtained from the titration are listed on the left side of the lower panel.

# Chapter 4 Discussion

## 4.1 Conserved topology of LS2 and dU2A50 RRM domains

RNA recognition motif (RRM) is most commonly found RNA binding domain in higher vertebrates (Venter, Adams et al. 2001). RRM is involved in a wide array of posttranscriptional gene expression processes. This domain displays remarkable ability to retain its canonical topology yet displaying versatile nature of single-stranded nucleic acid and protein binding.

Our structural and biophysical data shows that though LS2 and dU2AF50 RRM domains are structurally very similar, despite their functional differences. The solution NMR structure of LS2 RRM2 (Figure 3.11) and CS-ROSETTA structure of LS2 RRM1 (Figure 3.10), show that both LS2 RRM domains adopt canonical βαββαβ fold (Figure 4.1). Moreover, both RRM domains show the presence of conserved RNP1 and RNP2 motifs at respective β-3 and β-1 strands, which provides canonical RNA binding interface. Consistent with their high sequence identity and molecular function with U2AF65 RRM domains, dU2AF50 RRM domains also display the similar features as revealed by NMR data(Figure 3.52, Figure 3.53).

Recent studies show that RRM domains also display several structural variations in order to perform novel protein-RNA or protein-protein interactions. Most commonly found variations are the additional N-terminal and C-terminal secondary structural elements. For example, the additional helix in the C-terminal of the RRM is reported for La C-terminal RRM, which rests on RNA binding surface formed by β-sheet, thus preventing the canonical interaction with RNA nucleotides. (Jacks, Babon et al. 2003) (Figure 4.1, F). Similarly, in the case of PTB RRM2 and 3, the C-terminus forms an extra β5-strand, which extends the RNA-binding surface (Oberstrass, Auweter et al. 2005) (Figure 4.1, E). On the other hand, in the case of RRM4 of Prp24, both N- and C-terminal residues shows the presence of helix, which occludes the β-sheet surface in free form (Martin-Tumasz, Richie et al. 2011) (Figure 4.1, G). In addition, the length of loops in between the secondary structural elements or the secondary structural elements itself may vary as observed in the case of FUS RRM (Liu, Niu et al. 2013) (Figure 4.1, B) and α-1 helix of the U2AF35 RRM (Kielkopf, Rodionova et al. 2001) (Figure 4.1, H) respectively.

It is important to note that any such important structural difference is not observed in the case of LS2 RRM domains. Interestingly, the loop between α1-helix and a β-2 strand (loop 2) of LS2 RRM1 is 5 residues longer than $U2AF^{LS}$ paralogs and also contains three potential RNA-binding residues, which are similar to FUS RRM (Figure 4.1, A and B respectively). But, backbone dynamics data reveal that this loop is completely flexible (Figure 3.9) and subsequent RNA titration studies show that it is not involved in RNA binding (Figure 3.36). On the other hand, the novel transient helix formed by the LS2 specific residues at the boundary of LS2 RRM1 and the linker does not appear to be a part of RRM1 (Figure 3.16). Although the conserved topology of LS2 RRM domains with respect to $U2AF^{LS}$ RRM domains is a bit surprising, but nevertheless, it is in agreement with previously reported structural and

functional studies of other RRM domains, which are known to retain canonical topology yet display differential RNA binding properties (Clery, Blatter et al. 2008).

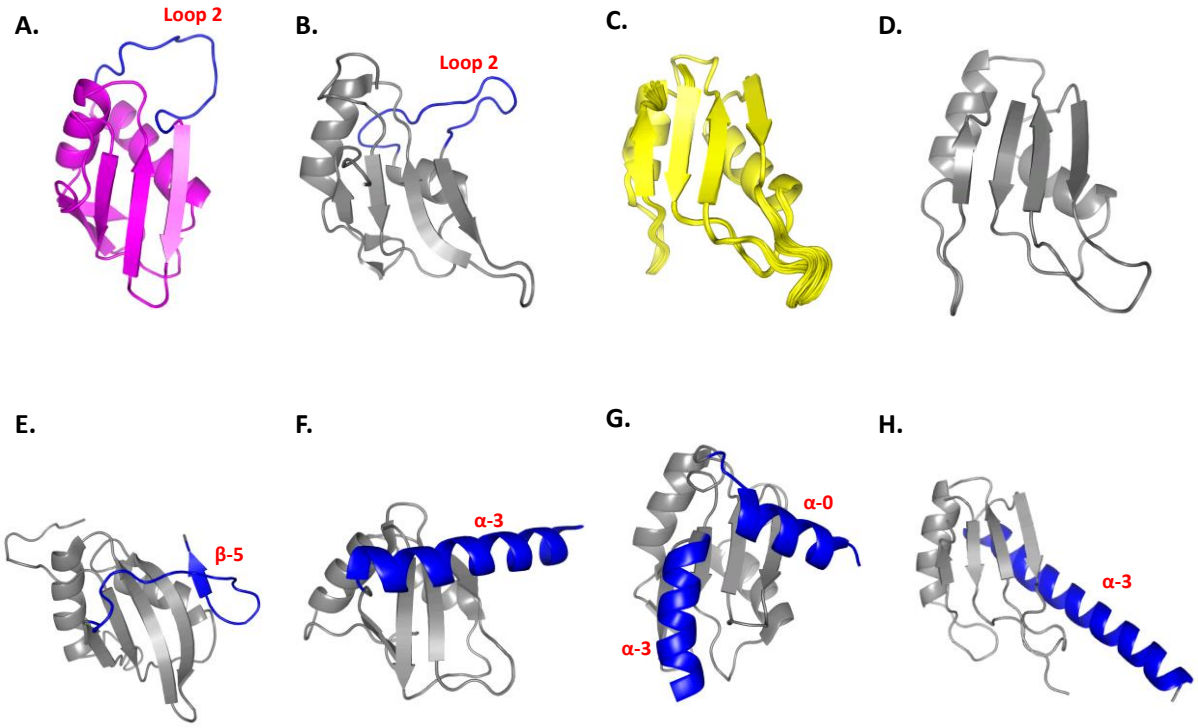



**Figure 4.1 Comparison of LS2 RRM domains with reported RRM types.** Schematic representation of RRM domains with additional elements colored in blue**. A**. LS2 RRM1 CS-ROSETTA structure, **B.** FUS RRM (Liu, Niu et al. 2013), **C.** LS2 RRM2 NMR structure, **D.** U2AF65 RRM2 structure as an example of canonical RRM fold, **E.** PTB RRM3 (Oberstrass, Auweter et al. 2005), **F.** La C-terminal RRM (Jacks, Babon et al. 2003), **G.** Prp24 oRRM4 (Martin-Tumasz, Richie et al. 2011), **H.** U2AF35 RRM (Kielkopf, Rodionova et al. 2001).

On the other hand, the interdomain conformational dynamics can also play an important role in achieving target specificity of L2, as reported for U2AF65 RNA binding domains (Mackereth, Madl et al. 2011). But, poor solubility of LS2 RRM1,2 poses difficulties in deducing this information.

## 4.2 Features of LS2 RRM1,2 interdomain linker

Generally, the interdomain linker between two domains is not well studied. In the case of LS2, the linker appears to influence RNA binding. This study identifies two important regions in the linker, which are located at the boundaries of RRM1-linker and RRM2-linker respectively. These regions are found to be highly conserved in the LS2 protein homologs found in the *Drosophila* species (Figure 3.15). Residues of these regions exhibit semi-rigidity as opposed to

the other residues of the linker showing flexibility. The first region is located in the N-terminal of the linker just after RRM1, which includes aromatic, charged residues. NMR data reveal that these residues show a helical propensity in the free form (Figure 3.13). This residual secondary structure might result in making these residues semi-rigid (Figure 3.14). NMR titration shows that some of the residues of this region are involved in RNA binding (Figure 3.44). Such transient helical regions are also reported in the case of other nucleic binding proteins, where it was shown that this transient helix gets ordered upon binding to its target nucleic acid sequences (Maris, Dominguez et al. 2005). Although it would be interesting to check whether the transient helix identified in this study becomes ordered upon RNA binding, line broadening of the residues in the presence RNA makes it difficult to determine whether this is the case. Chemical shifts of the residues in this region remain unchanged when they are part of the linker-RRM2 or RRM1,2 construct.

Interestingly, chemical shift derived secondary structure depicts that residues at the C-terminal end of the LS2 RRM1 (beginning of the linker), shows a α-helical tendency. This secondary structure formation is also supported by some degree of rigidity in the $^{15}N$ relaxation analysis. This region with helical propensity can be potentially identified as helix-C of RRM1 and may be missing the hydrophobic interaction of RRM1 and this helix-C was causing aggregation-prone behavior. But the inclusion of this region also did not improve the aggregation-prone behavior of RRM1 (Figure 3.16).

On the other hand, the second region showing semi-rigidity is located on the C-terminal of the linker just preceding RRM2. Our solution NMR structure reveals that hydrophobic residues located in this region interacts with β2 and β3 strands of RRM2, and thus gains partial rigidity (Figure 3.18). Interestingly, these two β strands also include RNA-binding residues and are involved in RNA-binding (Figure 3.37). The role of these regions of the linker in RNA binding is not clear yet and certainly requires further investigation. It could be possible that linker mediates autoinhibitory interaction by competing with weak RNA ligands for β-strands as observed for U2AF65 (unpublished data). On the other hand, this interaction may be required to bring N-terminal RNA-binding residues from the linker close to RRM2, in order to extend the RNA binding interface.

Apart from this, LS2 linker also contains five serine residues, which are highly conserved among the homologs from other Drosophila species. These residues are not found in the $U2AF^{LS}$. It is important to analyze if these residues also serve as phosphorylation sites. The possible Gene Ontology terms of possible functions of LS2 reveals that LS2 might play role gene regulation via phosphorylation (Taliaferro, Alvarez et al. 2011). Phosphorylation could also be involved in the regulation of aggregation, as observed in the case of elongation initiation factor 2α (Wolozin 2012).

## 4.3 Significance of the G-quadruplex formation by LS2 target RNA

*Cis*-regulatory elements can play a more active role in the regulation of splicing rather than just serving a site to recruit *trans*-acting RNA binding proteins. These elements could adopt secondary structures such as duplexes, hairpins, bulges, pseudoknots etc. and can regulate the splicing on its own (without the need of splicing factors). On the other hand, recruitment of splicing factors can be regulated by the equilibrium between the single-stranded or structured form of these elements (Buratti and Baralle 2004).

Guanosine-rich RNA is known to adopt G-quadruplex structures. Formation of G-quadruplex by *cis*-regulatory elements is also reported to have an impact on splicing regulation (Gomez, Lemarteleur et al. 2004). In recent years, many G-quadruplex forming motifs are identified near regulatory sites. Similarly, growing number of proteins are being identified which either interact with G-quadruplex structure, unwind or rather stabilize them (Phan, Kuryavyi et al. 2011). For example, Target RNA sequences of hnRNP F RRM domains adopt G-quadruplex topology. But, the detailed biophysical analysis of hnRNP F and target RNA interaction indicates that protein binds specifically to single-stranded rather than a G-quadruplex form of RNA. This means that formation of protein- single-stranded RNA complex is controlled by the rate of formation of G-quadruplex and can have an impact on mechanism controlled by hnRNP F protein (Samatanga, Dominguez et al. 2013). On the other hand, G-quadruplex formation is reported to enhance the splicing efficiency of the PAX9 intron1 (Ribeiro, Teixeira et al. 2015). Together, it suggests that G-quadruplex formation can either prevent or allow the recruitment of the protein factor at the particular regulatory site and can thereby influence mechanism of post-transcriptional regulation.

On the other hand, a recently published study suggests that in contrast to the presence of numerous endogenous regions with a propensity to form G-quadruplex structures as well as a high abundance of $K^+$ in the cell, G-quadruplex structures are globally unfolded in eukaryotic cells (Guo and Bartel 2016). The study suspects the presence of robust molecular machinery which maintains thousands of G-quadruplex forming regions in the unfolded state. The study employed dimethyl sulfate, which is known to methylate N7 position of guanosine in single-stranded form. The subsequent quantification of the stalling of reverse transcriptase identifies that G-quadruplex structures are in the unfolded state in the mammalian cells. Although this method highlights that G-quadruplex structures may not be prevalently represented by RNA molecules inside the cell, it does not take into account the fact that biomolecules display conformational equilibria. It is possible that the DMS modification shifts the equilibrium of G-quadruplex formation towards single-stranded form. The G-quadruplex formation can be dependent on the cell cycle, as revealed by DNA G-quadruplex structures in Human cells (Biffi, Tannahill et al. 2013), as well as cell types and states or subcellular compartments (Subramanian, Rage et al. 2011). The study performed by (Guo and Bartel 2016), does not rule out the possibility that the relative occurrence of of G-quadruplex forming regions in the untranslated regions might be folded and involved in the regulatory

functions. Moreover, authors also speculate that the use of steady-state measurements might not be able to detect the transient G-quadruplex formation in the G-quadruplex-forming region.

Our data shows that LS2 target RNA adopts G-quadruplex conformations in vitro. All the oligonucleotides that were designed according to the SELEX sequence (Figure 3.24), adopted G-quadruplex topologies. 8mer and 21mer adopted homogeneous population of G-quadruplex species at lower KCl concentration is which shifted to heterogeneous conformations with an increase in the salt concentration (Figure 3.26, Figure 3.28). On the other hand, 14mer just adopted heterogeneous G-quadruplex conformations (Figure 3.28). Further characterization of 21mer RNA with NMR, CD, SLS reveals that 21mer adopt dimeric, parallel, three planar G-quadruplex topology at lower KCl concentration.

Overall, these results provide evidence that G-quadruplex structures are formed by LS2 target RNA in vitro. The effect G-quadruplex formation on the regulation of alternative splicing by LS2 is being validated by performing in vivo studies.

## 4.4 Interaction between LS2 RRM domains and G-quadruplex RNA

Biochemical and structural studies reported on various RRM domains have shown its preference for single-stranded nucleic acid form (Afroz, Cienikova et al. 2015). The RRM domain utilizes its β-sheet surface for RNA recognition, but the mode of binding is not always same. In the case of canonical RRM domains, two aromatic residues located on RNP1 site (β3-strand) and 1 aromatic residue on RNP2 site (β1-strand) are involved in the base stacking interactions with RNA nucleotides. Additional contacts are made by positively charged side chain at RNP1 site.

In the light of G-quadruplex formation by LS2 target RNA sequences, it was tempting to assume that LS2 RRMs interact with a single-stranded form of RNA as reported for hnRNP F RRMs (Samatanga, Dominguez et al. 2013). However, the NMR titrations of single RRM domains, linker-RRM2 as well as RRM1,2 with various guanosine-rich RNA ligands presented, show that LS2 RRM domains interact with a G-quadruplex form of RNA. RRM1 interacts with all the guanosine-rich RNA oligonucleotides used in the study, irrespective of their homogeneous conformation, whereas, RRM2 shows specificity for the 21mer-specific uniform G-quadruplex conformation (Figure 3.40, Figure 3.42). Therefore, this is the first study, which reports novel RRM-G-quadruplex interaction.

It is intriguing that for interaction with G-quadruplex RNA, LS2 RRM2 uses canonical mode of binding, as in presence of RNA, β-1 and β-3 strands (containing RNP2 and RNP1 sites respectively) shows maximum chemical shift changes and line broadening (Figure 3.37). In addition to RNP sites, residues located on other β-strands also show an effect in presence of

RNA. On the other hand, in the case of LS2 RRM1, chemical shift perturbation analysis shows that residues located on the β-5 as well as C-terminal flexible part of LS2 RRM1 show maximum chemical shift changes (Figure 3.36). Chemical shift changes were also observed in the other parts of the protein, except α-2 helix and β-3 site harboring canonical RNP1 residues. This indicates that this RRM-RNA interaction is non-canonical, as RNP1 residues are not involved in RNA binding. The similar mode of binding is also reported for PTB RRM domains and Py tract interaction, in which β-3 strand doesn't participate in the RNA Binding instead there is additional β-5 strand which is involved in RNA binding (Oberstrass, Auweter et al. 2005). But, in contrast, in the case of LS2 RRM1, there is apparently no additional β-strand observed (Figure 3.19).

It was intriguing to see whether, in the presence of the RRM1-RRM2 linker, RRM2 can interact with the 21mer RNA, as this linker weakly interacts with the β2 and β3-strands in RRM2, which are also involved in RNA binding. However, similar to RRM2, linker-RRM2 showed interaction with 21mer, characterized by line broadening of both RNP sites and additional β-strands (Figure 3.44). A region with helical propensity also showed chemical shift perturbations as well as line broadening. It would be important to understand the contribution of the linker residues in RNA binding. It could be speculated that linker-RRM2 has potential to bind RNA more tightly (because of additional RNA binding patch) than RRM2. On the other hand, the interaction of the linker with the β-strands of the RRM2 might provide additional specificity. ITC experiments of RRM2 and linker-RRM2 with various RNA ligands should be performed to gain more insights about this.

NMR titration of linker-RRM2 with 21mer multiple conformations shows that protein does not bind to G-quadruplex RNA in multiple conformations. Whereas, upon a change in RNA conformation from multiple to uniform obtained through slow cooling, the protein-RNA binding was observed (Figure 3.46). This data again highlight that LS2 (in this case linker-RRM2) indeed shows specificity for G-quadruplex uniform conformation adopted by 21mer. Similarly, the addition of G-quadruplex-specific inhibitor TMPyP4, results into disruption of linker-RRM2-21mer complex, indicating that RNA is G-quadruplex form in the RNA and is not unfolded into single-stranded form upon protein binding (Figure 3.47).

Although the dimeric G-quadruplex topology of the 21mer RNA is intriguing, it is possible that the natural LS2 ligands may not necessarily be dimeric. The SELEX pattern obtained for LS2 target RNA sequences rather hints that they may adopt intramolecular G-quadruplex. The physiological relevance of our findings is being probed by performing in vivo splicing assays. For this purpose, the study further identifies possible mutants of LS2 RRM1,2 with retarded RNA-binding contribution from each RRM and linker respectively. These mutants will be used to perform in vivo splicing assays to validate the role of each RNA-binding element. The importance of G-quadruplex during in vivo splicing assays could be analyzed by using TMPyP4 to inhibit G-quadruplex formation as the intake inside the cell can be easily achieved.

Structural insights of RRM-G-quadruplex interaction will be important to understand this novel interaction. But, the aggregation-prone behavior of RRM1 and line-broadening of resonances observed for RRM2 and G-quadruplex interaction, poses difficulties in structural studies by NMR. It is possible that LS2 RRM domains bind to RNA in multiple registers, which

results in the line broadening of the resonances. Given the symmetry of G-quadruplex, it could also be possible that two or more molecules of RRM2 interact with each RNA molecule, which could increase the molecular mass of the complex. As larger complexes tumble more slowly in the solution, resulting in a shortened transverse relaxation time, thereby causing line broadening. Hence, the line broadening observed for both LS2 and 21mer RNA in bound form could be attributed to increased molecular weight. SLS data did not provide any information about the stoichiometry of the RRM2-G-quadruplex complex since the complex separated while passing through the column, indicating that the affinity of the protein-RNA complex may not be strong enough (Figure 3.43). With the lack of information about stoichiometry of the complex formation, as well as the susceptibility of G-quadruplex conformation for salt type and concentration, crystallization trials for the RRM-G-quadruplex complex may not be successful.

Although a number of G-quadruplex structures have been solved so far using X-ray crystallography and NMR, information regarding how proteins interact with G-quadruplex structures remains elusive. Recently, solution NMR structure (Phan, Kuryavyi et al. 2011) and crystal structure (Vasilyev, Polonskaia et al. 2015) are reported for RGG peptide of FMRP protein and RNA duplex-quadruplex junction. Both structures suggest that shape complementarity and hydrogen bonding interactions play a crucial role in the interaction between RGG peptide and duplex-quadruplex junction. The data reveals that two Arginine side chains form crucial hydrogen bonds with the RNA duplex residues, whereas no apparent interaction was detected between peptide and G-quadruplex.

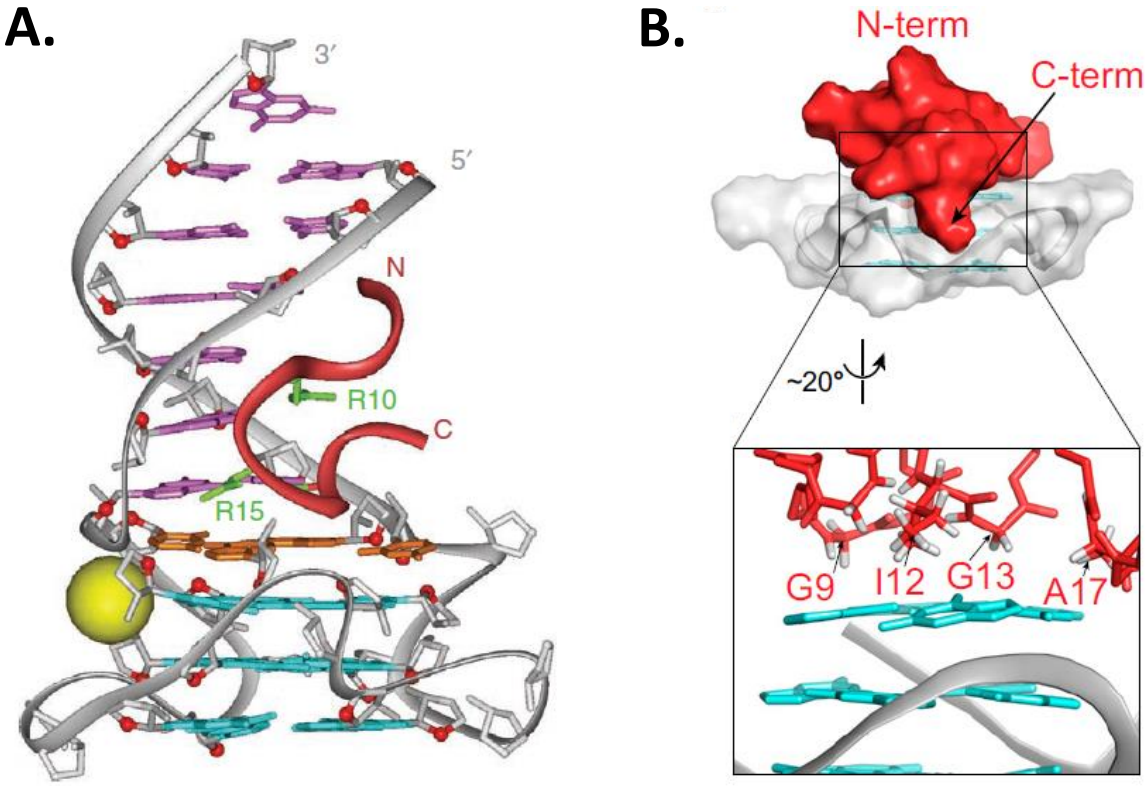



**Figure 4.2 Structures of G-quadruplex bound peptides. A.** A reported solution structure of RGG peptide-sc 1 RNA, showing the interaction between RGG peptide with G-quadruplex and G-quadruplex-duplex junction. The peptide is colored red with important Arginine side chains in green, the duplex, junction and quadruplex regions are represented by magenta, orange, and cyan, respectively whereas sugar-phosphate backbone is in silver color. [Figure adapted from (Phan, Kuryavyi et al. 2011)]. **B.** Side surface representation of solution structure of the Rhau18–T95-2T complex. The intermolecular interactions between peptide residues and DNA G-quadruplex are shown. [Figure adapted from (Heddi, Cheong et al. 2015)].

On the other hand, solution NMR structure of N-terminal RHAU region and G-DNA quadruplex shows that peptide covers the top most G-quartet plane (Heddi, Cheong et al. 2015). Positively charged Lysine and Arginine side chains mediate electrostatic interactions with negatively charged RNA phosphate groups. This interaction is similar to that of most G-quadruplex-specific ligands targeted for G-quadruplex binding.

## 4.5 RNA binding proteins and aggregation

Recent studies report that many RNA binding proteins show aggregation-prone behavior and undergo phase transitions to form hydrogels. These hydrogels are reported to exist as stress granules or RNA granules in the cell (Weber and Brangwynne 2012, Wolozin 2012). These granules form important constituents of ribonucleoprotein (RNP) bodies and exist in the nucleus in the form of Cajal bodies, nucleoli and PML bodies(Mao, Zhang et al. 2011). RNA binding proteins with RRM domains or KH domains are reported to be a major constituent of these RNA granules (Kato, Han et al. 2012). These bodies contain RNA as well as RNA binding

proteins in the sequestered form and carry out many fundamental processes related to the RNA metabolism including splicing (Weber and Brangwynne 2012).

It could be possible that LS2 is part of such granules and aggregation or hydrogel formation (observed in presence of RNA) is necessary for its function. Characterization of LS2 RRM1,2 shows that this construct has aggregation-prone behavior (Figure 3.5). Modulation of domain boundaries, as well as the inclusion of other domains, did not result in improvement in the solubility of this construct. The addition of binding partners such as dU2AF38 UHM domain as well as RNA (21mer) could also not achieve suitable solubility of this construct. Rather in presence of RNA, protein shows gel-like aggregation. LS2 RRM1 is liable for aggregation-prone behavior, as all the LS2 RRM1 containing constructs displayed aggregation-prone properties. It is interesting to note that G-quadruplex species are also known to exhibit aggregation-prone behavior (Davis 2004). But, the aggregated species observed during titrations appears to be caused by protein, which appears to be promoted by RNA (Figure 3.35). Whereas, successive addition of RNA in later stages of titration still shows G-quadruplex-specific signals similar to those observed in the absence of the protein.

It is also possible that the aggregation-prone behavior displayed by LS2 at in vitro may or may not observed in vivo. It should be taken into account that NMR experiments require much higher protein concentration than the endogenous expression of LS2 in cells. Or else, LS2 may require additional interaction with any auxiliary protein factors, which was missing in the in vitro NMR studies, causing protein aggregation.

# Conclusion

This study provides an NMR, biophysical and biochemical characterization of the LS2 protein from *Drosophila melanogaster* and its interaction with guanosine-rich RNAs by using a multi-disciplinary approach. Key findings are 1) LS2 RNA binding domains have a canonical fold. Whereas, RRM1-RRM2 linker is evolved to have a novel RNA binding α-helical region as well as an RRM2 interacting region. 2) LS2 target RNA sequences adopt G-quadruplex conformation in vitro. 3) LS2 RRM domains interact with a G-quadruplex form of the RNA, for which specificity is provided by RRM2 with additional contribution from the linker helical region. Substitution of key RNA binding residues explains the difference in the specificity for LS2 in comparison to dU2AF50. The significance of structural and in vitro interaction data presented in this thesis should be validated by performing in vivo studies.

# References


Abovich, N. and M. Rosbash (1997). "Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals." Cell **89**(3): 403-412.

Adrian, M., B. Heddi and A. T. Phan (2012). "NMR spectroscopy of G-quadruplexes." Methods **57**(1): 11-24.

Afroz, T., Z. Cienikova, A. Clery and F. H. Allain (2015). "One, Two, Three, Four! How Multiple RRMs Read the Genome Sequence." Methods Enzymol **558**: 235-278.

Agarwala, P., S. Pandey and S. Maiti (2015). "The tale of RNA G-quadruplex." Org Biomol Chem **13**(20): 5570-5585.

Agrawal, A. A., E. Salsi, R. Chatrikhi, S. Henderson, J. L. Jenkins, M. R. Green, D. N. Ermolenko and C. L. Kielkopf (2016). "An extended U2AF(65)-RNA-binding domain recognizes the 3' splice site signal." Nat Commun **7**: 10950.

Arnold, K., L. Bordoli, J. Kopp and T. Schwede (2006). "The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling." Bioinformatics **22**(2): 195-201.

Banerjee, H., A. Rahn, W. Davis and R. Singh (2003). "Sex lethal and U2 small nuclear ribonucleoprotein auxiliary factor (U2AF65) recognize polypyrimidine tracts using multiple modes of binding." RNA **9**(1): 88-99.

Banerjee, H., A. Rahn, B. Gawande, S. Guth, J. Valcarcel and R. Singh (2004). "The conserved RNA recognition motif 3 of U2 snRNA auxiliary factor (U2AF 65) is essential in vivo but dispensable for activity in vitro." RNA **10**(2): 240-253.

Barash, Y., J. A. Calarco, W. Gao, Q. Pan, X. Wang, O. Shai, B. J. Blencowe and B. J. Frey (2010). "Deciphering the splicing code." Nature **465**(7294): 53-59.

Biffi, G., M. Di Antonio, D. Tannahill and S. Balasubramanian (2014). "Visualization and selective chemical targeting of RNA G-quadruplex structures in the cytoplasm of human cells." Nat Chem **6**(1): 75-80.

Biffi, G., D. Tannahill, J. McCafferty and S. Balasubramanian (2013). "Quantitative visualization of DNA G-quadruplex structures in human cells." Nat Chem **5**(3): 182-186.

Black, D. L. (2003). "Mechanisms of alternative pre-messenger RNA splicing." Annu Rev Biochem **72**: 291-336.

Bloembergen, N., E. M. Purcell and R. V. Pound (1948). "Relaxation Effects in Nuclear Magnetic Resonance Absorption." Physical Review **73**(7): 679-712.

Buratti, E. and F. E. Baralle (2004). "Influence of RNA secondary structure on the pre-mRNA splicing process." Mol Cell Biol **24**(24): 10505-10514.

Busch, A. and K. J. Hertel (2012). "Evolution of SR protein and hnRNP splicing regulatory factors." Wiley Interdiscip Rev RNA **3**(1): 1-12.

Chen, K. and N. Tjandra (2012). "The use of residual dipolar coupling in studying proteins by NMR." Top Curr Chem **326**: 47-67.

Chen, M. and J. L. Manley (2009). "Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches." Nat Rev Mol Cell Biol **10**(11): 741-754.

Clery, A., M. Blatter and F. H. Allain (2008). "RNA recognition motifs: boring? Not quite." Curr Opin Struct Biol **18**(3): 290-298.

Clore, G. M. and J. Iwahara (2009). "Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes." Chem Rev **109**(9): 4108-4139.

Daubner, G. M., A. Clery and F. H. Allain (2013). "RRM-RNA recognition: NMR or crystallography...and new findings." Curr Opin Struct Biol **23**(1): 100-108.

Davis, J. T. (2004). "G-quartets 40 years later: from 5'-GMP to molecular biology and supramolecular chemistry." Angew Chem Int Ed Engl **43**(6): 668-698.

Delaglio, F., S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer and A. Bax (1995). "NMRPipe: a multidimensional spectral processing system based on UNIX pipes." J Biomol NMR **6**(3): 277-293.
Didiot, M. C., Z. Tian, C. Schaeffer, M. Subramanian, J. L. Mandel and H. Moine (2008). "The G-quartet containing FMRP binding site in FMR1 mRNA is a potent exonic splicing enhancer." Nucleic Acids Res **36**(15): 4902-4912.
Dominguez, C., M. Schubert, O. Duss, S. Ravindranathan and F. H. Allain (2011). "Structure determination and dynamics of protein-RNA complexes by NMR spectroscopy." Prog Nucl Magn Reson Spectrosc **58**(1-2): 1-61.
Farrow, N. A., R. Muhandiram, A. U. Singer, S. M. Pascal, C. M. Kay, G. Gish, S. E. Shoelson, T. Pawson, J. D. Forman-Kay and L. E. Kay (1994). "Backbone dynamics of a free and phosphopeptide-complexed Src homology 2 domain studied by 15N NMR relaxation." Biochemistry **33**(19): 5984-6003.
Fisette, J. F., D. R. Montagna, M. R. Mihailescu and M. S. Wolfe (2012). "A G-rich element forms a G-quadruplex and regulates BACE1 mRNA alternative splicing." J Neurochem **121**(5): 763-773.
Fu, X. D. and M. Ares, Jr. (2014). "Context-dependent control of alternative splicing by RNA-binding proteins." Nat Rev Genet **15**(10): 689-701.
Furtig, B., C. Richter, J. Wohnert and H. Schwalbe (2003). "NMR spectroscopy of RNA." Chembiochem **4**(10): 936-962.
Gobl, C., T. Madl, B. Simon and M. Sattler (2014). "NMR approaches for structural analysis of multidomain proteins and complexes in solution." Prog Nucl Magn Reson Spectrosc **80**: 26-63.
Gomez, D., T. Lemarteleur, L. Lacroix, P. Mailliet, J. L. Mergny and J. F. Riou (2004). "Telomerase downregulation induced by the G-quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing." Nucleic Acids Res **32**(1): 371-379.
Guntert, P. (2004). "Automated NMR structure calculation with CYANA." Methods Mol Biol **278**: 353-378.
Guo, J. U. and D. P. Bartel (2016). "RNA G-quadruplexes are globally unfolded in eukaryotic cells and depleted in bacteria." Science **353**(6306).
Hansel, R., S. Foldynova-Trantirkova, V. Dotsch and L. Trantirek (2013). "Investigation of quadruplex structure under physiological conditions using in-cell NMR." Top Curr Chem **330**: 47-65.
Heddi, B., V. V. Cheong, H. Martadinata and A. T. Phan (2015). "Insights into G-quadruplex specific recognition by the DEAH-box helicase RHAU: Solution structure of a peptide-quadruplex complex." Proc Natl Acad Sci U S A **112**(31): 9608-9613.
Heyd, F. and K. W. Lynch (2011). "Degrade, move, regroup: signaling control of splicing proteins." Trends Biochem Sci **36**(8): 397-404.
Huppert, J. L. and S. Balasubramanian (2005). "Prevalence of quadruplexes in the human genome." Nucleic Acids Res **33**(9): 2908-2916.
J. Sambrook, D. W. R. (2001). "Molecular Cloning: A Laboratory Manual." Cold Spring Harbour Laboratory Press.
Jacks, A., J. Babon, G. Kelly, I. Manolaridis, P. D. Cary, S. Curry and M. R. Conte (2003). "Structure of the C-terminal domain of human La protein reveals a novel RNA recognition motif coupled to a helical nuclear retention element." Structure **11**(7): 833-843.
Jenkins, J. L., A. A. Agrawal, A. Gupta, M. R. Green and C. L. Kielkopf (2013). "U2AF65 adapts to diverse pre-mRNA splice sites through conformational selection of specific and promiscuous RNA recognition motifs." Nucleic Acids Res **41**(6): 3859-3873.
Kanaar, R., S. E. Roche, E. L. Beall, M. R. Green and D. C. Rio (1993). "The conserved pre-mRNA splicing factor U2AF from Drosophila: requirement for viability." Science **262**(5133): 569-573.
Kato, M., T. W. Han, S. Xie, K. Shi, X. Du, L. C. Wu, H. Mirzaei, E. J. Goldsmith, J. Longgood, J. Pei, N. V. Grishin, D. E. Frantz, J. W. Schneider, S. Chen, L. Li, M. R. Sawaya, D. Eisenberg, R. Tycko and S. L. McKnight (2012). "Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels." Cell **149**(4): 753-767.
Keeler, J. (2002). "Understanding NMR Spectroscopy, Second Edition."

Kent, O. A., D. B. Ritchie and A. M. Macmillan (2005). "Characterization of a U2AF-independent commitment complex (E') in the mammalian spliceosome assembly pathway." Mol Cell Biol **25**(1): 233-240.
Kielkopf, C. L., N. A. Rodionova, M. R. Green and S. K. Burley (2001). "A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer." Cell **106**(5): 595-605.
Kikin, O., L. D'Antonio and P. S. Bagga (2006). "QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences." Nucleic Acids Res **34**(Web Server issue): W676-682.
Laemmli, U. K. (1970). "Cleavage of structural proteins during the assembly of the head of bacteriophage T4." Nature **227**(5259): 680-685.
Linge, J. P., M. Habeck, W. Rieping and M. Nilges (2003). "ARIA: automated NOE assignment and NMR structure calculation." Bioinformatics **19**(2): 315-316.
Lipay, J. M. and M. R. Mihailescu (2009). "NMR spectroscopy and kinetic studies of the quadruplex forming RNA r(UGGAGGU)." Mol Biosyst **5**(11): 1347-1355.
Liu, X., C. Niu, J. Ren, J. Zhang, X. Xie, H. Zhu, W. Feng and W. Gong (2013). "The RRM domain of human fused in sarcoma protein reveals a non-canonical nucleic acid binding site." Biochim Biophys Acta **1832**(2): 375-385.
Liu, Z., I. Luyten, M. J. Bottomley, A. C. Messias, S. Houngninou-Molango, R. Sprangers, K. Zanier, A. Kramer and M. Sattler (2001). "Structural basis for recognition of the intron branch site RNA by splicing factor 1." Science **294**(5544): 1098-1102.
Mackereth, C. D., T. Madl, S. Bonnal, B. Simon, K. Zanier, A. Gasch, V. Rybin, J. Valcarcel and M. Sattler (2011). "Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF." Nature **475**(7356): 408-411.
Malgowska, M., D. Gudanis, R. Kierzek, E. Wyszko, V. Gabelica and Z. Gdaniec (2014). "Distinctive structural motifs of RNA G-quadruplexes composed of AGG, CGG and UGG trinucleotide repeats." Nucleic Acids Res **42**(15): 10196-10207.
Mao, Y. S., B. Zhang and D. L. Spector (2011). "Biogenesis and function of nuclear bodies." Trends Genet **27**(8): 295-306.
Marcel, V., P. L. Tran, C. Sagne, G. Martel-Planche, L. Vaslin, M. P. Teulade-Fichou, J. Hall, J. L. Mergny, P. Hainaut and E. Van Dyck (2011). "G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms." Carcinogenesis **32**(3): 271-278.
Maris, C., C. Dominguez and F. H. Allain (2005). "The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression." FEBS J **272**(9): 2118-2131.
Martin-Tumasz, S., A. C. Richie, L. J. Clos, 2nd, D. A. Brow and S. E. Butcher (2011). "A novel occluded RNA recognition motif in Prp24 unwinds the U6 RNA internal stem loop." Nucleic Acids Res **39**(17): 7837-7847.
Massi, F., E. Johnson, C. Wang, M. Rance and A. G. Palmer, 3rd (2004). "NMR R1 rho rotating-frame relaxation with weak radio frequency fields." J Am Chem Soc **126**(7): 2247-2256.
Matera, A. G. and Z. Wang (2014). "A day in the life of the spliceosome." Nat Rev Mol Cell Biol **15**(2): 108-121.
Matlin, A. J., F. Clark and C. W. Smith (2005). "Understanding alternative splicing: towards a cellular code." Nat Rev Mol Cell Biol **6**(5): 386-398.
Merendino, L., S. Guth, D. Bilbao, C. Martinez and J. Valcarcel (1999). "Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3' splice site AG." Nature **402**(6763): 838-841.
Millevoi, S., H. Moine and S. Vagner (2012). "G-quadruplexes in RNA biology." Wiley Interdiscip Rev RNA **3**(4): 495-507.
Moore, M. J. (2000). "Intron recognition comes of AGe." Nat Struct Biol **7**(1): 14-16.
Neri, D., T. Szyperski, G. Otting, H. Senn and K. Wuthrich (1989). "Stereospecific Nuclear Magnetic-Resonance Assignments of the Methyl-Groups of Valine and Leucine in the DNA-Binding Domain of the 434-Repressor by Biosynthetically Directed Fractional C-13 Labeling." Biochemistry **28**(19): 7510-7516.

Oberstrass, F. C., S. D. Auweter, M. Erat, Y. Hargous, A. Henning, P. Wenter, L. Reymond, B. Amir-Ahmady, S. Pitsch, D. L. Black and F. H. Allain (2005). "Structure of PTB bound to RNA: specific binding and implications for splicing regulation." Science **309**(5743): 2054-2057.
Pan, Q., O. Shai, L. J. Lee, B. J. Frey and B. J. Blencowe (2008). "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing." Nat Genet **40**(12): 1413-1415.
Papasaikas, P. and J. Valcarcel (2016). "The Spliceosome: The Ultimate RNA Chaperone and Sculptor." Trends Biochem Sci **41**(1): 33-45.
Petoukhov, M. V., D. Franke, A. V. Shkumatov, G. Tria, A. G. Kikhney, M. Gajda, C. Gorba, H. D. Mertens, P. V. Konarev and D. I. Svergun (2012). "New developments in the ATSAS program package for small-angle scattering data analysis." J Appl Crystallogr **45**(Pt 2): 342-350.
Phan, A. T., V. Kuryavyi, J. C. Darnell, A. Serganov, A. Majumdar, S. Ilin, T. Raslin, A. Polonskaia, C. Chen, D. Clain, R. B. Darnell and D. J. Patel (2011). "Structure-function studies of FMRP RGG peptide recognition of an RNA duplex-quadruplex junction." Nat Struct Mol Biol **18**(7): 796-804.
Rain, J. C., Z. Rafi, Z. Rhani, P. Legrain and A. Kramer (1998). "Conservation of functional domains involved in RNA binding and protein-protein interactions in human and Saccharomyces cerevisiae pre-mRNA splicing factor SF1." RNA **4**(5): 551-565.
Raj, B. and B. J. Blencowe (2015). "Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles." Neuron **87**(1): 14-27.
Ribeiro, M. M., G. S. Teixeira, L. Martins, M. R. Marques, A. P. de Souza and S. R. Line (2015). "G-quadruplex formation enhances splicing efficiency of PAX9 intron 1." Hum Genet **134**(1): 37-44.
Robert, X. and P. Gouet (2014). "Deciphering key features in protein structures with the new ENDscript server." Nucleic Acids Res **42**(Web Server issue): W320-324.
Rudner, D. Z., K. S. Breger and D. C. Rio (1998). "Molecular genetic analysis of the heterodimeric splicing factor U2AF: the RS domain on either the large or small Drosophila subunit is dispensable in vivo." Genes Dev **12**(7): 1010-1021.
Salz, H. K. (2011). "Sex determination in insects: a binary decision based on alternative splicing." Curr Opin Genet Dev **21**(4): 395-400.
Samatanga, B., C. Dominguez, I. Jelesarov and F. H. Allain (2013). "The high kinetic stability of a G-quadruplex limits hnRNP F qRRM3 binding to G-tract RNA." Nucleic Acids Res **41**(4): 2505-2516.
Sattler, M., J. Schleucher and C. Griesinger (1999). "Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients." Progress in Nuclear Magnetic Resonance Spectroscopy **34**(2): 93-158.
Scotti, M. M. and M. S. Swanson (2016). "RNA mis-splicing in disease." Nat Rev Genet **17**(1): 19-32.
Selenko, P., G. Gregorovic, R. Sprangers, G. Stier, Z. Rhani, A. Kramer and M. Sattler (2003). "Structural basis for the molecular recognition between human splicing factors U2AF65 and SF1/mBBP." Mol Cell **11**(4): 965-976.
Shao, C., B. Yang, T. Wu, J. Huang, P. Tang, Y. Zhou, J. Zhou, J. Qiu, L. Jiang, H. Li, G. Chen, H. Sun, Y. Zhang, A. Denise, D. E. Zhang and X. D. Fu (2014). "Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome." Nat Struct Mol Biol **21**(11): 997-1005.
Shen, Y. and A. Bax (2013). "Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks." J Biomol NMR **56**(3): 227-241.
Shen, Y., O. Lange, F. Delaglio, P. Rossi, J. M. Aramini, G. Liu, A. Eletsky, Y. Wu, K. K. Singarapu, A. Lemak, A. Ignatchenko, C. H. Arrowsmith, T. Szyperski, G. T. Montelione, D. Baker and A. Bax (2008). "Consistent blind protein structure generation from NMR chemical shift data." Proc Natl Acad Sci U S A **105**(12): 4685-4690.
Shen, Y., R. Vernon, D. Baker and A. Bax (2009). "De novo protein structure generation from incomplete chemical shift assignments." J Biomol NMR **43**(2): 63-78.
Sickmier, E. A., K. E. Frato, H. Shen, S. R. Paranawithana, M. R. Green and C. L. Kielkopf (2006). "Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65." Mol Cell **23**(1): 49-59.

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson and D. G. Higgins (2011). "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega." Mol Syst Biol **7**: 539.
Simone, R., P. Fratta, S. Neidle, G. N. Parkinson and A. M. Isaacs (2015). "G-quadruplexes: Emerging roles in neurodegenerative diseases and the non-coding transcriptome." FEBS Lett **589**(14): 1653-1668.
Subramanian, M., F. Rage, R. Tabet, E. Flatter, J. L. Mandel and H. Moine (2011). "G-quadruplex RNA structure as a signal for neurite mRNA targeting." EMBO Rep **12**(7): 697-704.
Sun, H. and L. A. Chasin (2000). "Multiple splicing defects in an intronic false exon." Mol Cell Biol **20**(17): 6414-6425.
Taliaferro, J. M., N. Alvarez, R. E. Green, M. Blanchette and D. C. Rio (2011). "Evolution of a tissue-specific splicing network." Genes Dev **25**(6): 608-620.
Todd, A. K., M. Johnston and S. Neidle (2005). "Highly prevalent putative quadruplex sequence motifs in human DNA." Nucleic Acids Res **33**(9): 2901-2907.
Turunen, J. J., E. H. Niemela, B. Verma and M. J. Frilander (2013). "The significant other: splicing by the minor spliceosome." Wiley Interdiscip Rev RNA **4**(1): 61-76.
Vasilyev, N., A. Polonskaia, J. C. Darnell, R. B. Darnell, D. J. Patel and A. Serganov (2015). "Crystal structure reveals specific recognition of a G-quadruplex RNA by a beta-turn in the RGG motif of FMRP." Proc Natl Acad Sci U S A **112**(39): E5391-5400.
Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh and X. Zhu (2001). "The sequence of the human genome." Science **291**(5507): 1304-1351.

Vincentelli, R., S. Canaan, V. Campanacci, C. Valencia, D. Maurin, F. Frassinetti, L. Scappucini-Calvo, Y. Bourne, C. Cambillau and C. Bignon (2004). "High-throughput automated refolding screening of inclusion bodies." Protein Sci **13**(10): 2782-2792.
Vorlickova, M., I. Kejnovska, J. Sagi, D. Renciuk, K. Bednarova, J. Motlova and J. Kypr (2012). "Circular dichroism and guanine quadruplexes." Methods **57**(1): 64-75.
Wang, Z. and C. B. Burge (2008). "Splicing regulation: from a parts list of regulatory elements to an integrated splicing code." RNA **14**(5): 802-813.
Weber, S. C. and C. P. Brangwynne (2012). "Getting RNA and protein in phase." Cell **149**(6): 1188-1191.
Will, C. L., C. Schneider, R. Reed and R. Luhrmann (1999). "Identification of both shared and distinct proteins in the major and minor spliceosomes." Science **284**(5422): 2003-2005.
Wolozin, B. (2012). "Regulated protein aggregation: stress granules and neurodegeneration." Mol Neurodegener **7**: 56.
Wüthrich, K. (1986). "NMR of Proteins and Nucleic Acids." Wiley.
Yaku, H., T. Murashima, H. Tateishi-Karimata, S. Nakano, D. Miyoshi and N. Sugimoto (2013). "Study on effects of molecular crowding on G-quadruplex-ligand binding and ligand-mediated telomerase inhibition." Methods **64**(1): 19-27.
Yamazaki, T., J. D. Formankay and L. E. Kay (1993). "2-Dimensional Nmr Experiments for Correlating C-13-Beta and H-1-Delta/Epsilon Chemical-Shifts of Aromatic Residues in C-13-Labeled Proteins Via Scalar Couplings." Journal of the American Chemical Society **115**(23): 11054-11055.
Yoshida, H., S. Y. Park, T. Oda, T. Akiyoshi, M. Sato, M. Shirouzu, K. Tsuda, K. Kuwasako, S. Unzai, Y. Muto, T. Urano and E. Obayashi (2015). "A novel 3' splice site recognition by the two zinc fingers in the U2AF small subunit." Genes Dev **29**(15): 1649-1660.
Zambrano, R., M. Jamroz, A. Szczasiuk, J. Pujols, S. Kmiecik and S. Ventura (2015). "AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures." Nucleic Acids Res **43**(W1): W306-313.
Zamore, P. D., J. G. Patton and M. R. Green (1992). "Cloning and domain structure of the mammalian splicing factor U2AF." Nature **355**(6361): 609-614.
Zhang, M., P. D. Zamore, M. Carmo-Fonseca, A. I. Lamond and M. R. Green (1992). "Cloning and intracellular localization of the U2 small nuclear ribonucleoprotein auxiliary factor small subunit." Proc Natl Acad Sci U S A **89**(18): 8769-8773.
Zhang, Y., T. Madl, I. Bagdiul, T. Kern, H. S. Kang, P. Zou, N. Mausbacher, S. A. Sieber, A. Kramer and M. Sattler (2013). "Structure, phosphorylation and U2AF65 binding of the N-terminal domain of splicing factor 1 during 3'-splice site recognition." Nucleic Acids Res **41**(2): 1343-1354.

# Appendix

## oligonucleotide sequences

**Table 0-1 List of DNA oligonucleotides.**

| No. | Name | Sequence 5'-3' |
|---|---|---|
| **1** | LS2_110_fwd | CATG CCATGG GA GCT CGC CGC CTG TAT GTG G |
| **2** | LS2_319_rev | GATC CTCGAG TTA AAT AGA TCT CTG GAC CACC |
| **3** | LS2_101_fwd | CATG CCATGG GA TCC GCC GCG ATT TC |
| **4** | LS2_325_rev | GATC CTCGAG TT ATG CGT TCT TTC CGC CTG GAA TAG |
| **5** | LS2_204_rev | GATC CTCGAG TT ACG AGG GCA CTG GCT GAT AG |
| **6** | LS2_242_fwd | CATG CCATGG GA AAT AAG ATC TAC GTA GGT GG |
| **7** | LS2_449_rev | GAT CGC GGC CGC TTA TTG TAG ATC ATC CGC CAG GTA C |
| **8** | LS2_53_fwd | CATG CCATGG GA CGA CGT TTC AGT CGG CCT CC |
| **9** | LS2_82_rev | GATC CTCGAG TTA GGC CAG CAT GGC CTT GTA CTG |
| **10** | LS2_204_fwd | CATG CCATGG GA  TCG ATT TCG GTC TCT GCA ATG GAG G |
| **11** | LS2_221_rev | GATC CTCGAG TTA GAT GGC AGG GAC CC |
| **12** | U2AF50_92_fwd | CATG TCATGA GA GCG CGT CGC CTG TAC GTT GG |
| **13** | U2AF50_286_rev | GATC CTCGAG TTA GCC CAC ACT GGC TCG TTG G |
| **14** | U2AF50_85_fwd | CATG TCATGA GA GGA TCG ACA ATT ACC CGA CAG G |
| **15** | U2AF50_290_rev | GATC CTCGAG TTA GGC ATT CTT GGC GCC CAC AC |
| **16** | U2AF50_37_fwd | CATG CCATGG GA AGG CGC AAG CCG TCG CTTTAT TGG |
| **17** | U2AF50_64_rev | GATC CTCGAG TTA CGC CTG CAT GGC TTT GTA TTG CAT C |
| **18** | U2AF38_43_fwd | CATG CCATGG GA TCG CAG ACG GTG CTT CTC CA |
| **19** | U2AF38_148_rev | GATC CTCGAG TTA CGG CGA TAG TTC CGA GTA CAC |
| **20** | LS2_RRM1_RNP1_fwd | CGA ATC TGG AGA AGA ACG ATG CCG ATC TGG AAT TCC GAT CC |
| **21** | LS2_RRM1_RNP1_rev | GGA TCG GAA TTC CAG ATC GGC ATC GTT CTT CTC CAG ATT CG |
| **22** | LS2_RRM1_RNP2_fwd | CAT GGG AGC TCG CGA CCT GGA TGT GGG TAA TAT TCC |
| **23** | LS2_RRM1_RNP2_rev | GGA ATA TTA CCC ACA TCC AGG TCG CGA GCT CCC ATG |
| **24** | LS2_RRM1_beta5_fwd | CCG TGG ACA AAC CTT GGA GAT AGA ACG GCC ACA CGA C |
| **25** | LS2_RRM1_beta5_rev | GTC GTG TGG CCG TTC TAT CTC CAA GGT TTG TCC ACG G |
| **26** | LS2_RRM2_RNP1_fwd | CGA ACC TGA ACA AGG GTG ACG CCG ACT TTG AGT ACT GCG |
| **27** | LS2_RRM2_RNP1_rev | CGC AGT ACT CAA AGT CGG CGT CAC CCT TGT TCA GGT TCG |
| **28** | LS2_RRM2_RNP2_fwd | GAT TCC CCC AAT GAG ATC GAC GTA GGT GGC TTG |
| **29** | LS2_RRM2_RNP2_rev | CAA GCC ACC TAC GTC GAT CTC ATT GGG GGA ATC |

# Abbreviations

**Table 0-2 Abbreviations**

| | |
|---|---|
| 1D, 2D, 3D | one-, two-, three- dimensionals |
| Å | Angstrom |
| ATP | Adenosine tri-phosphate |
| BME | β-mercaptoethanol |
| BPS | Branch point site |
| BSA | Bovine serum albumin |
| CD | Circular Dichroism |
| CSP | Chemical shift perturbation |
| Da | Dalton |
| $D_{max}$ | Maximum dimension |
| DMSO | Dimethyl Sulfoxide |
| DNA | Deoxyribonucleic acid |
| DTT | Dithiothreitol |
| *E. coli* | *Escherichia coli* |
| EDTA | Ethylenediaminetetraacetic acid |
| EMSA | Electrophoretic Mobility Shift Assay |
| FID | Free induction decay |
| G | Gravitational acceleration |
| GTP | Guanosine Triphosphate |
| HMQC | Heteronuclear Multiple Quantum Coherence |
| hnRNP | heterogenous nuclear ribonucleoprotein |
| HSQC | heteronuclear single quantum correlation |
| Hz | Hertz |
| IB | Inclusion body |
| IPTG | Isopropyl β-D-1-thiogalactopyranoside |
| ITC | Isothermal titration calorimetry |
| $K_d$ | Dissociation constant |
| LB | Lysogeny broth |
| LS2 | Large subunit 2 |
| MES | 2-(*N*-morpholino)ethanesulfonic acid |
| min | Minute |
| Ni-NTA | Nickel-Nitrilotriacetic acid |
| NMR | Nuclear Magnetic Resonance |
| NOE | Nuclear Overhauser Effect |
| NOESY | Nuclear Overhauser Effect spectroscopy |
| ns | Nanoseconds |
| nt | Nucleotide |
| OD600 | Optical density at 600 nm wavelength |
| PAGE | Polyacrylamide Gel Electrophoresis |
| PCR | polymerase chain reaction |
| PDB | protein data bank |

| | |
|---|---|
| pI | Isoelectric point |
| Poly-G | Poly guanosine |
| Poly-U | Poly uracil |
| ppm | parts per million |
| PRE | Paramagnetic relaxation enhancement |
| pre-mRNA | precursor messenger RNA |
| Py tract | Pyrimidine tract |
| RDC | residual dipolar coupling |
| Rg | Radius of gyration |
| RMSD | Root mean square deviation |
| RNA | Ribonucleic Acid |
| RRM | RNA recognition motif |
| RT | Room temperature |
| SAXS | Small angle x-ray scattering |
| SDS-PAGE | sodium dodecyl sulfate polyacrylamide gel electrophoresis |
| SF1 | Splicing factor 1 |
| SOFAST | band-selective optimized flip-angle short-transient |
| TEV | tobacco etch virus |
| TMPyP4 | meso-Tetra (N-methyl-4-pyridyl) porphine tetra tosylate |
| TOCSY | total correlation spectroscopy |
| Tris | Tris (hydroxymethyl) aminomethane |
| U2AF | U2 snRNP auxiliary factor |
| UHM | U2AF homology motifs |
| ULM | UHM ligand motifs |

# Acknowledgement

I would like to begin by expressing my sincere gratitude to Prof. Dr. Michael Sattler for giving me the opportunity to do my doctoral studies in his group. I truly appreciate all the support and freedom which he provided during my course of Ph.D. His positivity and passion towards science always motivated me to give my best. I deeply admire his patience when the project was going nowhere.

I feel fortunate to have Dr. Hyun-Seo Kang, as a co-supervisor and mentor during my doctoral studies. He bestowed upon me his generous help by being my "go-to guy" during the project. His top-down approach for problem-solving as well as constructive criticism boosted the project forward. I also appreciate his help in proofreading my thesis manuscript.

I would like to thank my thesis advisory committee members, Dr. Jürg Müller, and Dr. Gregor Witte for their valuable advice. I would also like to thank Dr. Mathew Taliaferro, Prof. Donald Rio and Prof. Christopher Burge for the excellent ongoing collaboration. I am also thankful to GRK 1721 graduate program for sponsoring all my conference visits as well as providing a platform to interact with Ph.D. students from different structural biology groups.

I would like to express my gratitude towards all the past and present members of the Sattler lab, Madl lab, and Reif lab. I really appreciate the help provided by Dr. Arie Geerlof in protein purification and expression, PD Dr. Gerd Gemmecker in NMR techniques and Dr. Ralf Stehle in SAXS experiments. I sincerely thank Christoph Hartmüller for performing the CS-ROSETTA structure calculation. Waltraud Wolfson took care of all the administrative work and cheered me whenever required, for which I will be always grateful to her. I am thankful to Dr. Divita Garg, Dr. Hamed Kooshapur for helping me to get settled in the lab and Dr. Alexander Beribisky for introducing me to the RNA work. Johannes Günther and Martin Rübbelke generously shared their experience in RNA and NMR, respectively. I am also deeply grateful for their help with German translation, without which thesis completion would have been impossible. My special thanks to Carolina Sanchéz Rico, Leonidas Emmanouilidis, Diana Rodriguez, Mohanraj Gopalswamy, Komal Soni, Pravin Jagtap, Eleni Kyriakou and Nishtha Gulati for all the memorable times spent together.

I am deeply grateful to my parents, without their support and patience, this work would not have been possible. My special thanks go to my wife Nimitha Rose Mathew, for providing me support, encouragement and suggestions whenever required. My monthly Freiburg visit provided a much-needed break from the lab work. I offer my regards to all my friends for their love, friendship and back up.

# Curriculum vitae

**Ashish Ashok Kawale**
Doctoral student,
Department of Chemistry,
Technical University of Munich,
Lichtenbergstrasse 4,
85747 Garching
E-mail: ashish.kawale@tum.de

---

**Education: -**

- 2011-present Doctoral student at Department of Chemistry, TU Munich
- 2009-2011 M.Sc. Biotechnology from IIT Roorkee, India
- 2006-2009 B.Sc. Biotechnology from R. J. college, Mumbai University, India
- 2006 Higher Secondary School, India

---

**Academic experience: -**

- 2011-present **Doctoral thesis at Department of Chemistry, TU Munich**
  Title: Structural and functional evolution of the alternative splicing factor LS2 from *Drosophila melanogaster*.
  Supervisor: Prof. Dr. Michael Sattler

- 2010-2011 **M.Sc. thesis at Indian Institute of Technology, Roorkee, India**
  Title: Purification and characterization of serine protease from *Euphorbia hirta*.
  Supervisor: Dr. Ashwani kumar Sharma

- 2012-present **Teaching assistant, Department of Chemistry, TU Munich**
  - Supervision and conducting B.Sc. Organic Chemistry practical's
  - M.Sc. practical's for Biomolecular NMR spectroscopy

**Publication: -**

- **Purification and physicochemical characterization of a serine protease with fibrinolytic activity from latex of a medicinal herb *Euphorbia hirta***.
  Girijesh Kumar Patel, Ashish Ashok Kawale, Ashwani Kumar Sharma. Plant Physiology and Biochemistry, Volume 52, March 2012, pages 104-111

---

**Conferences: -**

- International Conference on Magnetic Resonance in Biological Systems (Kyoto, Japan, 2016)
- Euromar (Zurich, Switzerland,2014)
- 35$^{th}$ FGMR meeting for advanced magnetic resonances-methods and application (Frauenchiemsee, Germany, 2013)

---

**Workshops: -**

- SAXS workshop organized by GRK1721 graduate school, 2013
- Adobe illustrator methods workshop organized by GRK1721 graduate school, 2013
- Crystallography workshop organized by GRK1721 graduate school, 2012