DISS. ETH No. 29297

Hybrid method for studying dynamic conformational ensemble of multidomain proteins: SRSF1 tandem RRMs in the free and bound states

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES

(Dr. sc. ETH Zürich)

presented by

CRISTINA KIM XUAN NGUYEN

MSc, University of Milano-Bicocca (Milan, Italy)

born on 04.10.1990

citizen of Italy

accepted on the recommendation of

Prof. Frédéric H.-T. Allain

Prof. Remco Sprangers

Prof. Karsten Weis

Dr. Antoine Cléry

2023

Summary

The genetic information present in the DNA is transcribed into RNA, which must undergo a series of maturation steps before it can be translated into functional proteins. Constitutive splicing is a major step in pre messenger ribonucleic acid (pre-mRNA) maturation, during which introns are removed and exons are ligated to form mature mRNA. Alternative splicing, on the other hand, allows for the inclusion or exclusion of exons and introns depending on various factors, such as developmental stage, cell type, and environmental conditions. This process helps generate diversity without increasing the size of an organism's genome. RNA binding proteins (RBPs) play a crucial role in regulating all post-transcriptional processing events, including splicing.

SRSF1 (Serine/arginine-rich splicing factor 1) is an important factor involved both in constitutive and alternative splicing and it was the first member of the family to be identified. SRSF1 is composed of two tandem RNA Recognition motifs (RRMs) and the RS domain at the C-terminal enriched in arginine and serine. The first RRM (or RRM1) is canonical while the second (or RRM2) is a pseudo-RRM and they are connected by a flexible linker domain.

Previous studies performed in our laboratory showed that RRM1 can bind CA/CG motifs using the canonical β-sheet interface, while RRM2 binds the GGA motif using the α_1-helix. In addition, a bimodal mode of interaction has been shown for SRSF1 tandem RRMs. More specifically, SELEX experiments demonstrated that RRM1 binds the cytosine located at -4 or +6 from the GGA motif recognized by RRM2.

Here, we aim to combine nuclear magnetic resonance (NMR) and electron paramagnetic resonance (EPR) to determine the structural ensemble of the tandem RRMs (SRSF1 RRM1+2) in the free state and in complex with RNA (5′-UCAUUGGAU-3′ and 5′-UGGAUUUUUCAU-3′ designed based on the SELEX experiments). Previous studies in the lab used the standard CYANA calculation to obtain structures of protein or protein-RNA complexes based on NMR data.

In the present study, we show the first application of the Multistate CYANA calculation method on a multidomain and dynamic system which was used previously only on single globular domains. In addition, to refine the ensemble we performed the EnsembleFit step from the MMMx toolbox to include data from EPR.

The ensembles obtained indicate that SRSF1 tandem RRMs do not behave as independent domains, but they already show preferred conformations in the free state; it has been observed that these conformations can promote the binding to 5′-UCAUUGGAU-3′ RNA via conformational selection and to 5′-UGGAUUUUUCAU-3′ RNA via both conformational selection and induced fit.

SRSF1 was also found in nuclear speckles (NSs), membrane-less organelles located in the nucleus that act as a central hub to coordinate various steps of nuclear gene expression regulation. The assembly and maintenance of NSs depend on interactions among their different components, many of which contain flexible low-complexity regions (LCRs); high concentrations of macromolecules promote phase separation, resulting in the formation of liquid droplets when concentrated proteins with LCRs are present. We then focused on investigating SRSF1 as one of the main components of the nuclear speckles in the context of phase separation. Due to solubility challenges, we decided to analyze the behavior of the SRSF1 tandem RRMs using light microscopy, turbidity measurements, and NMR Diffusion-ordered spectroscopy (DOSY) experiments. Our results showed that even in the absence of the RS domain, the two RRMs can form droplets *in vitro*, and RNA can play an important role in dissolving the droplets.

Riassunto

L'informazione genetica contenuta nel DNA viene trascritta in RNA, che successivamente deve subire una serie di passaggi di maturazione prima di poter essere tradotto in proteine funzionali. Lo splicing costitutivo è un passaggio importante della maturazione del pre-mRNA, durante il quale gli introni vengono rimossi e gli esoni vengono uniti per formare l'mRNA maturo. Lo splicing alternativo, invece, consente l'inclusione o l'esclusione di esoni e introni a seconda di vari fattori, come lo stadio di sviluppo, il tipo di cellula e le condizioni ambientali. Questo processo contribuisce a generare diversità senza aumentare le dimensioni del genoma di un organismo. Le proteine che legano l'RNA (RNA Binding Protein, RBP) svolgono un ruolo cruciale nella regolazione di tutti gli eventi di elaborazione post-trascrizionale, compreso lo splicing.

La proteina ricca in serina/arginina (Serine/arginine-rich splicing factor 1, SRSF1) è un fattore importante coinvolto sia nello splicing costitutivo che alternativo ed è stato il primo membro della famiglia delle proteine SR ad essere stato identificato. SRSF1 è composta da due motivi di riconoscimento dell'RNA in tandem (RNA Recognintion Motif, RRM) e dal dominio RS ricco in arginine e serine al C-terminale. Il primo RRM (o RRM1) è canonico, mentre il secondo (o RRM2) è uno pseudo-RRM.

Studi precedenti condotti nel nostro laboratorio hanno dimostrato che RRM1 può legare motivi CA/CG utilizzando la canonica interfaccia composta da β-foglietto, mentre RRM2 lega motivi GGA utilizzando l'α_1-elica**.** Inoltre, è stata dimostrata una modalità di interazione bimodale per SRSF1. In particolare, gli esperimenti SELEX hanno dimostrato che RRM1 lega la citosina situata a -4 o +6 dal motivo GGA riconosciuto da RRM2.

In questo lavoro, ci proponiamo di combinare la risonanza magnetica nucleare (RMN o NMR) e la risonanza paramagnetica elettronica (EPR) per determinare il complesso (ensemble) delle strutture degli RRM in tandem (SRSF1 RRM1+2) allo stato libero e legato all'RNA (5′-UCAUUGGAU-3′ e 5′-UGGAUUUUUCAU-3′ disegnati sulla base degli esperimenti SELEX). Precedenti studi condotti in laboratorio hanno utilizzato il metodo CYANA standard per ottenere strutture di proteine o complessi proteina-RNA sulla base di dati NMR. Nel presente studio mostriamo la prima applicazione del metodo Multistate CYANA su un sistema dinamico e composto da multipli domini, utilizzato in precedenza solo su domini globulari singoli. Inoltre, per perfezionare l'ensemble abbiamo eseguito il passaggio di EnsebleFit del MMMx toolbox per includere i dati di EPR.

I complessi strutturali ottenuti indicano che gli RRM in tandem di SRSF1 non si comportano come domini indipendenti, ma mostrano già preferite conformazioni allo stato libero; è stato osservato che queste conformazioni possono promuovere il legame all'RNA 5′-UCAUUGGAU-3′ tramite selezione conformazionale e all'RNA 5′-UGGAUUUCAU-3′RNA tramite selezione conformazionale e adattamento indotto.

SRSF1 è stato inoltre trovato negli speckles nucleari (Nuclear speckles, NS), organelli senza membrana situati nel nucleo che agiscono come un hub centrale per coordinare varie fasi della regolazione dell'espressione genica nucleare. L'assemblaggio e il mantenimento dei NS dipendono dalle interazioni tra i loro diversi componenti, molti dei quali contengono regioni flessibili a bassa complessità (Low Complexity Regions, LCR); elevate concentrazioni di macromolecole promuovono la separazione di fase, con conseguente formazione di "droplet" liquidi (goccioline) quando sono presenti alte concentrazioni di proteine contenenti LCR. Ci siamo quindi concentrati sullo studio di SRSF1, uno dei principali componenti degli speckles nucleari, nel contesto della separazione di fase. A causa di problemi di solubilità, abbiamo deciso di analizzare il comportamento dei RRM in tandem di SRSF1 utilizzando la microscopia ottica, le misure di torbidità e gli esperimenti di spettroscopia NMR a diffusione ordinata (DOSY). I nostri risultati hanno dimostrato che, anche in assenza del dominio RS, i due RRM possono formare droplet *in vitro* e che l'RNA può svolgere un ruolo importante nella dissoluzione dei droplet.

Table of Content

Chapter 1: Introduction 1

1.1 RNA and RNA Binding Proteins (RBPs) 2

1.2 Splicing and alternative splicing of the pre-mRNA 3

1.3 Splicing regulators 6

1.4 The SR protein family 8

1.5 Conformational changes of protein-ligand binding 11

1.6 Nuclear Magnetic Resonance (NMR) Spectroscopy and CYANA structure calculation 13

1.6.1 Chemical shift differences 14

1.6.2 Spin relaxation 14

1.6.3 Paramagnetic Relaxation Enhancement (PRE) 15

1.6.4 Structure calculation using CYANA 17

1.7 Electron Paramagnetic Resonance (EPR) Spectroscopy and MMMx modeling 18

1.7.1 DEER experiments 18

1.7.2 MMM modeling 19

1.8 Liquid-Liquid Phase Separation and Membraneless organelles 20

1.8.1 Diffusion experiments 22

1.9 Objectives and Overview 24

1.9.1 Chapter 2: Structural investigation of SRSF1 tandem RRMs 24

1.9.2 Chapter 3: SRSF1 and phase separation 24

Chapter 2: Structural investigation of SRSF1 tandem RRMs 25

2.1 Introduction 27

2.1.1 SRSF1 27

2.1.2 Bimodal mode of interaction of SRSF1 tandem RRMs with RNA 28

2.2 Results 32

2.2.1 Construct design and characterization of SRSF1 tandem RRMs 32

2.2.2 Analysis of SRSF1 tandem RRMs using DEER 35

2.2.2.1 Inter-RRM and RRM-linker distance distributions 35

2.2.2.2 Protein-RNA and Intra-RNA distance distributions 38

2.2.3 Analysis of SRSF1 tandem RRMs using PRE 39

2.2.4 *RigiFlex* modeling of SRSF1 tandem RRMs in the free form and in complex with RNA 41

2.2.5 Structure determination of SRSF1 tandem RRMs in the free form and in complex with RNA using the hybrid modeling 45

2.2.5.1 Ensemble of SRSF1 tandem RRMs in the free form 47

2.2.5.2 Ensemble of SRSF1 tandem RRMs in complex with 5′-UCAUUGGAU-3′ RNA and 5′-UGGAUUUUUCAU-3′ RNA 50

2.2.5.3 Comparison: MMMx *versus* Hybrid ensembles 55

2.3 Discussion 57

2.4 Material and methods 62

2.4.1 Site-directed mutagenesis 62

2.4.2 Expression and purification of the recombinant proteins 62

2.4.3 Site-directed spin labeling of the protein samples 63

2.4.4 Site-directed spin labeling of the RNA samples 63

2.4.5 NMR experiments 64

2.4.5.1 Chemical shift differences analysis 64

2.4.5.2 ^{15}N-{^{1}H} Heteronuclear NOE experiments 64

2.4.5.3 Protein dynamics study: ^{15}N spin-relaxation experiments 64

2.4.5.4 PRE experiments 65

2.4.6 EPR experiments 66

2.4.6.1 Continuous-Wave (CW) experiments 66

2.4.6.2 DEER sample preparation and measurements 66

2.4.6.3 DEER data analysis 67

2.4.7 Modeling with *RigiFlex* 67

2.4.8 Modeling with Hybrid method 68

2.4.8.1 DUMMY atoms to reproduce the MTSL spin label and calculation of the starting structure 68

2.4.8.2 Multistate CYANA calculation: PREP, CALC and SPLIT 69

2.4.8.3 MMMx EnsembleFit 70

Chapter 3: SRSF1 and Phase Separation 71

3.1 Introduction 73

3.2 Results 75

3.2.1 Phase separation of SRSF1 tandem RRMs and influence of the RNA 75

3.2.2 NMR Diffusion experiments on stabilized droplets *in vitro* 78

3.3 Discussion 80

3.4 Material and methods 82

3.4.1 Protein expression and purification 82

3.4.2 Light microscopy 82
3.4.3 Turbidity measurements 83
3.4.4 Sample preparation in agarose gel 83
3.4.5 NMR experiments 83
3.4.5.1 ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) experiments 83
3.4.5.2 Diffusion-ordered spectroscopy (DOSY) experiments 83
Chapter 4: Conclusions and Outlook 85
Chapter 5: Appendix 89
Appendix Chapter 2 90
Appendix Chapter 3 113
Chapter 6: References 117

Abbreviations

The three- and one-letter codes for amino acids and the one-letter code for nucleotides are used and are presumed to be known to the reader

CW	Continuous wave
CYANA	Combined assignment and dynamics algorithm for NMR applications
DEER	Double electron-electron resonance
DNA	Deoxyribonucleic acid
DOSY	Diffusion-ordered spectroscopy
EPR	Electron paramagnetic resonance
ESE/ESS	Exonic splicing enhancer/silencer
Exon	Expressed region of the messenger RNA
FUS	Fused in sarcoma
GB1	B1 domain of protein G
hetNOE	Heteronuclear nuclear overhauser effect
hnRNP	Heterogeneous nuclear ribonucleoprotein
HSQC	Heteronuclear single quantum coherence
IAP	3-(2-Iodoacetamido)-PROXYL
Intron	Intergenic region of the pre-messenger RNA
ITC	Isothermal titration calorimetry
Max	The maximal violation
MMM	Multiscale modeling of macromolecules
mRNA	Messenger RNA
MLO	Membraneless organelles
MTSSL	(1-oxyl-2,2,5,5-tetramethylpyrroline-3-methyl) methanethiosulfonate spin label
NMR	Nuclear magnetic resonance
NOE	Nuclear overhauser effect
Npl3	Nuclear shuttling protein 3
NS	Nuclear speckles
PRE	Paramagnetic relaxation enhancement
pre-mRNA	Precursor messenger RNA
PTBP1	Polypyrimidine tract binding protein 1
RBD	RNA binding domain
RBP	RNA binding protein
RNA	Ribonucleic acid
RNP	Ribonucleoprotein
RRM	RNA recognition motif

RS domain	Arginine/serine-rich domain
RT	Room temperature
SELEX	Systematic evolution of ligands by exponential enrichment
snRNP	Small nuclear ribonucleoprotein
SR protein	Serine/arginine-rich protein
SRSF1	Serine/arginine-rich splicing factor 1
SRSF1 RRM1+2	SRSF1 construct lacking the C-terminal domain, with an N-terminal GB1- and His6 tag, and Y37S-Y72S point mutations
TEV	Tobacco etch virus
TF	Target function
U2AF	U2 auxiliary factor
# viol	Number of restraints that are violated

Table of Figures

Figure 1.1 Constitutive and alternative splicing events
Figure 1.2 Human protein families of alternative splicing regulators and their domain composition
Figure 1.3 SR proteins and splicing regulation: cooperative and competitive binding
Figure 1.4 Mechanisms of protein-ligand binding
Figure 1.5 Paramagnetic Relaxation Enhancement (PRE)
Figure 1.6 The EPR-DEER experiment
Figure 1.7 Schematic visualization of MMMx pipelines
Figure 1.8 Overview of MLOs of eukaryotic cells and liquid–liquid droplets
Figure 1.9 DOSY pulse sequence
Figure 2.1 SRSF1 sequence and the RRM structures
Figure 2.2 SRSF1 RRM1 interacts with RNA on both sides of RRM2-binding site
Figure 2.3 NMR characterization of SRSF1 RRM1+2
Figure 2.4 Rotational correlation time of protein-RNA complexes
Figure 2.5 Inter-RRM and RRM-linker distance distribution restraints
Figure 2.6 Protein-RNA distance distribution restraints
Figure 2.7 PRE effect on SRSF1 RRM1+2 on the free form and in complex with RNA
Figure 2.8 Ensembles of SRSF1 RRM1+2 in the free form and in complex with RNA calculated using the *RigiFlex* modeling
Figure 2.9 Restraint fit of the ensemble of SRSF1 RRM1+2 in the free form calculated using the *RigiFlex* modeling
Figure 2.10 Scheme of the hybrid structural modeling approach
Figure 2.11 Structure calculation of SRSF1 RRM1+2 in the free form and in complex with RNA using standard CYANA
Figure 2.12 Ensembles of SRSF1 RRM1+2 in the free form calculated using the hybrid modeling
Figure 2.13 Distance distribution restraint fulfillment of SRSF1 RRM1+2 in the free state
Figure 2.14 Ensembles of SRSF1 RRM1+2 in complex with 5′-UCAUUGGAU-3′ RNA calculated using the hybrid modeling
Figure 2.15 Ensembles of SRSF1 RRM1+2 in complex with 5′-UGGAUUUUUCAU-3′ RNA calculated using the hybrid modeling
Figure 2.16 Comparison between ensembles obtained with *RigiFlex* and the hybrid modeling in the free state
Figure 2.17 Npl3 tandem RRMs in the free form and in complex with RNA

Figure 2.18 Comparison of the ensembles of SRSF1 RRM1+2 free form and in complex with RNA
Figure 3.1 Characterization of SRSF1 RRM1+2 construct in absence of GB1
Figure 3.2 Phase separation of SRSF1 RRM1+2 in presence of different RNAs *in vitro*
Figure 3.3 Agarose stabilization and DOSY experiments
Figure 3.4 Interfacial splicing model

Figure A1 SRSF1-RRM1(YS) construct
Figure A2 SRSF1 RRM1+2 single-cysteine mutants for PRE experiments
Figure A3 PRE experiments of SRSF1 RRM1+2 single-cysteine mutants in the free form
Figure A4 PRE experiments of SRSF1 RRM1+2 single-cysteine mutants in complex with 5′-UCAUUGGAU-3′ RNA
Figure A5 PRE experiments of SRSF1 RRM1+2 single-cysteine mutants in complex with 5′-UGGAUUUUUCAU-3′ RNA
Figure A6 Restraint fit of the ensemble SRSF1 RRM1+2 in complex with 5′-UCAUUGGAU-3′ RNA calculated using the *RigiFlex* modeling
Figure A7 Restraint fit of the ensemble SRSF1 RRM1+2 in complex with 5′-UGGAUUUUUCAU-3′ RNA calculated using the *RigiFlex* modeling
Figure A8 Ensembles of SRSF1 RRM1+2 in the free form calculated using Multistate CYANA calculation
Figure A9 PRE restraint fit of the ensemble SRSF1 RRM1+2 in the free state after Multistate CYANA calculation
Figure A10 PRE restraint fit of the ensemble SRSF1 RRM1+2 in the free state after the EnsembleFit step
Figure A11 Restraint fit of the ensemble SRSF1 RRM1+2 in complex with 5′-UCAUUGGAU-3′ RNA after the EnsembleFit step
Figure A12 Restraint fit of the ensemble SRSF1 RRM1+2 in complex with 5′-UGGAUUUUUCAU-3′ RNA after the EnsembleFit step
Figure A13 Comparison between ensembles of SRSF1 RRM1+2 in complex with RNA obtained with *RigiFlex* and the hybrid modeling
Figure B1 Phosphorylation of the RS domain
Figure B2 SRSF1 RRM1+2 without GB1 in presence of RNA
Figure B3 ^{1}H-^{15}N HSQC of SRSF1 RRM1+2 without GB1 stabilized in 0.5% agarose

Chapter 1: Introduction

1.1 RNA and RNA Binding Proteins (RBPs)

In accordance with the Central Dogma of Molecular Biology, proposed by Francis Crick in 1957, genetic information encoded in DNA is transmitted to functional proteins via RNA, specifically messenger RNA (mRNA) (Cech and Steitz 2014). The transcription of pre-mRNA by RNA polymerase II is the initial step, after which the mRNA undergoes various maturation processes in the nucleus before being transported to the cytoplasm for translation into proteins. Maturation processes include capping the 5' end of the mRNA, polyadenylation at the 3' end, and removal of introns during splicing. Once mature, mRNA is transported across the nuclear membrane to the cytoplasm, where it is translated into proteins by ribosomes and eventually degraded. However, RNA molecules are not limited to messenger RNA and can also have other functions, such as enzymatic activities (e.g. ribozymes such as rRNA), structural roles, and regulatory activities (Kruger et al. 1982; Guerrier-Takada et al. 1983). Non-coding RNA, such as small interfering RNA (siRNA), microRNA (miRNA), and long non-coding RNA (lncRNA), play a role in regulating gene expression (Wilusz, Sunwoo, and Spector 2009; Eddy 2001). In all cases, RNA requires modifications to function properly, and the interaction with RNA-binding proteins (RBPs) is crucial for the proper function of RNA and the regulation of the DNA-RNA-protein pathway.

In general, RBPs contain canonical RNA binding domains (RBDs) such as RNA recognition motifs (RRMs), which are the most common (Afroz et al. 2015) , K-homology (KH), DEAD-box helicases, and zinc-finger (ZnF) domains (Gerstberger, Hafner, and Tuschl 2014; Castello et al. 2016). Intrinsically disordered regions (IDRs) are also common in RBPs and can also bind RNA (Castello et al. 2016). The RBDs often occur in multiple repeats or combinations, allowing for coordinated and enhanced binding to RNA (Clery, Schubert, and Allain 2012); they are combined with different IDRs such as the glycine-rich or arginine-serine rich (RS) domains. By containing multiple RBDs, a protein achieves higher sequence specificity and affinity than having a single domain.

The structural study of multi-domain RBPs has largely focused on those containing two RBDs. Structures of RBPs with two RRMs showed that both domains can adapt a variety of conformations with respect to each other (Afroz et al. 2015) it has been reported that tandem RRM recognize RNA primarily in three main ways. The tandem RRMs can be independent of each other in their free state and adopt a rigid structure upon binding to RNA. In the second case, the tandem RRMs maintain a fixed orientation in their free state, which is preserved upon RNA binding. In the last case, a conformational change in the structure of the tandem RRMs occurs upon RNA binding.

Despite this diversity, RBPs play crucial roles in coordinating many steps of gene expression through their ability to bind RNA with high specificity and affinity.

1.2 Splicing and alternative splicing of the pre-mRNA

Pre-mRNA splicing was discovered in the late 1970s when it was demonstrated that eukaryotic pre-mRNAs molecules contain both protein-coding sequences (expressed regions or exons) and non-coding sequences (intragenic regions or introns) that were not present in the mature mRNA (Berget, Moore, and Sharp 1977; Chow et al. 1977).

It is now established that the number and length of introns vary significantly among different eukaryotic species (Rogozin et al. 2012). For example, in *S. cerevisae*, only about 4% of genes contain introns, and those that do usually have a single intron with an average length of 100-400 nt (Hooks, Delneri, and Griffiths-Jones 2014; Parenteau et al. 2008). In contrast, the average number of introns in human is eight per gene, with some introns spanning several thousand bases in length with an average of 3365 nt, making them significantly larger than exons which usually range from 50-250 nt (Sakharkar et al. 2005; Chen and Manley 2009). During pre-mRNA processing, over 90% of the pre-mRNA is removed as introns, and only about 10% of the average pre-mRNA is joined together as exonic sequences via pre-mRNA splicing. Almost all protein-coding genes in eukaryotic cells contain introns, which are removed by RNA splicing in the nucleus during pre-mRNA processing (Tazi, Bakkour, and Stamm 2009).

Through the splicing event, the introns are removed by the spliceosome, a macromolecular complex composed by several protein components and five snRNPs (small nuclear ribonucleoproteins particles): U1, U2, U4, U5 and U6 (Kramer 1996; Will and Luhrmann 2001, 2011). The assembly of the spliceosome is firstly initiated by the recognition of the 5´ ss (splice site) by the U1 snRNP and the 3´ ss by the heterodimeric U2AF (U2 snRNP auxiliary factor) forming the E complex. Second, the U2 snRNP is recruited to the BP (branch-point), in an ATP-dependent step, forming the A complex. Third, the recruitment of the U4/U6-U5 tri-snRNP forms the B complex, followed by structural rearrangements to form finally the active spliceosomal C complex (Matlin and Moore 2007). The spliceosome is a dynamic structure and in the active C complexes more than 300 proteins have been identified (Rappsilber et al. 2002; Zhou et al. 2002; Bessonov et al. 2008; Jurica and Moore 2003).

There is another type of spliceosome which is less common that is the U12-dependent spliceosome (for the splicing of the U12-type introns). U2-type and U12-type introns differ for their consensus sequence of their splice site: U2-type introns have GURGU (R stands for A or G) and YAG (Y stands for U or C) for the 5´ and the 3´ splice sites, respectively, while U12-type introns have RUAUCCUU and YAS (S stands for G or C) (Patel and Steitz 2003; Turunen et al. 2013). The U12-dependent splicing (or minor splicing) share with the U2-dependent the U5 snRNP but he has analogous U11, U12, U4atac, and U6atac snRNPs (Patel and Steitz 2003).

In both cases, the snRNAs are bound by seven Sm proteins and a varying composition of other particle-specific proteins to form the small nuclear ribonucleoproteins (snRNPs). In addition to the snRNPs, there are numerous non-snRNP proteins involved in the spliceosome assembly (Will and Luhrmann 2011).

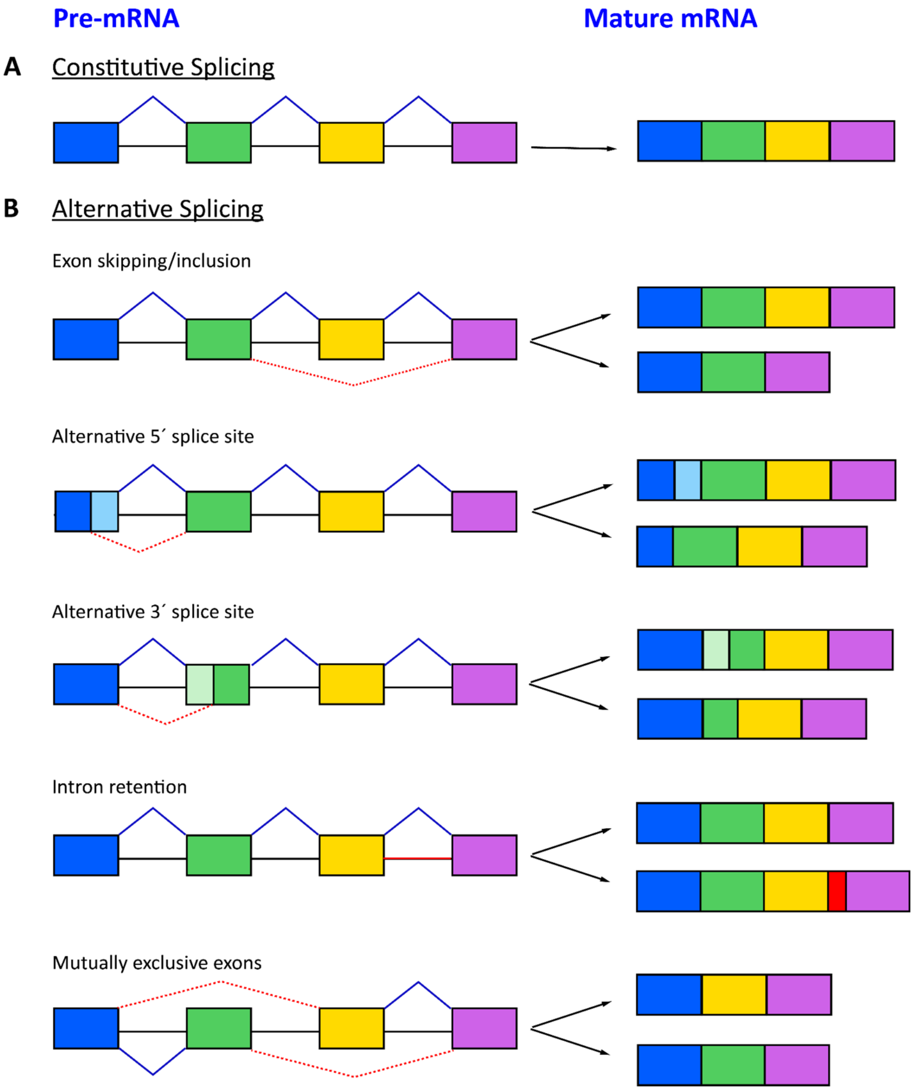


Figure 1.1 ***Constitutive and alternative splicing events****.* *(A) Constitutive splicing. (B) Five most common modes of alternative splicing: exon skipping/inclusion, alternative 5´splice-site selection, alternative 3´ splice-site selection, intron retention and mutually exclusive exons. On the right the mature mRNA derived from each event is shown. Modified from (Frankiw, Baltimore, and Li 2019).*

There are two types of splicing mechanisms in higher eukaryotes: constitutive and alternative splicing (Figure 1.1). In the constitutive splicing the introns are removed, and exon ligated in the same order they are in the gene (Figure 1.1A); in the alternative splicing certain exons may be skipped resulting in various forms of mature mRNAs derived from a single pre-mRNA transcript (Figure 1.1B). Each variant of mRNA generated by the alternative splicing encodes for a different protein isoform that differ in protein-protein interactions, subcellular localization or catalytic ability (Stamm et al. 2005); the alternative splicing is a crucial mechanism for generating proteomic diversity in higher eukaryotes (Ule and Blencowe 2019). It is estimated that more than

88% of the human protein-coding genes are affected by alternative splicing (Kampa et al. 2004).

Various types of alternative splicing events lead to the generation of distinct transcripts (Figure 1.1B). The most common events include cassette exon inclusion or skipping, mutually exclusive exons, intron retention, and the use of different 3' or 5' splice sites (Baralle and Giudice 2017). Besides conventional alternative splicing, there are emerging non-canonical splicing reactions such as the production of circular RNA (Baralle and Giudice 2017) or chimeric transcripts, which consist of exons from different genes (Babiceanu et al. 2016).

Since only a few reports have documented mutations in core splicing machinery elements that lead to human diseases, it is possible that defects in the general splicing machinery are generally incompatible with life. In contrast, changes in alternative splicing (which can affect numerous genes) may be tolerated by an organism, although these changes could result in a disease. The number of discovered diseases associated with changes in alternative splicing has increased dramatically in recent years (Tazi, Bakkour, and Stamm 2009).

Spinal Muscular Atrophy (SMA) is the most well-known disease associated with alternative splicing, which is further discussed in Section 1.4. Another group of diseases called Tauopathies, characterized by abnormal intracellular accumulations of abnormal filaments containing the microtubule-associated protein tau, affect the central nervous system. The gene MAPT encodes the tau protein, which undergoes extensive alternative splicing, including eight alternatively spliced exons out of a total of sixteen. Genetic studies have identified 42 rare dominant mutations in the tau gene that cause frontotemporal dementia with Parkinson linked to chromosome 17 (FTDP-17). The majority of these mutations affect the regulation of exon 10 splicing, altering its normal inclusion fraction and leading to changes in the pre-mRNA encoding 3R and 4R repeat tau isoforms, which have been associated with FTDP-17. Abnormal intracellular tau aggregates are also found in other tauopathies, such as Alzheimer's disease (Glatz et al. 2006; Tazi, Bakkour, and Stamm 2009).

Frontotemporal lobar dementias are caused by the loss of the splicing factor TDP43 (TAR DNA-binding protein 43 kd), a member of hnRNP family. In this disease, TDP43 is cleaved by caspase-3, and the resulting cleavage fragments accumulate in the cytosol, where they form aggregates. Progranulin inhibits the caspase-3 reaction, which explains why mutations reducing progranulin expression cause FTLD (Zhang et al. 2007). Whether the disease is caused by a loss of nuclear function of TDP43 or a possible cytotoxic accumulation is not clear (Tazi, Bakkour, and Stamm 2009); in addition, mutations in the gene encoding TDP43 are found in families with amyotrophic lateral sclerosis, as well as in sporadic cases (Sreedharan et al. 2008).

Furthermore, it has been observed that alterations in alternative splicing are closely associated with cancer (Venables 2006). The expression of alternative or even tumor-specific splice variants significantly affects several essential cellular processes critical for cancer biology like cell proliferation, motility, and drug response (Skotheim and Nees 2007). One of the causes of these changes can be attributed to alterations in the concentration, localization, composition, or activity of trans-acting regulatory factors (as hnRNP or SR proteins), which can influence splice selection. Moreover, mutations in splicing regulatory elements that affect splice site selection provide a clear link between pre-mRNA processing and cancer development (Skotheim and Nees 2007).

1.3 Splicing regulators

There are two primary families of splicing regulators: the heterogeneous nuclear ribonucleoproteins (hnRNPs) (Gallinaro et al. 1981; Dreyfuss et al. 1993) and the serine/arginine-rich (SR) proteins (Braberg et al. 2013; Long and Caceres 2009); SR proteins will be discussed in more detail in the next section (Section 1.5). There are 20 hnRNP proteins named hnRNP A-U (Figure 1.2B), which contain one or more RNA binding domains such as RNA recognition motifs (RRM), hnRNP K homology (KH) domains, or arginine-glycine-glycine (RGG) boxes (Geuens, Bouhy, and Timmerman 2016), as well as other auxiliary domains that are glycine-, proline-, or acid-rich (Dreyfuss, Kim, and Kataoka 2002). They associate with newly transcribed RNA and help to stabilize it and, are involved not only in alternative splicing but in many steps of RNA metabolism.

HnRNPs are often found to bind to intronic splicing silencers (ISSs) to prevent exon inclusion. Their primary competitors are SR proteins, which often bind to exonic and intronic splicing enhancers (ESEs and ISEs). Generally, hnRNPs act as repressors by binding to exonic and intronic silencers, mainly by multimerizing along exons and blocking spliceosome assembly. This blocking occurs either by hindering small nuclear ribonucleoprotein (snRNP) recruitment or by looping out the exon (Zhu, Mayeda, and Krainer 2001; House and Lynch 2006; Martinez-Contreras et al. 2006). However, in some specific cases, hnRNPs can act as activators (Douglas and Wood 2011). For example, hnRNP G can bind to exon 7 and activate its inclusion in the survival of motor neurons (SMN) gene transcripts (Moursy, Allain, and Clery 2014; Hofmann and Wirth 2002).

In addition to the these two families of regulators, cells also contain tissue-specific splicing factors such as CELF (CUGBP, Elav-like family) (Dasgupta and Ladd 2012), Nova (neuro-oncological ventral antigen) (Irimia et al. 2011), and MBNL (muscleblind-like) (Pascual et al. 2006). Although the hnRNPs are a diverse family of abundant proteins, CELF and MBNL proteins are sometimes considered as hnRNP-like proteins. The CELF family consists of six members (Figure 1.2C), among which CELF3-6 are tissue-specific. All CELF proteins contain three RNA recognition motifs (RRMs) arranged in a unique configuration, with two N-terminal RRMs followed by a linker and a third C-terminal RRM. Human CELF genes are particularly involved in alternative splicing during brain and nervous system development (Ladd 2013).

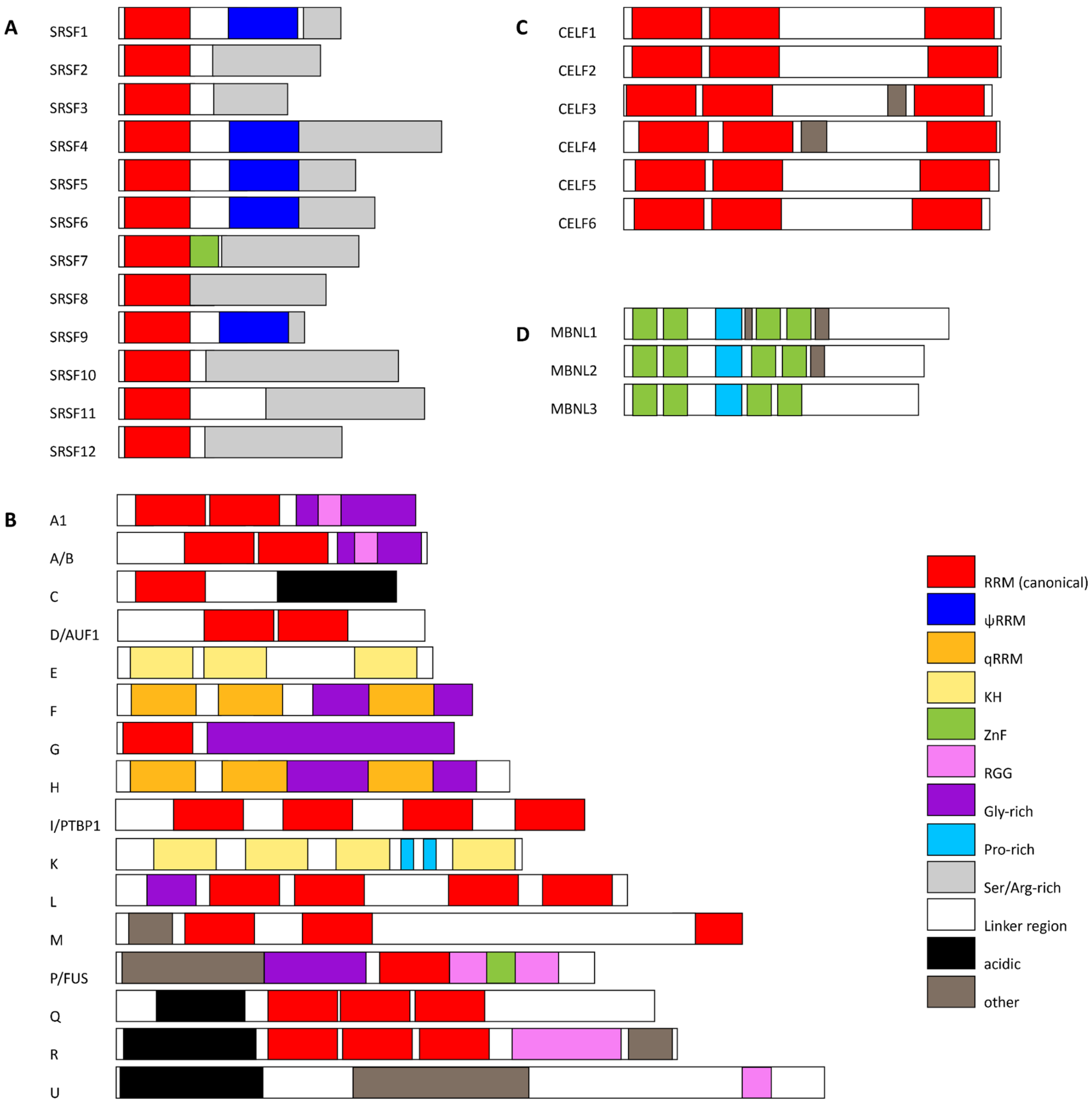


Figure 1.2 Human protein families of alternative splicing regulators and their domain composition. *(A) SR protein family, (B) hnRNP family, (C) CELF family, (D) MBNL family. RNA Binding Domains (RBDs) = RRM (canonical), ψRRM, qRRMs, KH, ZnF; auxiliary domains = RGG, Gly-rich, Pro-rich, Ser/Arg-rich, linker region, acidic. Color codes of the domains are explained within the figure.*

The MBNL family includes three members that bind RNA with two pairs of tandem CCCH-type zinc finger domains separated by a long linker (Figure 1.2D). They are tissue-specific alternative splicing regulators, for example, in heart and muscle, and bind to intronic splicing elements. The MBNL genes consist of 10 exons, some of which can be alternatively spliced, resulting in different isoforms (Fardaei et al. 2002; Kino et al. 2004; Patryk Konieczny et al. 2014).

1.4 The SR protein family

The SR proteins were first discovered as splicing factors in the early 1990s in *Drosophila*, where genetic screens identified SWAP (suppressor-of-white-apricot), Tra (transformer) and Tra-2 (transformer-2) as splicing factors (Chou, Zachar, and Bingham 1987; Boggs et al. 1987; Amrein, Gorman, and Nothiger 1988). The common feature of these three proteins was a domain rich in arginine and serine dipeptides called the arginine/serine (RS) domain. Subsequent identification of SF2/ASF (splicing factor 2/alternative splicing) and SC35 (spliceosomal component 35) from human cell lines also revealed the presence of extended RS domains (Ge and Manley 1990; Krainer, Conway, and Kozak 1990; Fu and Maniatis 1992). The term "SR protein" was coined following identification of additional proteins containing an RS domain recognized by a monoclonal antibody, mAb 104, which binds active sites of RNA polymerase II transcription (Roth, Murphy, and Gall 1990).

In human there are 12 canonical SR proteins (Fig 1.2A). All SR proteins share a modular structure consisting of at least an RNA Recognition Motif (RRM) and an RS domain. The RRM domains located at the N-terminal of SR proteins are responsible for the specificity of RNA binding, while the C-terminal RS domain promotes mainly protein-protein interactions (Figure 2.1 B).

SRSF2, SRSF3, SRSF7, SRSF8, and SRSF10-12 are composed of a single RRM1, while SRSF1, SRSF4-6, and SRSF9 contain an additional RRM known as a pseudo-RRM or RRMH (Figure 1.2B). The two RRMs of SR proteins exhibit distinct RNA binding mechanisms and have different binding specificities when tested in isolation. The presence of two RRMs could allow SR proteins to associate with a broader range of RNAs than those containing only one RRM. Alternatively, the two RRMs could increase the specificity of SR proteins if both RRM recognition sequences are needed for RNA targeting. Lastly, one of the RRMs may have evolved additional functions that do not contribute to RNA binding *in vivo* (Anko 2014).

The RS domains of SR proteins participate in interactions with other splicing factors that contain RS domains, SR-related proteins, and components of the general splicing machinery (Zahler et al. 1992; Wu and Maniatis 1993; Shen and Green 2004; Shen, Kan, and Green 2004). The RS domain also functions as a nuclear localization signal by interacting with transportin-SR, the SR protein nuclear import receptor (Caceres et al. 1997; Kataoka, Bachorik, and Dreyfuss 1999; Lai et al. 2000). However, structural characterization of SR proteins has not yet been achieved due to the poor solubility of these proteins in their free state, the phosphorylation state of the serines within the RS domain, and the degenerate RNA-binding sequences recognized by SR proteins. Consequently, only isolated RRM domains of SR proteins have been structurally analyzed using NMR spectroscopy, and no structure of the RS domain has been solved.

The SR proteins are a significant constituent of nuclear speckles (Manley and Tacke 1996; Fu 1995) and are localized to these membraneless organelles by the RS domain (Caceres et al. 1997; Hentze et al. 2018). The intranuclear organization of SR proteins is dynamic and they are mobilized from the interchromatin granule clusters (IGC) to the sites of co-transcriptional splicing, the perichromatin fibrils. The mobilization of SR proteins from the IGCs to the perichromatin fibrils requires both the RNA-binding domains and RS domains, as well as the phosphorylation of the RS domain. SR proteins are involved in several steps of the RNA processing and regulatory process due to their ability to bind different types of RNAs and their localization in both the nucleus and cytoplasm.

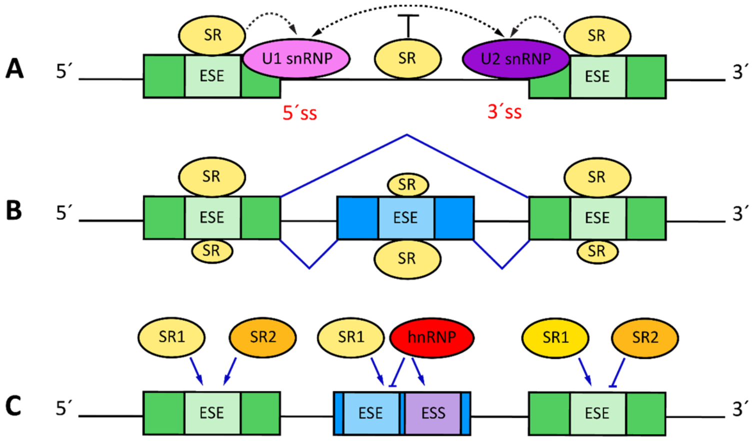


***Figure 1.3 SR proteins and splicing regulation: cooperative and competitive binding**. (A) SR protein binding to exonic splicing enhancer (ESE) stimulates the recognition of the nearby 5' ss by U1 snRNP and 3' ss by U2 snRNP. On the other hand, SR protein binding to intronic regions inhibits splicing, likely through interfering with the communication between the functional 5' ss and 3' ss. (B) SR protein-dependent exon inclusion or skipping. (C) Cooperative and competitive binding of SR proteins; ESS = Exonic Splicing Silencer.. Modified from (Zhou and Fu 2013).*

The activity of SR proteins is linked to their shuttling between the nucleus and cytoplasm, which is regulated by the phosphorylation status of the RS domain. The SR protein phosphorylation/dephosphorylation cycle is necessary for spliceosome assembly and splicing catalysis. After splicing, SR proteins are dephosphorylated, and only hypo-phosphorylated SR proteins can interact with the nuclear export machinery. Re-phosphorylation is necessary for SR proteins to return to the nucleus. The phosphorylation status of SR proteins determines their cellular localization and activity, and phosphorylation can represent an important mechanism *in vivo* to integrate signals from cellular pathways to coordinate gene expression.

In the context of splicing mechanism, SR proteins play a crucial role by binding to splicing enhancers located in both exons and introns, and promoting the recruitment of U1 snRNP to 5´ splice site (5´ss) and U2 auxiliary factor (U2AF) to 3´ splice site (Figure 1.3A). This recruitment occurs in a phosphorylation-dependent manner via interactions with the U1 subunit U1-70k and U2AF35 through their RS domains.

Inhibition of U1 and U2 recruitment is observed when SR proteins bind to intronic splicing silencers. Additionally, SR proteins stabilize the base-pairing of U2 snRNP with the branchpoint and facilitate the recruitment of the U4/U6–U5 tri-snRNP and U6 snRNP binding. The formation of the catalytically active spliceosome occurs through extensive remodeling of RNA–RNA and RNA–protein interactions, which is coupled to the dephosphorylation of SR proteins. SR proteins typically promote splice site usage, depending on their binding strength, expression levels, and extent of cooperation and competition with other SR proteins and hnRNP proteins (Figure 1.3B). However, not all SR proteins promote splicing. SRSF10 and SRSF12 also act as global repressors of splicing, depending on their phosphorylation state (Cowper et al. 2001). In the case of alternative splicing, SR proteins play an essential role in the inclusion of internal alternative exons by binding to them (Figure 1.3C). Conversely, SR protein binding to flanking competing exons causes exon skipping (Zhou and Fu 2013). The interaction between SR proteins and members of the hnRNP family, which includes several well-established splicing repressors mediating the repressive effects of exonic splicing silencers (ESSs), is also a crucial aspect of alternative splicing mechanisms. The functional antagonism between SR proteins and hnRNP proteins was first observed between SRSF1 and hnRNP A1 on various alternative splicing modalities. The underlying molecular mechanisms of this antagonism are distinct, with SRSF1 promoting selection of the proximal 5′ splice site (closest to the 3′ ss) in the case of competing splice site donors, while hnRNP A1 promotes the usage of more distal sites by reducing the binding of U1 snRNP at the proximal site.

SR proteins exhibit functions that are independent of their splicing activity (Wagner, 2021). These non-redundant functions in splicing indicate their crucial roles in various less-defined mechanisms of transcriptional activation, such as mRNA export, nonsense-mediated decay (NMD), and translation. The non-canonical functions of SR proteins suggest that they are essential factors involved in coordinating multiple steps of gene expression, underscoring their significance in RNA processing (Sanford, Ellis, and Caceres 2005; Zhong et al. 2009).

The disruption of the various roles of SR family proteins could led to different diseases. There is now growing evidence connecting the mis-expression of SR proteins with the development of cancerous tissues. Specifically, while SR protein levels are downregulated during cell differentiation, their abnormal overexpression has been shown to promote dedifferentiation, tumorigenesis, and metastasis (Zheng et al. 2020). For instance, SRSF1 has been linked to acute lymphoblastic leukemia, prostate, lung, and breast cancer (Zheng et al. 2020; Wagner and Frye 2021); SRSF2 and SRSF4 have been associated with acute myeloid leukemia (Zheng et al. 2020; Li and Wang 2021; Tan, Wang, and Ma 2018). On the other hand, SRSF3 is connected to colon cancer and osteosarcoma (More and Kumar 2020; Che and Fu 2020), while SRSF5 is associated with lung and breast cancer and, SRSF6 with breast and skin cancer (Zheng et al. 2020; Cerasuolo et al. 2020). Moreover, the dysregulation of the canonical and non-canonical

functions of SR proteins also contributes to various neurological disorders, liver disease, as well as coronary and cardiac diseases (More and Kumar 2020; Ortiz-Sanchez et al. 2019; Larrasa-Alonso et al. 2021; Kumar et al. 2019).
SR proteins have been demonstrated to play a significant role in the regulation of various splicing events that impact the different transcripts of HIV-1. The virus employs a combination of several alternative 5' and 3' splice sites to produce over 40 distinct mRNAs from its complete genomic pre-mRNA (Stoltzfus and Madsen 2006). Additionally, HIV infection alters the levels of splicing factors, including SR proteins, which regulate viral alternative splicing and thereby the replication of the virus (Dowling et al. 2008). Consequently, a promising alternative strategy to overcome the issue of resistance of HIV-1 to current inhibitors is to target the involvement of SR proteins in HIV pre-mRNA splicing (Soret, Gabut, and Tazi 2006).

Spinal Muscular Atrophy (SMA) is a neurodegenerative disease caused by the degeneration of motor neurons. This leads to muscle denervation and a reduction in the number of motor neuron units present in the spinal cord and lower brainstem (Crawford and Pardo 1996). The underlying cause of this disease is the absence of functional SMN1 (survival of motor neuron 1) gene product, which plays an important role in the biogenesis of small nuclear ribonucleoproteins (snRNPs). The SMN2 gene is a paralogue of SMN1 that is located centromeric to it. It differs from SMN1 by a single nucleotide change, a C > U transition in exon 7. This change leads to the skipping of exon 7 and the production of a non-functional protein. The exon-skipping event has been attributed to the loss of an exonic splicing enhancer that is bound by SRSF1 protein (Cartegni and Krainer 2002) or the creation of an exonic splicing silencer that is dependent on the hnRNP A/B proteins (Kashima and Manley 2003).

1.5 Conformational changes of protein-ligand binding

Structural and biochemical investigations of protein-RNA interactions aim to elucidate the mechanism by which a protein specifically recognizes and interacts with an RNA site, and how this interaction affects the structure and function of both protein and RNA (Draper 1995; Reyes and Kollman 2000). In general, both protein binding and function frequently involve conformational changes, which can occur in the absence or presence of ligands. Ligands may also "select" protein conformations for binding or unbinding (Weikl and Paul 2014). The conformations with high energy are in dynamic, thermally-activated interchange with the ground-state conformations of lower energy, which correspond to the most stable conformations. The energy-landscape perspective, which was originally developed for protein folding, provides a theoretical basis for the conformational dynamics of proteins in the native, folded state (Dill and Chan 1997; Bryngelson et al. 1995; Dill 1985; Bryngelson and Wolynes 1987). This perspective, incorporates key concepts such as the population shift of protein conformations during binding or chemical reactions, and conformational selection, which posits that ligands can "select" pre-existing, higher-energy conformations of proteins for binding (Ma et al. 1999). The central aspects of these concepts were already evident in early models of protein allostery. Overall, both conformational changes and binding/unbinding events are thermally activated processes that require overcoming free-energy barriers.

The protein-ligand interaction involves two components: specific binding of P to L and associated conformational changes that may occur before and/or after the binding step. The combination of binding events and conformational transitions in a given recognition mechanism generates a range of kinetic behaviors that can be measured experimentally. The challenge is to understand the nature of the conformational transitions involved in the recognition process through analysis of the transient behavior of the system as it relaxes to equilibrium (Eigen 1957, 1968).

The initial model proposed by Fischer, known as the lock-and-key model (Fischer 1894), postulated that the protein or in general a macromolecule in both the free and bound states existed as the same species. Although the binding may stabilize the protein, it was assumed that the conformational distribution remained unaltered. However, with the advent of x-ray crystallography (Pozzi et al. 2012), NMR spectroscopy (Boehr et al. 2006; Tang, Schwieters, and Clore 2007), and single-molecule fluorescence detection, it is now evident that the lock-and-key model oversimplifies protein-ligand binding. This model envisions a rigid collision between the protein and the ligand without accounting for the conformational flexibility of the different macromolecules involved (Henzler-Wildman et al. 2007).

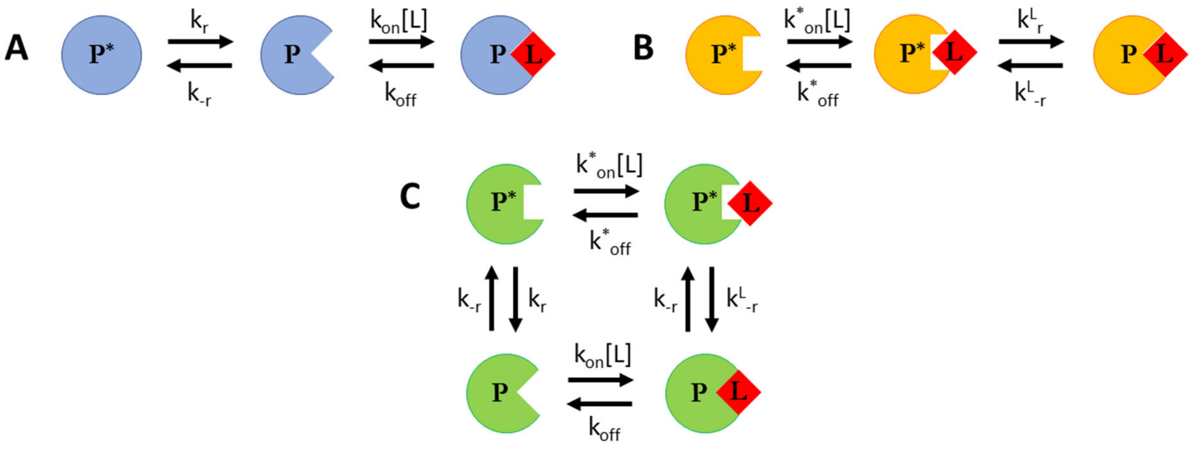


Figure 1.4 Mechanisms of protein-ligand binding. *(A) Conformational selection. (B) Induced fit. (C) Linkage scheme: conformational selection and induced fit are the extreme case; P = protein; L =Ligand. Modified from (Vogt et al. 2014).*

Considering the fact that protein-ligand interactions often involve changes in protein conformation and structure, two major models have been proposed to account for these changes: conformational selection (Figure 1.4A) and induced fit (Figure 1.4B). According to the conformational selection model, the protein exists in multiple conformations in equilibrium, and ligand binding occurs when the ligand selectively stabilizes one of these pre-existing conformations. In contrast, the induced fit model proposes that the protein undergoes a conformational change upon ligand binding, leading to a tighter fit between the two molecules. In this model, the initial interaction between the protein and ligand is weaker, and the full binding affinity is achieved only after the conformational change. Differently to the lock-and-key model, both conformational selection and induced fit models require an additional step which involve the transition between different conformational states.

The current view considers conformational selection and induced fit as the extremes of a more general scheme (Figure 1.4C): the protein exists as multiple conformations capable of interacting with the ligand; after binding, the ligand may change the relative stabilities of the two conformations or may alter the barrier for conversion between the two bound conformations (Vogt et al. 2014).

1.6 Nuclear Magnetic Resonance (NMR) Spectroscopy and CYANA structure calculation

There are three primary techniques for determining macromolecular structures at atomic resolution: cryo-electron microscopy (cryo-EM), X-ray crystallography, and nuclear magnetic resonance (NMR) spectroscopy. These techniques differ in the size of the molecules that can be studied. Cryo-EM is best suited for larger macromolecular complexes, while X-ray crystallography can study macromolecules of any size, but requires the crystallization of the sample, which is not always achievable. NMR spectroscopy is unique in that it allows for the study of macromolecules in solution, allowing the determination of their structure as well as characterization of their dynamic behavior. This technique requires the presence of NMR active molecules with spin-½ nuclei.

In the biological context, ^{1}H and ^{31}P isotopes are naturally abundant, while ^{13}C and ^{15}N must be intentionally enriched. For proteins, this is commonly achieved by growing *E. coli* cells in a minimal medium containing $^{15}NH_4Cl$ as the nitrogen source and $^{13}C_6$-glucose as the sole carbon source. RNA is produced via *in vitro* transcription using ^{15}N and ^{13}C labeled NTPs.

In NMR, the signal-to-noise ratio is directly proportional to the number of observed spins and the gyromagnetic ratio of the excited and detected spins (Lundstrom, Ahlner, and Blissing 2012). Due to its high gyromagnetic ratio and prevalence in biological macromolecules, protons are the most sensitive NMR probes. However, the high density of ^{1}H atoms in biological molecules often results in significant signal overlap, which can limit the utility of NMR investigations. Multidimensional heteronuclear NMR provides improved spectral resolution by correlating proton spins with those of heteroatoms such as ^{15}N, ^{13}C, and ^{31}P, allowing for the study of larger molecular systems (up to 100 kDa).

1.6.1 Chemical shift differences

Within an NMR spectrometer, NMR-active nuclei undergo precession at a specific frequency (called the Larmor frequency), which depends on the external magnetic field and the gyromagnetic ratio of the nucleus. The chemical shift (CS) is defined as the difference in resonance frequency (v_i) between a nucleus in a molecule and a reference nucleus in a standard compound, expressed in parts per million (ppm) of the operating frequency. The chemical shift is influenced by different factors, including the electronic environment around the nucleus, the hybridization and geometry of the surrounding atoms, and the presence of nearby magnetic or electric fields. For example, the electron densities of nuclei in residues of the same amino acid differ if the protein is folded, leading to unique resonance positions or chemical shifts in the NMR spectrum.

For its sensitivity, the CS is used to investigate conformational changes or to compare different forms of the same protein (e.g., individual domains vs. full-length protein) or binding events by mapping chemical shift perturbations performing titration experiments (Williamson 2013). The most common protein spectrum is the ^{1}H-^{15}N Heteronuclear Single Quantum Coherence (HSQC), which identifies cross-peaks of resonances associated with the protein backbone's amides (Waudby et al. 2016). Typically, the backbone amide signal is used and chemical shift differences and chemical shift perturbations are determined using the formula

$$\Delta CS = \sqrt{(\delta\mathrm{HN})^2 + (\delta\mathrm{N}\,/\,6.51)^2}$$

where δHN is the chemical shift difference of the amide proton and δN represents the chemical shift difference of nitrogen atom.

In addition, the CS difference between two states is influenced not only by the nature of the two species but also by the spectrometer frequency and experimental conditions (e.g temperature and salt concentration).

1.6.2 Spin relaxation

NMR is frequently used for the investigation of dynamics processes. One useful application is studying the spin relaxation, which provides important information about general tumbling time, mobility, structural constraints, and exchange.

The macroscopic magnetization is a result of all observable spins present. After excitation, the longitudinal (z) and transverse (x, y) components of the magnetization return to equilibrium with time constants T_1 and T_2, respectively. The relaxation of the magnetization is influenced by fluctuating fields generated by global and internal motion. Therefore, the dynamical behavior of individual spins influences the macroscopic magnetization, and T_1 and T_2 measurements provide insight into protein motion. Assuming the macromolecule of interest has a spherical shape, the correlation time (τ_c, the time required for a 1-radian rotation) can be estimated from the T_1 and T_2 time constants to obtain information of global motion using

$$\tau_c = \frac{1}{4\pi\nu_N}\sqrt{6\frac{T_1}{T_2} - 7}$$

where ν_N is the ^{15}N resonance frequency**.** As a general rule, the correlation time (τ_c) of globular molecules tends to increase linearly with their size and can be also affected when different molecules interact (Cavanagh et al. 2007).

Another relaxation mechanism is the nuclear Overhauser effect (NOE), a dipolar interaction that consists in magnetization transfer between neighboring spins where the transfer efficiency has a $1/d^6$ dependency (d is the distance between the spins). For this reason, NOE is a powerful source both for structural information and for dynamics information. ^{15}N heteronuclear NOE experiments (hetNOE) is used to measure the dynamics of the backbone of a protein. By measuring the NOE of the nitrogen nucleus, information about the dynamics of the amide bond between the nitrogen and the adjacent carbon atom can be obtained. If the nitrogen atom is surrounded by rigid structures, then the NOE will be large showing a value around 0.8. However, if the nitrogen atom is in a flexible region of the protein or if there is internal motion in the protein, the NOE will be smaller than 0.5.

1.6.3 Paramagnetic Relaxation Enhancement (PRE)

Short range distance experiments (NOE), which have a limit of 6 Å, are commonly used to study globular proteins due to the short inter-proton distances between residues that are far apart in the linear amino acid sequence (Clore and Gronenborn 1989). However, for macromolecules or multi domain proteins, short range distances are insufficient, and long-range information must be obtained (Clore and Gronenborn 1998b, 1998a; Clore and Venditti 2013). Paramagnetic relaxation enhancement (PRE) is an experiment that provides long range information by detecting interactions between an unpaired electron of a paramagnetic site and protons up to ~25 Å away (Clore and Iwahara 2009).

The presence of a paramagnetic center in a protein (or other molecules), causes an enhancement of the signal intensity of the surrounding nuclei (Figure 1.5). This effect arises from the magnetic dipolar interaction between the unpaired electron spins of the paramagnetic center and the neighboring nuclear spins. In the case of a protein, the interaction leads to an increase in the rate of spin relaxation of the protein nuclei, which can be quantified by measuring the relaxation rates of the protein protons with and without the addition of the paramagnetic species. The enhancement is inversely proportional to the distance between the nuclei of the protein and the paramagnetic center, with the effect being stronger at shorter distances (Figure 1.5C).

The paramagnetic center can be an intrinsic component of the system, such as metalloproteins, or an extrinsic component that is added by chemical reactions. There are two classes of paramagnetic labels: nitroxide free radicals (Kosen 1989) and suitable metal ions, such as Mn(II), Cu(II), or Gd(III), chelated to EDTA (Iwahara et al. 2003). Site-directed spin labeling of proteins is generally achieved by conjugating the paramagnetic tag to a surface-exposed cysteine residue introduced by site-directed mutagenesis, to form a disulfide bond between the cysteine and the paramagnetic tag (Altenbach et al. 1990).

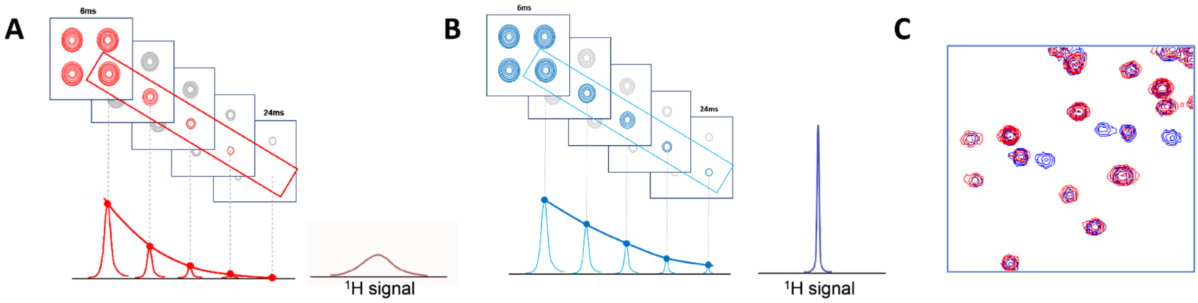


Figure 1.5 Paramagnetic Relaxation Enhancement (PRE). *(A) ^{1}H relaxation in presence of the paramagnetic center. (B) ^{1}H relaxation in absence of the paramagnetic center. (C) Overlay of ^{1}H-^{15}N HSQC spectra of the sample in presence of the paramagnetic center (red) and in absence of the paramagnetic center (blue).*

The PRE is measured by taking the difference in nuclear relaxation rates between the paramagnetic sample (Figure 1.5A) and a diamagnetic control (Figure 1.5B). Longitudinal relaxation (Γ_1) is the rate at which the nuclear spins return to their equilibrium state along the direction of the external magnetic field (the z-axis) while transverse relaxation (Γ_2) is the rate at which the nuclear spins lose coherence between their transverse components (x and y) due to interactions with their environment, including the paramagnetic center. While both longitudinal (Γ_1) and transverse (Γ_2) PRE rates can be measured, Γ_2 measurements provide the most reliable and accurate data (Clore and Iwahara 2009; Iwahara, Schwieters, and Clore 2004; Iwahara, Tang, and Marius Clore 2007). The distance between the paramagnetic center and the nucleus is calculated from Γ_2:

$$r = \left[\frac{K}{\Gamma_2}\left(4\tau_c + \frac{3\tau_c}{1 + {\omega_h}^2{\tau_c}^2}\right)\right]^{1/6}$$

where K is the magnetic susceptibility difference between the paramagnetic center and the surrounding medium, Γ_2 is the transverse relaxation rate of the nucleus, τ_c is the correlation time of the protein, ω_h is the Larmor frequency of the nucleus.

PRE data can be finally used in structural calculation to improve the accuracy and precision of protein structure determination. Since PRE data provide long-range distance restraints between the paramagnetic center and the protein residues, they can be used in combination with other structural data, such as NOEs, and residual dipolar couplings (RDCs), to calculate high-resolution protein structures. The distance restraints can be defined as upper (.upl) and lower (.lol) distance between the paramagnetic center and the protein residue, based on the magnitude of the PRE effect. It is pertinent to note that while utilizing PRE data as restraints for structural calculation, it is necessary to consider the presence of the paramagnetic center when the distance are calculated. In our specific instance, we employ dummy-atoms to delineate the location of the center of the spin-label cloud (Session 2.4.8.1).

1.6.4 Structure calculation using CYANA

CYANA (Combined assignment and dYnamics Algorithm for NMR Applications) is a program used to determine the three-dimensional structure of biomolecules such as proteins or nucleic acids using NMR data (Wurz et al. 2017). The input data used in CYANA include chemical shift assignments, NOESY (Nuclear Overhauser Effect Spectroscopy) peak lists, and constraints or restraints on the conformational parameters. Chemical shift assignments provide information on the resonance frequencies of individual atoms in a biomolecule, which can be used to determine their chemical environment and neighboring atoms. This information is used to assign a unique identifier to each atom in the biomolecule. NOESY peak lists provide information on the distances between pairs of atoms in a biomolecule, which can be used to generate distance constraints or restraints. These constraints specify the minimum and maximum distances between atoms that are consistent with the NOESY data. In addition to NOESY peak lists, other types of experimental data can also be used to generate additional constraints or restraints. CYANA employs a combination of simulated annealing and algorithms to generate a diverse set of conformations that satisfy the input restraints. The program iteratively improves the ensemble until it converges on a final set of structures that best fit the experimental data. The resulting structures are evaluated based on various criteria such as energy minimization, stereochemistry, and agreement with experimental data. The final output includes a set of structures that are representative of the biomolecule in solution, along with statistical information on the quality of the structure (Wurz et al. 2017).

On the other hand, Multistate CYANA, extends the capabilities of standard CYANA by allowing the calculation of the structures in multiple states or conformations (Strotz et al. 2017; Vogeli, Guntert, and Riek 2013). This is particularly useful in cases where a biomolecule exhibits dynamic, conformational flexibility, such as in the case of intrinsically disordered or multidomain proteins. The resulting set of conformations represents the ensemble of conformations that the biomolecule can adopt in solution. In summary, while standard CYANA calculates the three-dimensional structure of a biomolecule in a single conformation, Multistate CYANA extends this capability by allowing the calculation of the ensemble of conformations that a biomolecule can adopt in solution. Until now the multistate CYANA calculation has been used only on single domains to observe different conformations of disordered regions (e.g. loops) (Vogeli, Guntert, and Riek 2013; Okada et al. 2021). Multistate CYANA calculation will be covered more in detail in Session 2.2.5.

1.7 Electron Paramagnetic Resonance (EPR) Spectroscopy and MMMx modeling

Electron Paramagnetic Resonance (EPR) spectroscopy can provide valuable long-distance structural restraints for protein-RNA complexes in a manner similar to PRE experiments (Duss et al. 2015). This technique involves the use of one or two unpaired electrons (Gmeiner et al., 2017; Duss, Yulikov, et al., 2014). Like the PRE-NMR experiment, site-specific labeling can be used to introduce spin labels in both the protein and RNA of interest. Methanethiosulfonate (MTSL), iodoacetamidoproxyl (IAP) and Gd(III)-maleimido-DOTA are commonly used spin labels for probing protein and RNA.

Previous research has shown that EPR and NMR techniques complement each other in determining the structures of free proteins, protein-RNA complexes (Duss et al. 2014; Masliah et al. 2018), and in studying phase separation (Emmanouilidis et al. 2021).

1.7.1 DEER experiments

The distances between two paramagnetic labels is determined using four-pulse double electron-electron resonance (DEER) also called pulsed electron electron double resonance (PELDOR) (Jeschke 2012; Schiemann et al. 2007). DEER experiments are conducted on frozen glassy samples, which results in distance distributions instead of single distances, since the sample include all the different orientation of the individual molecules with respect to the magnetic field. The width of the distribution reflects the flexibility of the object, and the shape reflects the conformational distribution. In addition, a broad distribution indicates a smooth and slow dipolar evolution signal decay, whereas a narrow distribution indicates clear oscillations (Jeschke 2013).

The primary data obtained from a DEER experiment is a dipolar evolution curve, which represents the time-dependent changes in the dipolar interaction between the unpaired electrons of two spin-labeled molecules (Figure 1.6A). The dipolar evolution curve is then processed to obtain the form factor, which is a mathematical representation of the distance distribution between the two spin labels (Figure 1.6B). The form factor is then transformed into a distance distribution using the Tikhonov regularization method, which is a mathematical algorithm that allows the inversion of the form factor to obtain the distance distribution (Figure 1.6C). The distance distribution represents the probability density of finding the two spin labels at a given distance, and is typically presented as a histogram or a probability density plot. The distance distribution can then be analyzed to obtain information about the structure and dynamics of the protein or their complexes (Jeschke 2012; Schiemann et al. 2007) in combination with other experimental techniques.

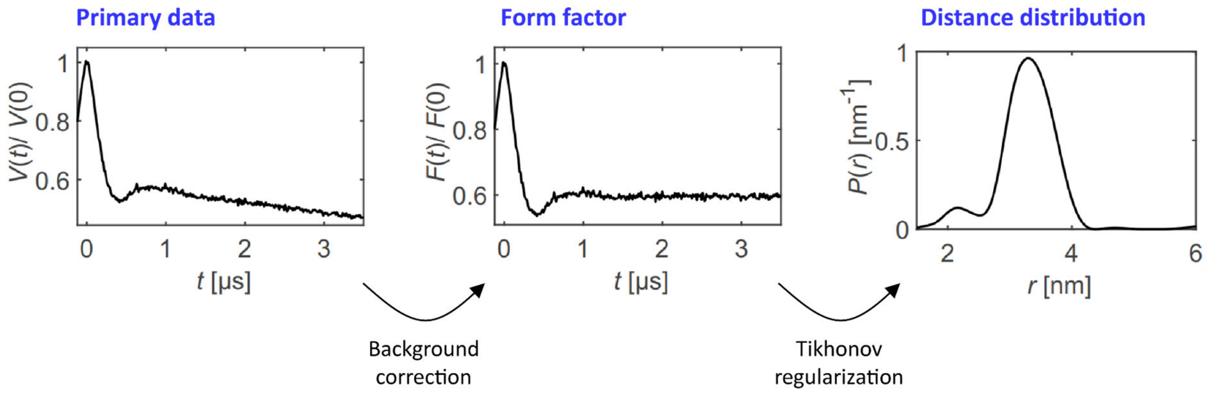


Figure 1.6 The EPR-DEER experiment. *(A) Normalized primary DEER data V(t)/V(0). (B) Form factor F(t)/F(0) obtained by division with the background function. (C) Corresponding distance distribution. Modified from (Esteban-Hofer 2022)).*

1.7.2 MMM modeling

Multiscale Modeling of Macromolecules (MMM) is a Matlab-based open-source modeling toolbox for the structural characterization of proteins and their complexes (Figure 1.7). The method combines data from various experimental techniques, with a focus on distance distribution restraints obtained EPR experiments. The approach, developed by Prof. Gunnar Jeschke of ETH Zürich, has been extensively reviewed in several publications (Jeschke 2016, 2018, 2021; Jeschke and Esteban-Hofer 2022).

RigiFlex is the tool that is used to obtain the raw ensemble and it uses mainly EPR restraints (DEER). The approach involves dividing the system into rigid bodies (folded domains) and flexible sections (linkers and loops). Each rigid body is defined by three reference points, which are selected to form the largest equilateral triangle. By performing DEER measurements between the reference points of the rigid bodies, the *Rigi* module allows the placement of the rigid bodies relative to each other. Flexible peptide or nucleic-acid linkers are then inserted using the *Flex* and *FlexRNA* modules, and DEER measurements between the reference points of the rigid bodies and the flexible regions are also included.

Once a raw ensemble has been generated, EsnsembleFit step is performed to obtain the representative ensemble. This step can include data for different techniques. A subset of conformers from the initial ensemble is selected for optimization. This subset should contain enough conformers to adequately represent the experimental data but should also be small enough to avoid overfitting.

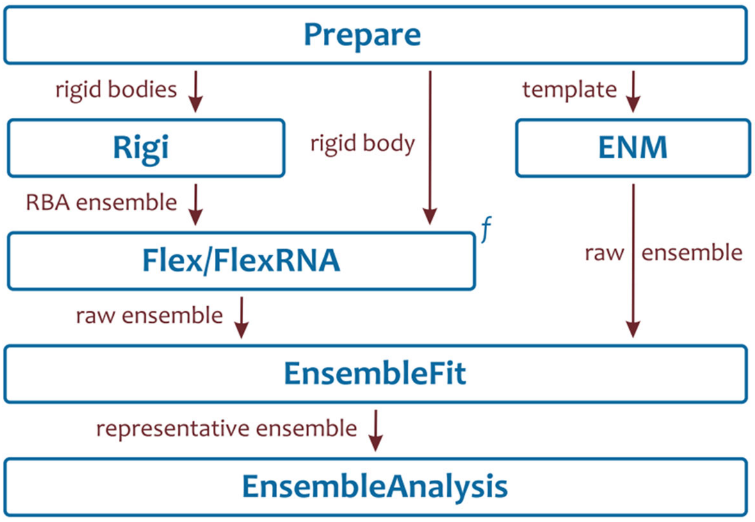


Figure 1.7 Schematic visualization of MMMx pipelines. *(Jeschke and Esteban-Hofer 2022)*

The selected conformers are optimized to minimize the difference between the experimental data and the calculated data for each conformer. The optimized ensemble is evaluated to ensure that it accurately represents the experimental data. This involves performing the backcalculation analysis and compare the calculated data with the experimental data for each conformer in the ensemble. Depending on the fitting a probability (population or statistical weight) is assigned to each conformer in the representative ensemble. The EnsembleFit step is a critical part of the MMMx modeling process as it confirms that the ensemble generated by *RigiFlex* accurately represents the experimental data.

1.8 Liquid-Liquid Phase Separation and Membraneless organelles

Eukaryotic cells are comprised of two distinct types of organelles: membrane-bound organelles and membraneless organelles. The first is surrounded by a membrane that can have different microenvironments suitable for their specific functions, while the latter lacks an enclosing membrane, thus facilitating the exchange of molecules (Figure 1.8A). Throughout the years, these compartments have been referred to by various names, including cellular bodies, nuclear bodies, non-membrane-bound compartments, granules, speckles, aggregates, assemblages and membrane puncta.

The first observation of a membraneless compartment within the neuronal cell nucleus in the 1830s, which was later named the nucleolus (Pederson 2011). Subsequently, numerous such compartments have been identified in the nucleus and cytoplasm of all eukaryotic cells. These compartments comprise many molecular components, including proteins and nucleic acids, which are capable of retaining stable concentrations within the structures for prolonged periods. However, many of these compartments can undergo molecular exchange with the surrounding environment (Phair and Misteli 2000; Weidtkamp-Peters et al. 2008).

The discovery that P granules in *C. elegans* germ cells exhibit liquid-like properties has provided a significant insight into the physical mechanisms underlying the formation of membraneless compartments; (Banani et al. 2017; Shin and Brangwynne 2017). The assembly of these condensates is mediated by the molecular process of liquid-liquid phase separation (LLPS), in which a supersaturated solution separates into two distinct phases, the condensed and dispersed phases (Figure 1.8B). Recent research has established the crucial role of LLPS in several essential biological processes, including gene expression and signaling pathways. LLPS also drives the assembly of all membraneless organelles such as nucleoli and stress granules, promoting the compartmentalization of cellular matter and facilitating spatiotemporal regulation of biological reactions (Peran and Mittag 2020; Martin et al. 2020).

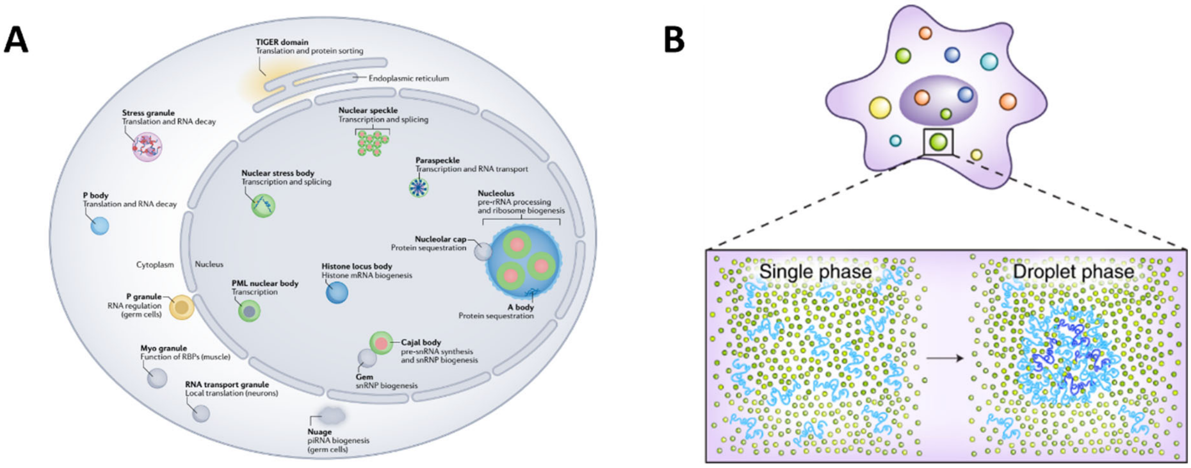


Figure 1.8 ***Overview of MLOs of eukaryotic cells and liquid–liquid droplets.*** *(A) Membraneless organelles (MLOs) in the nucleus and cytoplasm (Hirose et al. 2022). (B) Intracellular liquid–liquid phase separation (Tang 2019).*

Membraneless organelles exhibit a vast diversity in their physical characteristics, dimensionality (i.e., membrane-associated or soluble), molecular composition, subcellular localization, and functions. Among these organelles, there are ribonucleoprotein granules, which contain high concentrations of protein and RNA.

Proteins that contain large intrinsically disordered regions (IDRs) are the primary macromolecules that undergo phase separation under physiological conditions. IDRs lack a well-defined, folded structure, but frequently feature repeated sequence elements that facilitate weakly adhesive multivalent intermolecular interactions. These proteins are abundant in many biomolecular condensates, particularly those that concentrate RNA. IDRs typically exhibit low sequence complexity, with a limited number of amino acid types such as glycine, serine, glutamine, asparagine, phenylalanine, and tyrosine (Brangwynne, Tompa, and Pappu 2015). Additionally, some IDRs are enriched in charged residues such as lysine, arginine, glutamate, and aspartate. The lack of sequence diversity results in the presence of multiple Gly/Ser-

Phe/Tyr-Gly/Ser sequences, poly-Gln and poly-Asn tracts, as well as blocks of positive or negative charge in these molecules, which are critical for their targeting to RNA and for in vitro and in vivo phase separation. Another critical factor that regulates the phase separation of certain proteins in membraneless organelles is their post-translational modifications, like the phosphorylation of the RS domain in SRSF1, which regulates the localization of SRSF1 in nuclear speckles.

1.8.1 Diffusion experiments

Molecular diffusion (or simply diffusion) refers to the movement of all particles. The rate of diffusion is influenced by several factors including the temperature, particle size and shape, and viscosity of the fluid. Diffusion can occur among various types of particles, such as ions, molecules, intermolecular complexes, organometallic complexes, and micelles, and can be studied using NMR.

The use of pulsed field gradients in NMR spectroscopy allows for the measurement of the diffusion rates of nuclear spins. The technique, known as Self-Diffusion (SD)-NMR or Diffusion Ordered SpectroscopY (DOSY), enables the spectroscopic resolution of different compounds in a mixture based on their varying diffusion coefficients, which are dependent on the size and shape of the molecules. The general principle is that smaller molecules diffuse faster, while larger molecules diffuse more slowly.

The pulse program of a DOSY experiment is shown in Figure 1.9.

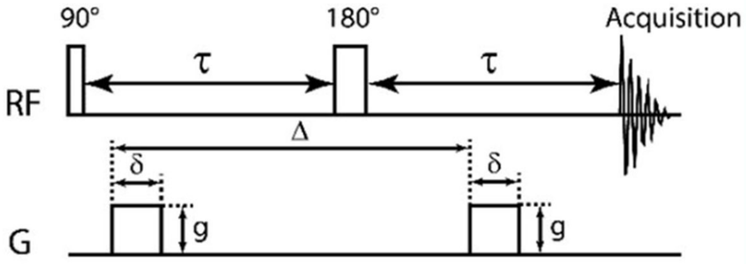


Figure 1.9 DOSY pulse sequence. *Modified from: https://nmr.chem.ucsb.edu/education/*

In this experiment, a 90-degree pulse flips the spins into the transverse plane. The position and phase of the molecules are then encoded by a gradient pulse depending on their position in the sample tube. A 180-degree pulse is applied to reverse the phase changes from the first gradient pulse, which are then nullified by the second gradient pulse, unless the spins have diffused changing position over the diffusion time (Δ). The measured signal is the integral over the whole sample volume and the intensity is attenuated depending on the diffusion time Δ and the gradient parameters (g = gradient strength, δ = delay).

The experiment is repeated multiple times, incrementing the gradient strength and keeping the delays constant, and a plot is made of the signal intensity against the gradient strength. The value of the diffusion coefficient is obtained by fitting the signal attenuation curve using the equation

$$I = I_0\, exp(-\delta^2 g^2 \Delta D \gamma^2)$$

Where I is the observed intensity, I_0 is the reference intensity, δ is the delay, g is the gradient strength, Δ is the diffusion time, D is the diffusion constant, and γ is the gyromagnetic ratio

Considering that diffusion depends on the size and shape of a molecule, DOSY experiments are a useful technique to investigate differences in diffusion between the dispersed phase and droplet state in the context of phase separation. This aspect is discussed in Section 3.2.2.

1.9 Objectives and Overview

1.9.1 Chapter 2: Structural investigation of SRSF1 tandem RRMs

The second chapter focuses on the structural characterization of SRSF1 tandem RNA-Binding Domains (SRSF1 RRM1+2) in the free state and in complex with two different RNAs (5′-UCAUUGGAU-3′ and 5′-UGGAUUUUUCAU-3′) designed after performing SELEX experiments. We applied an integrative structural modeling approach using restraints from NMR, and EPR to obtain a structural ensemble of the protein and protein-RNA complexes. First, *RigiFlex* modeling was used to calculate the ensembles using EPR (DEER experiments) as main restraints and NMR (PRE experiments) only for the ensemble fitting.

In this work, we combine the new multistate CYANA method with the MMMx refinement step to obtain ensembles using PRE data as main restraints and then understand the molecular interactions.

The ensembles obtained suggest that SRSF1 tandem RRMs have preferred conformations in the free form that can facilitate the interaction with RNA accommodating both the two RNAs designed from the SELEX experiments through both conformational selection and induced fit.

1.9.2 Chapter 3: SRSF1 and phase separation

In the third chapter of this project, we present the initial findings regarding SRSF1 and its role in phase separation. SRSF1 is present in membraneless organelles, specifically nuclear speckles located within the nucleus. A key aspect of this part of the project involved removing the GB1 tag, which was previously used to improve protein solubility. We optimized the protocol to obtain the protein without the GB1 tag and then examined the truncated protein in the context of phase separation in the presence of RNA. Microscopy analysis, turbidity measurements, and NMR Diffusion-ordered spectroscopy (DOSY) experiments were employed in our investigation. Our results indicate that the protein has a tendency to undergo phase separation *in vitro*. However, we found that the addition of RNAs, which include the binding sequences for the two RRMs, resulted in the dissolution of the droplets. Further experiments are ongoing to study the full-length protein and its modifications, such as phosphorylation of the RS domain, and their impact on the phase separation mechanism.

Chapter 2: Structural investigation of SRSF1 tandem RRMs

Structural investigation of SRSF1 tandem RRMs

Abstract

Serine-arginine (SR) rich proteins form an RNA-binding protein family which is involved in multiple steps of RNA metabolism including the regulation of constitutive and alternative splicing events. SRSF1 was the first member of the family to be identified; it is composed of two tandem RNA Recognition motifs (RRMs) and the RS domain enriched in arginine and serine at the C-terminal. The first RRM (or RRM1) is canonical while the second (or RRM2) is a pseudo-RRM and they are connected by a flexible and disordered linker domain.

Previous studies performed in our laboratory showed that RRM1 binds CA/CG motifs using the canonical β-sheet interface, while RRM2 binds GGA motif using the α_1-helix. In addition, a bimodal mode of interaction of SRSF1 tandem RRMs with RNA has been shown; more specifically, SELEX experiments demonstrated that RRM1 binds the cytosine located at -4 or +6 from the GGA motif recognized by RRM2.

Here, we aim to combine nuclear magnetic resonance (NMR) and electron paramagnetic resonance (EPR) to determine the structural ensemble of the tandem RRMs (SRSF1 RRM1+2) in the free state and in complex with two RNAs (5′-UCAUUGGAU-3′ and 5′-UGGAUUUUUCAU-3′) designed based on the SELEX experiments. In the present study we show the first application of the Multistate CYANA calculation method on a multidomain and dynamic system which was used previously only on single globular domains. In addition, we refined the ensemble including the EPR data, using the EnsebleFit step from the MMMx. The ensembles obtained indicate that SRSF1 tandem RRMs do not behave as independent domains but they already show preferred conformations in the free state; we observed that these conformations can promote the binding to the 5′-UCAUUGGAU-3′ RNA via conformational selection and to the 5′-UGGAUUUUUCAU-3′ RNA via both conformational selection and induced fit.

Author contributions

Cristina K. X. Nguyen (C.K.X.N.) prepared protein samples, performed NMR experiments, and performed the structure calculation using the hybrid modeling. Laura Esteban-Hofer (L.E.H.) prepared labeled RNA samples, performed and analyzed the EPR experiments, and generated the *RigiFlex* ensembles. Antoine Cléry (A.C.) performed preparatory work. A.C. and Fred Franz Damberger (F.F.D) assisted with the design and analysis of the NMR experiments and CYANA calculation. Gunnar Jeschke (G.J.) developed the MMMx software package. A.C. and Maxim Yulikov (M.Y.) provided guidance. G.J. and Frédéric H.- T. Allain (F.H.-T.A.) provided financial support and overall supervision of the project. C.K.X.N. wrote the text with input from L.E.H. and contributions from all the involved authors.

2.1 Introduction

2.1.1 SRSF1

Serine Arginine Splicing Factor 1 (SRSF1 or SF2/ASF) is the prototype and the first discovered member of the SR proteins family. It contains a canonical RNA-Recognition-Motif (RRM1) and a pseudo-RRM (ψRRM2 or RRM2) at its N-terminal, followed by the RS domain at its C-terminal part, as all the other SR proteins (Figure 1.2A, Figure 2.1A). The RRM domains consist of approximately 90 amino acids and adopt a β_1–α_1–β_2–β_3–α_2–β_4 secondary structure arrangement, which forms a four-stranded β-sheet packed against the two α-helices; the loops between secondary structure elements can vary in length. The canonical RRM1 contains conserved aromatic residues on the β_1 and β_3 strands that are essential for RNA binding, while the pseudo-RRM2 lacks these residues and instead contains a conserved heptapeptide (SWQDLKD) on the α_1-helix that is responsible for RNA interaction.

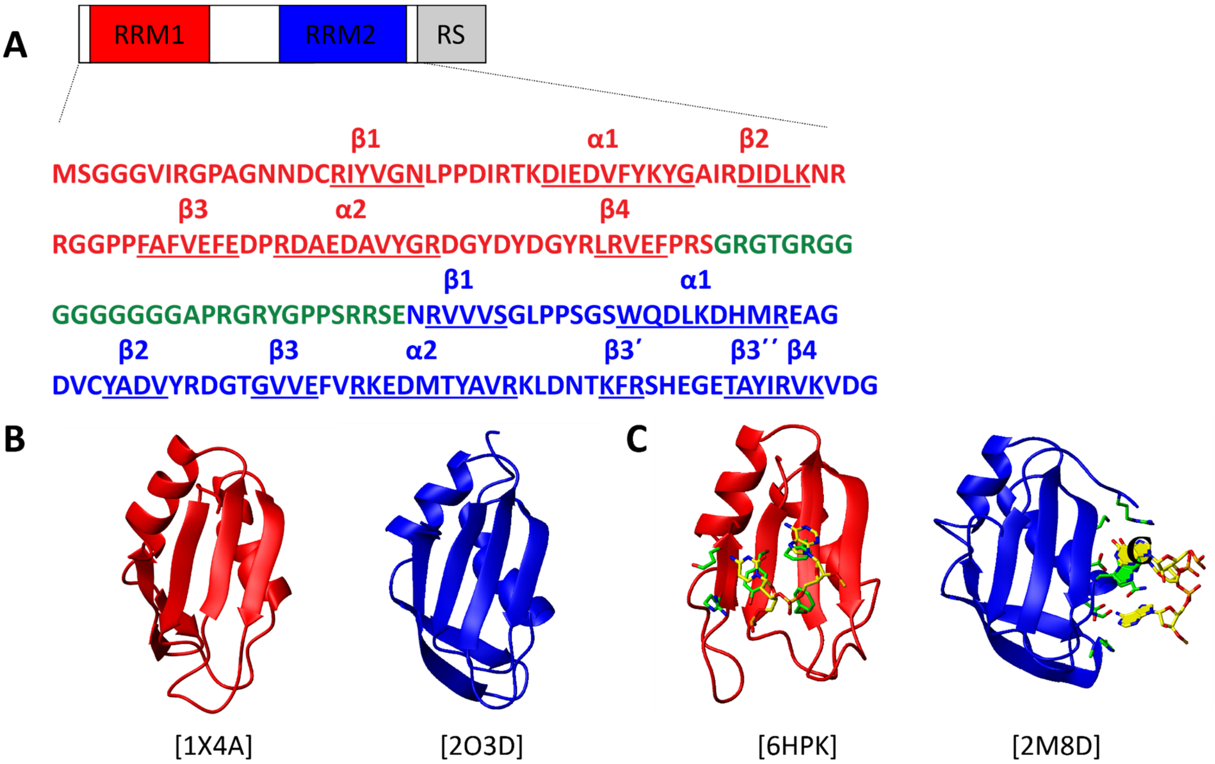

A
RRM1
RRM2
RS
β1
α1
β2
MSGGGVIRGPAGNNDCRIYVGNLPPDIRTKDIEDVFYKYGAIRDIDLKNR
β3
α2
β4
RGGPPFAFVEFEDPRDAEDAVYGRDGYDYDGYRLRVEFPRSGRGTGRGG
β1
α1
GGGGGGGAPRGRYGPPSRRSENRVVVSGLPPSGSWQDLKDHMREAG
β2
β3
α2
β3′
β3′′
β4
DVCYADVYRDGTGVVEFVRKEDMTYAVRKLDNTKFRSHEGETAYIRVKVDG
B
C
[1X4A]
[2O3D]
[6HPK]
[2M8D]

Figure 2.1 SRSF1 sequence and the RRM structures. *(A) SRSF1 comprises two RRMs (blue and red), connected by a flexible and disordered linker and an RS domain at the C-terminal (grey). The sequence of SRSF1 RRM1+2 is shown (1-196); amino acid numbering is according to the PDB sequence. Amino acids involved in the formation of β-strands and a-helices are underlined. (B) Structures of the isolated RRM1 (red) and RRM2 (blue) in the free form (He et al. 2005; Tintaru et al. 2007). (C) Structures of the isolated RRM1 bound to CA (red) and RRM2 bound to GGA (blue) (Clery et al. 2013; Clery et al. 2021). The PDB accession codes are indicated at the bottom of each structure.*

The structures of RRM1 and RRM2 in free form (Figure 2.1B; PDB number: 1X4A for RRM1 and 2O3D for RRM2; (He et al. 2005; Tintaru et al. 2007)) and in complex with

RNA (Figure 2.1C; PDB number: 6HPK for RRM1 and 2M8D for RRM2; (Clery et al. 2013; Clery et al. 2021)) have been solved. RRM1 binds preferentially to CN (N stands for any nucleotides) motifs using its canonical β-sheet surface, with the largest extent of chemical shift perturbations (CSPs) observed for CA dinucleotide (Clery et al. 2021). The pseudo-RRM2 binds GGA motifs using the α_1-helix. Although the heptapeptide is conserved in other pseudo-RRMs, they show a lower affinity for the GGA motif than SRSF1, possibly due to the lack of conservation of Arg117 and His183 that interact with RNA or different amino acid environments around the binding site, or a combination of both (Clery et al. 2013). The two RRMs are connected by a glycine-enriched linker of 30 residues (89-119) that provides high flexibility. The linker sequence is not conserved among other SR proteins, suggesting that it may play a crucial role in binding specificity (Anko 2014; Tacke and Manley 1999). In fact, studies have shown that the linker is essential for binding to exonic splicing enhancers (ESEs) in a construct lacking the C-terminal RS domain and that the N- and C-terminal linker extremities cooperate with RRM2 during binding to ESEs, while the RRMs remain in proximity (Cho, Hoang, Chakrabarti, et al. 2011; Cho, Hoang, Sinha, et al. 2011).

2.1.2 Bimodal mode of interaction of SRSF1 tandem RRMs with RNA

A previous study conducted by our laboratory, which primarily investigated the interaction of RRM1 with RNA, yielded some results on the SRSF1 tandem RRMs in complex with RNA (Clery et al., 2021). The experimental construct utilized lacked the RS domain and contained two point mutations in RRM1, as well as an N-terminal GB1-tag for solubility purposes and a His6-tag for purification purposes (Figure 2.2A). The two point mutations, Tyr37Ser (Y37S) and Tyr72Ser (Y72S), not only improved the solubility of the protein, but also did not affect its ability to bind to RNA (Figure A1), as demonstrated in (Clery et al. 2021). This construct will be referred to as SRSF1 RRM1+2.

SELEX experiments were performed to understand how the two RRMs interact with RNA when linked by their natural interdomain linker, and if there is any sequence preference. The RNAs utilized in the SELEX experiments contained an invariant GGA motif in the middle, flanked by 12 degenerate nucleotides on both sides (Figure 2.2B). A clear enrichment in cytosines was observed at positions −4 and +6 of sequences containing a single GGA motif, whereas cytosines were primarily selected downstream of the GGANGGA motif (Figure 2.2C). This suggests that SRSF1 RRM1 can bind either upstream or downstream of the RRM2 binding site, when only a single GGA motif is present, which implies high flexibility of the inter-RRM linker. To validate this result, two short RNAs were designed containing the CA motif recognized by the RRM1 at positions −4 and +6 from the two ends of the GGA motif, respectively: 5′-U<u>CA</u>UU<u>GGA</u>U-3′ and 5′-U<u>GGA</u>UUUUU<u>CA</u>U-3′.

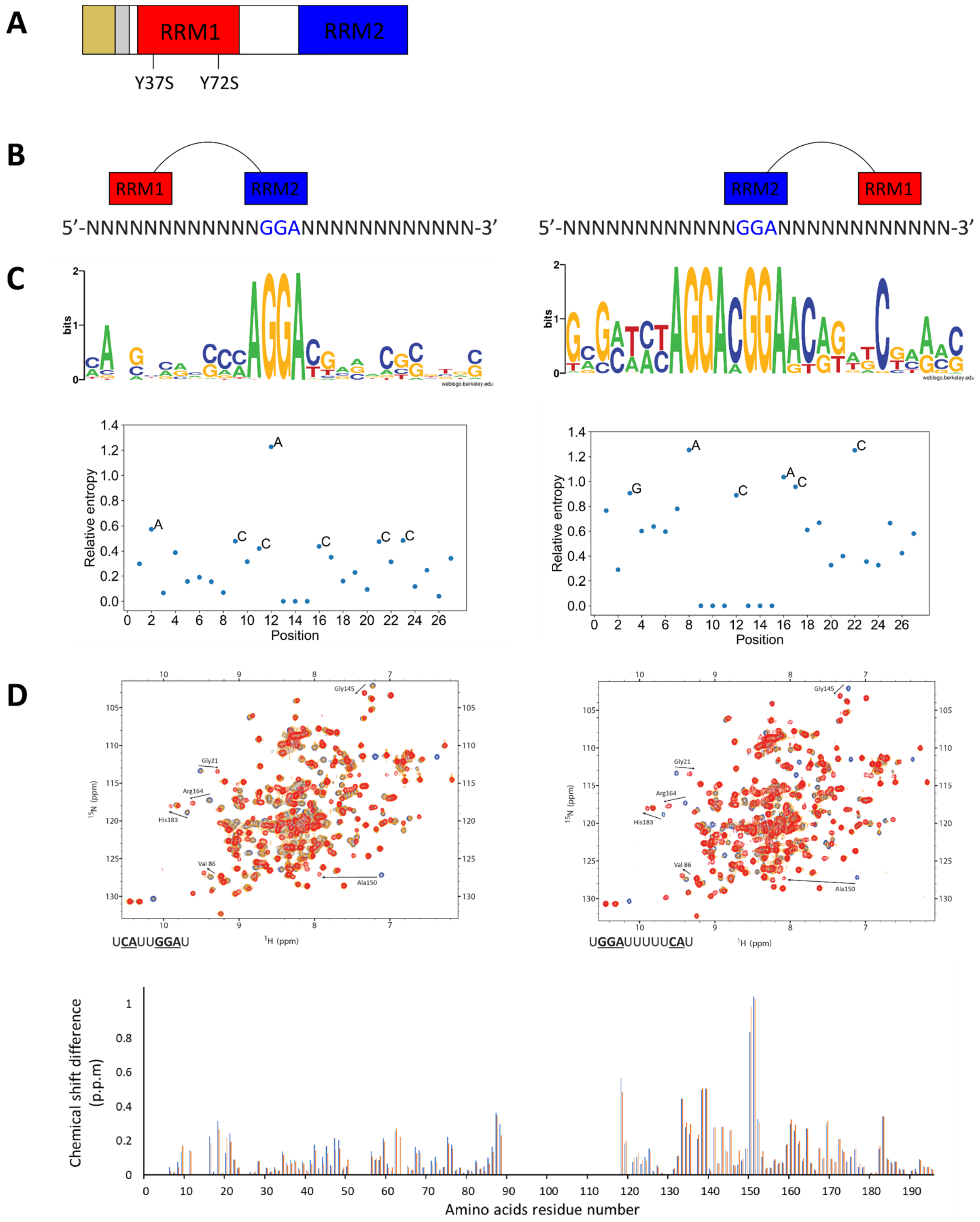


Figure 2.2 SRSF1 RRM1 interacts with RNA on both sides of RRM2-binding site. *(A) The SRSF1 RRM1+2 (1-196) construct used in our experiments: the protein contains two point mutations in RRM1 (Y37S and Y72S) and an N-terminal GB1-tag (yellow) followed by a His6-tag (grey) for solubility and purification purposes, respectively. (B) Schematic representation of the RNA molecules used for the SELEX experiment: invariant GGA nucleotides were inserted in the middle of the degenerated sequence to recruit SRSF1 RRM2 and allow the selection of potential RRM1 binding sites on both sides of the motif. (C) The sequences selected by SELEX with SRSF1 RRM1+2 followed two main patterns containing either the GGA or a GGANGGA motif. To determine the positions with the most stringent selection during SELEX, we calculated the relative entropy of position-dependent nucleotide frequency distributions of foreground and background sequences (lower panels). (D) Overlay of* ^{1}*H-*^{15}*N HSQC spectra and mapping of the combined CSPs. The spectra were measured with SRSF1 RRM1+2 free form (blue) and in the presence of 5'-UCAUUGGAU-3' or 5'-UGGAUUUUUCAU-3' RNA at 0.3:1 (orange) and 1:1 (red) protein:RNA ratios. SRSF1 RRM1 binds equally well the cytosine located at −4 or +6 from the GGA motif. Modified from (Clery et al. 2021).*

NMR titration experiments revealed that the saturation was reached at 1:1 protein:RNA ratio for both RNAs, and combined CSPs had similar intensities for both RRM1 and RRM2 amides, indicating that the two domains bind to a single molecule of RNA (Figure 2.2D). In good agreement with the NMR data, SRSF1 RRM1+2 had a similar affinity for the 5′-UCAUUGGAU-3′ and 5′-UGGAUUUUUCAU-3′ RNAs (K_d of 58 and 55 nM, respectively), which similarly decreased when the CA was mutated to UU in both RNAs (K_d of 164 and 145 nM, respectively). These data also suggest a cooperative mode of interaction of both RRMs with RNA, as the K_d values obtained were around 20 μM for RRM1 and 0.7 μM for RRM2. Taken together, these results demonstrate that RRM1 binds surprisingly equally well to RNA containing cytosines at two fixed positions (−4 and +6 of the edges of the GGA RRM2 binding site). However, this equal affinity appears to be lost when two consecutive GGA motifs are present in the sequence (Figure 2.2C). In fact, RRM1 only binds downstream in such cases, indicating that the number of GGA motifs not only increases the affinity of SRSF1 for RNA, but also influences the relative position of the two RRMs of SRSF1 on the RNA (Clery et al., 2021).

A recent study investigated SRSF1 tandem RRMs presented results that are partially in contrast with our previous findings (De Silva et al. 2022). According to this study, the main task of the linker is to allow flexibility for RNA binding without playing a key role in the binding itself. In addition, the flexibility of the linker allows considerable length variation in the spacer between the CA and GGA binding sites and this spacer can influence the affinity. The study concluded that SRSF1 has a preference for RNA sequences with shorter spacers, and that the RRM1 cognate motif is typically located upstream.

Through the use of mutagenesis and fluorescence polarization (FP) binding assays, the study demonstrated that deleting either RRM1 or RRM2 significantly reduces binding affinity, while mutating basic residues in the linker only slightly decreases RNA binding. The main role of the linker in RNA binding is to connect RRM1 and RRM2. The study also measured binding affinities to RNA sequences with different spacer lengths between the bipartite motifs (ranging from 0 to 10 nucleotides), as well as swapping the CA and GGA binding sites. The results revealed that RRM2 plays a dominant role in RNA-binding specificity of SRSF1. Although RRM1 prefers cytidine, it can bind to other nucleotides with lower affinity. SRSF1 favors RNA sequences with shorter spacer motifs and RNA sequences with the upstream RRM1 cognate motif. Using paramagnetic relaxation enhancement (PRE) experiments, the study labeled residue E120C (on the edge between the linker and RRM2) to investigate the SRSF1 tandem RRMs in complex with uuuCAuuGGAuu or uGGAuuuuuCAu RNA. The results indicated that paramagnetic perturbations are spread around RRM1, which cannot be explained by a single conformation, as the tandem RRMs do not have a fixed relative orientation. Furthermore, the perturbation patterns were different for SRSF1 bound to the two RNA sequences, suggesting that the domain arrangement is different when these bipartite motifs are swapped. Finally, Xplor-NIH simulations were performed, which demonstrated that the PRE data can only be fitted with a conformational ensemble rather than a single conformation. The study suggests that the binding diversity of SRSF1 is achieved through the flexibility of both the SRSF1 linker domain and the RNA backbones.

In our study we also conducted PRE experiments using different labeled sites on the protein, as elaborated in subsequent sections, to obtain restraints to use in the structure determination process. Furthermore, we employed various software to obtain the structure of SRSF1 tandem RRMs both in isolation and in complex with RNA (5′-UCAUUGGAU-3′ and 5′-UGGAUUUUUCAU-3′), with the aim of gaining insight into the underlying mechanism. The results obtained indicate that despite the apparent flexibility implied by the linker sequence, the two domains of SRSF1 are pre-arranged in favorable orientations in the free form, thereby facilitating RNA binding through either conformational selection or induced fit.

2.2 Results

2.2.1 Construct design and characterization of SRSF1 tandem RRMs

The protein SRSF1 contains two RNA recognition motifs (RRMs), RRM1 and RRM2, which are connected by a 30-residue linker. This low complexity domain is composed of a high percentage (48%) of glycine residues, suggesting that it has high flexibility. Although both RRMs, among the other SR proteins, recognize similar RNA sequences (C for RRM1 and GGN for RRM2), the linker sequence is not conserved in length or composition, suggesting that it plays a role in recognizing RNA in cell. To better understand the role of the linker in RNA binding, the structure of the tandem RRMs of SRSF1 was investigated using a truncated construct of SRSF1 lacking the RS domain (SRSF1 RRM1+2, residues 1-196) and containing a GB1 tag and two point mutations (Y37S and Y72S) to improve protein solubility (Clery et al. 2021).

First, the folding of the truncated construct was compared to that of the isolated RRMs, for which the structures were already solved (Figure 2.1B and Figure 2.3A). Using ^{1}H-^{15}N heteronuclear single quantum spectroscopy (HSQC) spectra, chemical shift differences were measured. The high percentage of glycines complicated the assignment of the linker in the NMR spectra causing overlapping of the signals but we could transfer the assignment of almost all the residues of the RRMs from the spectra of the individual domains (Clery et al. 2013; Clery et al. 2021). From the NMR spectra we conclude that the two RRMs show a similar folding when they are linked in tandem, some chemical shift differences were observed for the residues close to the linker of both RRMs. The RRM2 construct contains some linker residues (107-119) and, so in this case the chemical shift differences depend on a different linker arrangement.

To assess the flexibility of the linker, we conducted ^{15}N-spin relaxation experiments (Figure 2.3B). Such experiments leverage the ability of the ^{1}H-^{15}N backbone amide bond vector to reveal movements of the peptide chain. The range of motion within the protein chain is influenced by its structural state, with residues within secondary structures exhibiting limited motion (Morin 2011). Consistent with this, low hetNOE values were observed for residues located at the N-terminal of the protein (1-15) and the inter-RRMs linker region, whereas the hetNOE values for the folded regions of both RRMs were higher (around 0.7-0.8), indicating well-structured regions.

Additionally, given that the spectral overlay and chemical shift differences of the tandem RRMs of SRSF1 did not suggest independent behavior of the subdomains with respect to each other, we measured rotational correlation to confirm this hypothesis (Figure 2.3C). Sets of longitudinal T_1 and transverse T_2 relaxation times were measured and used to calculate τ_c (Section 2.4.5.3). Individual RRMs have a τ_c value of 7.5 ns and 6.8 ns respectively, calculated for the rigid parts (16-88 for RRM1 and 120-195 for RRM2) (Table 2.1).

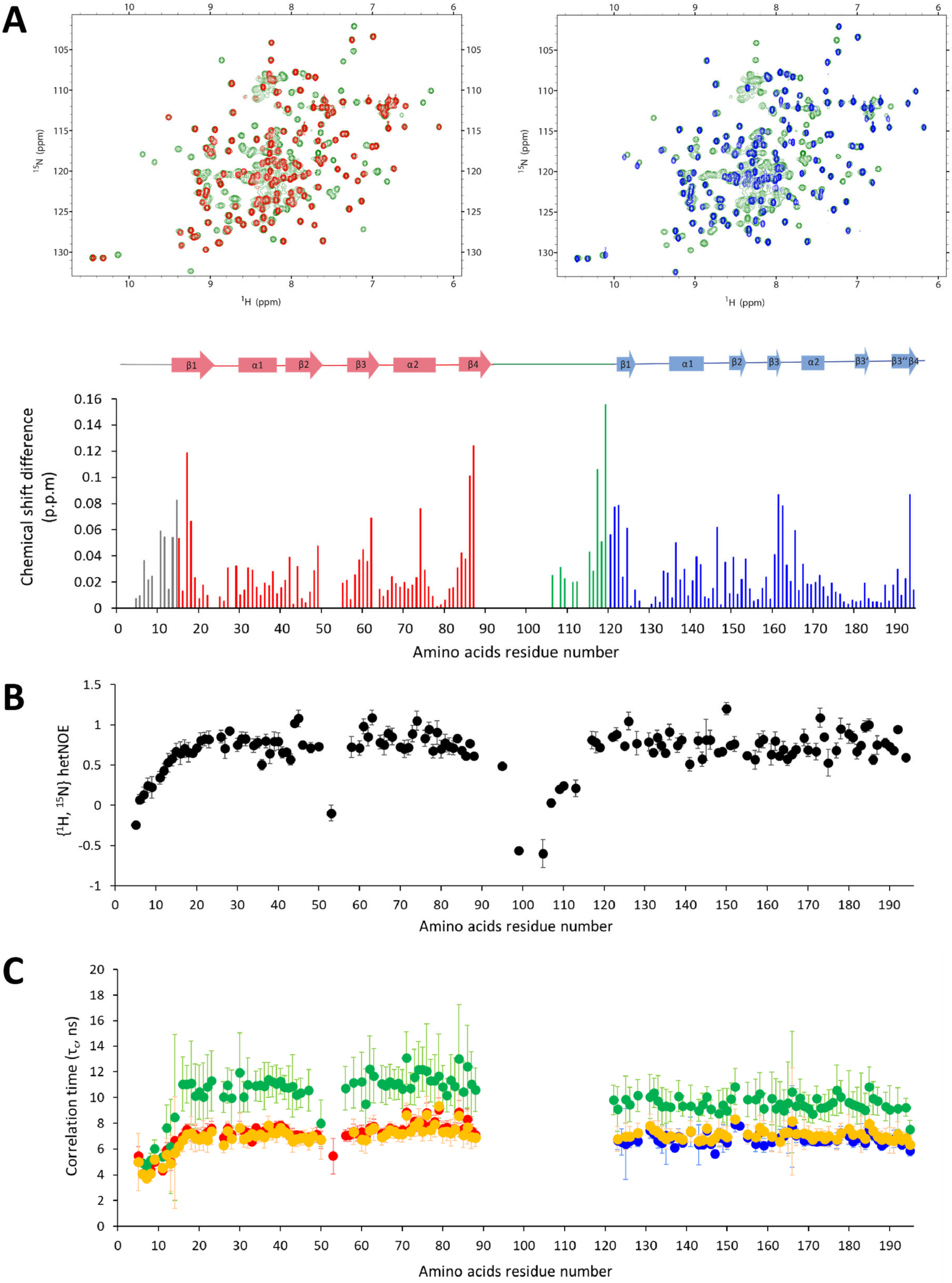


Figure 2.3 NMR characterization of SRSF1 RRM1+2. *(A) Overlay of ^{1}H-^{15}N HSQC spectra of SRSF1 RRM1+2 (residue 1-196, green) and RRM1 (residues 1-90, red) and RRM2 (residues 107-203, blue); the proteins contain an N-terminal GB1-tag followed by a His6-tag for solubility and purification purposes, respectively. The graph shows the chemical shift differences of SRSF1 RRM1+2 compared to individual. Gray denotes the N-terminal of RRM1 peaks, red folded RRM1 peaks, green linker peaks and blue corresponds to RRM2 peaks. Secondary-structure elements of the protein domain are displayed at the top of the graph. (B) HetNOE values of SRSF1 RRM1+2: RRM1 and RRM2 domains are well structured, while the linker is dynamic. (C) Rotational correlation times of SRSF1 RRM1+2 (green), RRM1 (red), RRM2 (blue), and isolated RRMs mixed in the same tube (yellow). Error bars denote standard deviations.*

The similar values obtained by the isolated RRMs can be explained by their similar size, as τ_c correlates with the molecular weight for globular domains (Lee et al. 2006). These values are slightly larger than expected (~6 ns, according to the approximation $\tau_c \approx$ MW x 0.6 ns/kDa), but still in and acceptable range (Gossert and Jahnke 2016). The same τ_c values were obtained analyzing the individual RRMs mixed in the same sample (7.2 ns for RRM1 and 7.0 ns for RRM2) suggesting that in these conditions (not in tandem) the two RRMs still behave as independent domains. A different result was obtained for SRSF1 RRM1+2 as we observed an increase of the τ_c value to 11 ns for RRM1 and 9.6 ns for RRM2. This indicates that when the two domains are connected by the inter-RRMs linker, they are not completely independent from each other. In addition, since τ_c values are lower than the sum of the two RRMs (~14 ns), we hypothesize that the linked domains do not have a strong interaction and a single conformation.

	RRM1 τ_c (ns)	**RRM2 τ_c (ns)**
RRM1	7.5 ± 0.5	-
RRM2	-	6.8 ± 0.7
RRM1+2	11 ± 1.9	9.6 ± 1.5
RRM1 + RRM2	7.2 ± 1.1	7.0 ± 0.8

Table 2.1 Rotational correlation time of SRSF1 constructs. *Rotational correlation times of RRM1 (16-88) and RRM2 (120-195) in the context of the isolated domains, both RRMs mixed in the same tube and SRSF1 RRM1+2.*

This result indicates that the linker is important for the regulation of the direct interactions between RRM1 and RRM2 or between the RRMs and the linker itself.

Rotational correlation time was also measured for the protein-RNA complexes (Figure 2.4 and Table 2.2), and a slight increase in τ_c values was observed for both RRMs (~11.5 ns for each domain). As for the free protein, this result indicates that when the protein is bound to the different RNAs the two RRMs are not in a fixed configuration but they still adopt different conformations, confirming the flexibility of the linker also in the protein-RNA complex.

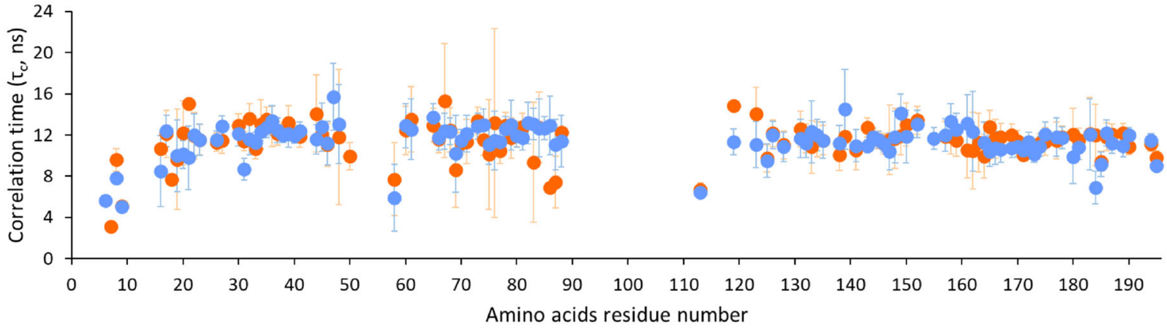


Figure 2.4 Rotational correlation time of protein-RNA complexes. *Rotational correlation times of SRSF1 RRM1+2 bound to 5'-UCAUUGGAU-3' (orange) and SRSF1 RRM1+2 bound to 5'-UGGAUUUUUCAU-3' (light blue).*

	RRM1 τ_c (ns)	**RRM2 τ_c (ns)**
SRSF1 RRM1+2 bound to 5′-UCAUUGGAU-3′	11.7 ± 2.4	11.5 ± 1.6
SRSF1 RRM1+2 bound to 5′-UGGAUUUUUCAU-3′	11.9 ± 1.8	11.4 ± 1.6

Table 2.2. Rotational correlation time of protein-RNA complexes. *Rotational correlation times of RRM1 (16-88) and RRM2 (120-195) in the context of SRSF1 RRM1+2 in complex with RNA: bound to 5'-UCAUUGGAU-3' RNA and bound to 5'-UGGAUUUUUCAU-3' RNA.*

2.2.2 Analysis of SRSF1 tandem RRMs using DEER

The protein samples utilized in the DEER experiments were prepared by myself (Section 2.4.3). Dr. Laura Esteban-Hofer, from the group of Prof. Gunnar Jeschke at ETH Zürich, performed the measurements and analysis of the DEER distance restraints. The methodology and detailed analysis are outlined in the PhD Thesis (Esteban-Hofer 2022), while we report the main results in this study.

2.2.2.1 Inter-RRM and RRM-linker distance distributions

The original plan was to determine the structure of the SRSF1 tandem RRMs in isolation and in complex with RNA utilizing *RigiFlex* modeling, a method developed by Prof. Gunnar Jeschke (ETH Zürich) (Jeschke 2016, 2018, 2021; Jeschke and Esteban-Hofer 2022). This approach divides the system into rigid bodies (such as folded domains) and flexible sections (such as linkers and loops). High-resolution structures of the rigid bodies are required, which can be obtained experimentally using various techniques, including X-ray crystallography, NMR spectroscopy, cryo-electron microscopy, or computational methods such as AlphaFold (Jumper et al. 2021). Each rigid body is defined by three reference points; Double Electron Electron Resonance (DEER) measurements, an Electron Paramagnetic Resonance (EPR)-based technique that provides information on a distance range of approximately 20 to 80 Å, was mainly used to obtain distances between inter-domain reference points. For each DEER experiment, two paramagnetic centers were required. The *Rigi* module allows the placement of the rigid bodies relative to each other. Flexible peptide or nucleic-acid linkers are introduced using the *Flex* and *FlexRNA* modules, and DEER measurements between the reference points of the rigid bodies and the flexible regions are also incorporated. The generated conformers are refined using YASARA (Krieger et al. 2009). Finally, the ensembles can be contracted to representative ensembles that best match the full shape of the distance distributions and other experimental techniques that provide mean value restraints.

To obtain the structure of SRSF1 tandem RNA recognition motifs (RRMs) using *RigiFlex* modeling, we treated the two RRMs as rigid bodies with a flexible inter-domain linker (Section 2.4.7). Since the protein does not contain any unpaired electrons, paramagnetic probes (MTSL or Gd(III)-maleimido-DOTA) were introduced via site-directed spin labeling of cysteine residues.

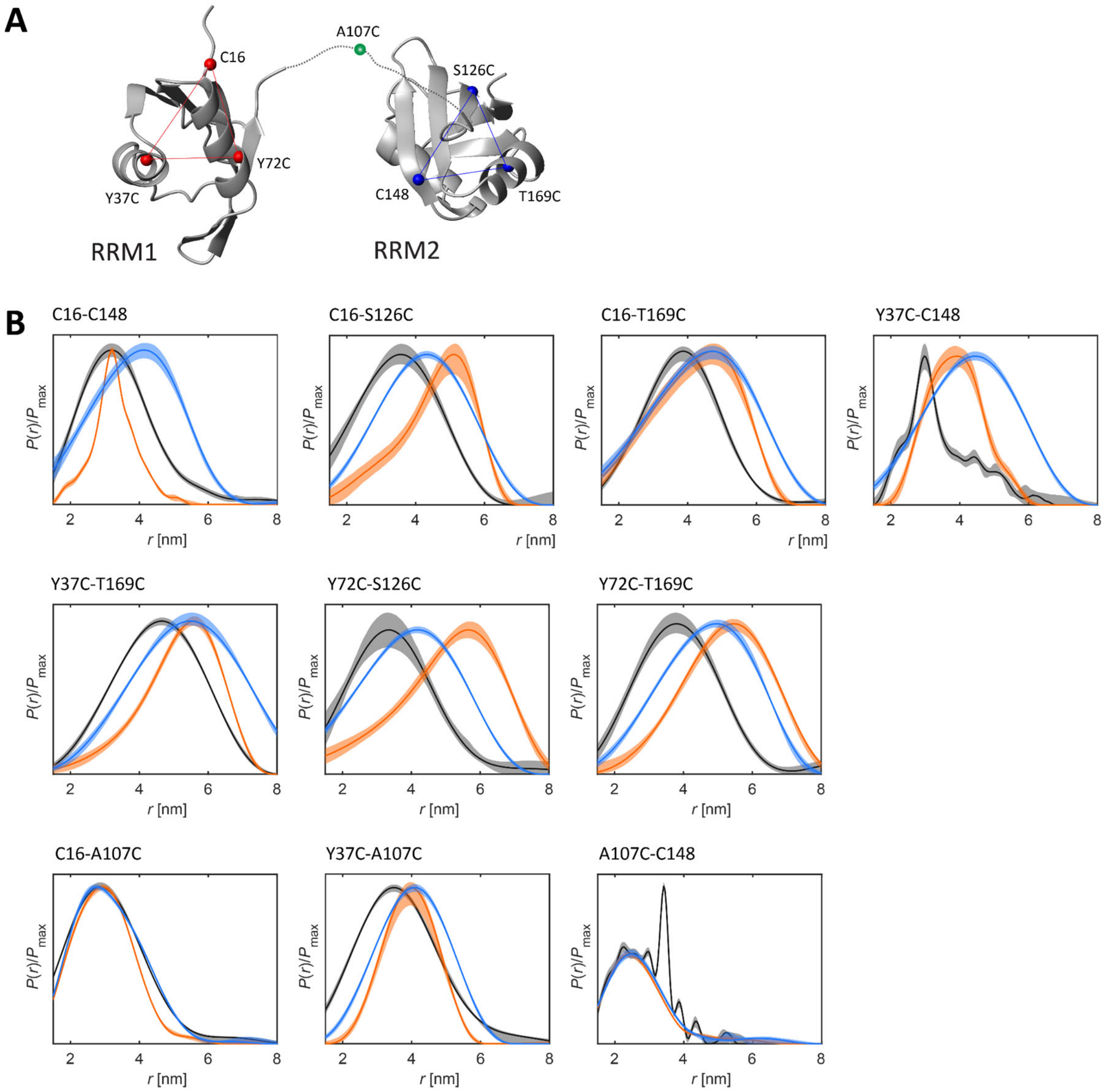


***Figure 2.5 Inter-RRM and RRM-linker distance distribution restraints**. (A) Schematic visualization of the sites involved in the acquisition of inter-RRMs or linker-RRMs distance restraints (RRM1 = red spheres, linker = green sphere, RRM2 = blue spheres). (B) Distance distributions of spin-labeled C16-C148, C16-S126C, C16-T169C, Y37C-C148, Y37C-T169C, Y72C-S126C, Y72C-T169C, C16-A107C, Y37C-A107C, and A107C-C148 respectively. Solid lines are the medians and shaded areas the 95% confidence intervals obtained by bootstrapping of 1000 samples. Black denotes the free protein, orange the protein in complex with 5'-UCAUUGGAU-3' RNA, and light blue the protein in complex with 5'-UGGAUUUUUCAU-3' RNA. The distance distribution of A107C-C148 free protein was normalized with respect to the broad feature to facilitate a comparison. Modified from (Esteban-Hofer 2022).*

Since for the *Rigi* module three reference points for each rigid bodies were needed, positions for site-directed spin labeling on the RRMs were chosen based on simulating the spin label attachment to existing NMR solution structures of individual RRMs of SRSF1 (He et al. 2005; Tintaru et al. 2007; Clery et al. 2013; Clery et al. 2021) using the software package Multiscale Modelling of Macromolecular systems (MMM) (Jeschke 2018). The following criteria were considered while scanning for potential labeling sites within the RRMs: potential mutation sites need to be accessible and in a folded region (α-helix or β-strand), the mutation to cysteine should not disrupt the structure of the protein, and the mutation should not affect the binding to RNA. Ideally, the three reference sites are positioned in a nearly equilateral triangle with the largest possible distance between the vertices. In addition to the native cysteines (C16 on RRM1 and C148 on RRM2) other two reference sites for each RRMs were chosen replacing the native residue in cysteine. During all experiments, a maximum of two cysteines were present at a time, requiring in some cases to mutate the native cysteines to the non-reactive amino acid alanine (e.g. C16 S126C C148A). From now we will omit these background mutations in the nomenclature and only explicitly indicate the sites directly involved in the distance measurements (e.g. C16-S126C). Seven of the nine possible combinations between the reference sites were selected: C16-C148, C16-S126C, C16-T169C, Y37C-C148, Y37C-T169C, Y72C-S126C, and Y72C-T169C (Figure 2.5A, Table A1).

For the *Flex* module, additional mutants were created to measure distance between RRMs and the linker. Alanine 107 was selected in the linker region since it does not show any influence in RNA binding and to not induce minimal perturbation on the linker due to the chemical similarity of the residues. Three distance restraints between the linker and the RRMs, involving the A107C linker mutation, C16 and Y37C in RRM1, and C148 in RRM2, were measured (Figure 2.5A, Table A1).
For all mutants created, DEER experiments were performed to characterize each combination of cysteines in the free form and in complex with two different RNAs. Once all mutants were expressed and labelled, Continuous-Wave (CW) EPR experiments were performed to confirm the labelling (data not shown) and DEER measurements were performed. An overview of the DEER results is shown in Figure 2.5B. All the inter-RRMs DEER distances presented differences between SRSF1 RRM1+2 in the free form and in complex with the two RNAs (Figure 2.5B upper part), except for the C16-C148 combination, where the free protein and the complex with the 5′-UCAUUGGAU-3′ RNA had a similar mean distance. The distance distribution of the free protein always presented the shortest distances, while it increased in the protein-RNA complexes. In two cases, the mean distance of protein-RNA complexes was similar (C16-T169C and Y37C-T169C); in some cases, the protein bound to 5′-UCAUUGGAU-3′ RNA showed longer distances (C16-S126C, Y72C-S126C and Y72C-T169); and in other cases, the protein bound to 5′-UGGAUUUUUCAU-3′ RNA showed longer distances (C16-C148 and Y37C-C148). The widths of the distance distributions were generally broad, confirming the freedom provided by the flexibility of the linker except for C16-C148 in complex with 5′-UCAUUGGAU-3′ RNA and Y37C-C148 in the free form. Based on these inter-RRMs DEER restraints, we could hypothesize that the RRMs are closer together in the free form and more distant when bound to RNA. The three distance distributions that involve the linker residue A107C (Figure 2.5B lower part) showed a very unexpected result: in fact, the RRM-linker distributions are strikingly similar for the free and RNA-bound states; only the mutant Y37C-A107 shows a small difference for the bound form with both RNA.

2.2.2.2 Protein-RNA and Intra-RNA distance distributions

In addition to the inter-protein restraints, we also collected distance distributions involving the RNA. Specifically, we measured distance distributions from the RNA extremities to specific locations in RRM1 (C16), the linker (A107C), or RRM2 (C148), producing a total of six possible combinations per protein-RNA complex (Figure 2.6A, Table A2). To make these measurements, we labeled the protein with Gd(III)-maleimido-DOTA and the RNA with IAP (Section 2.4.6). The protein samples utilized in the DEER experiments were prepared by myself (Section 2.4.3) while the labelling of the RNAs was performed by Dr. Laura Esteban Hofer (Section 2.4.4).

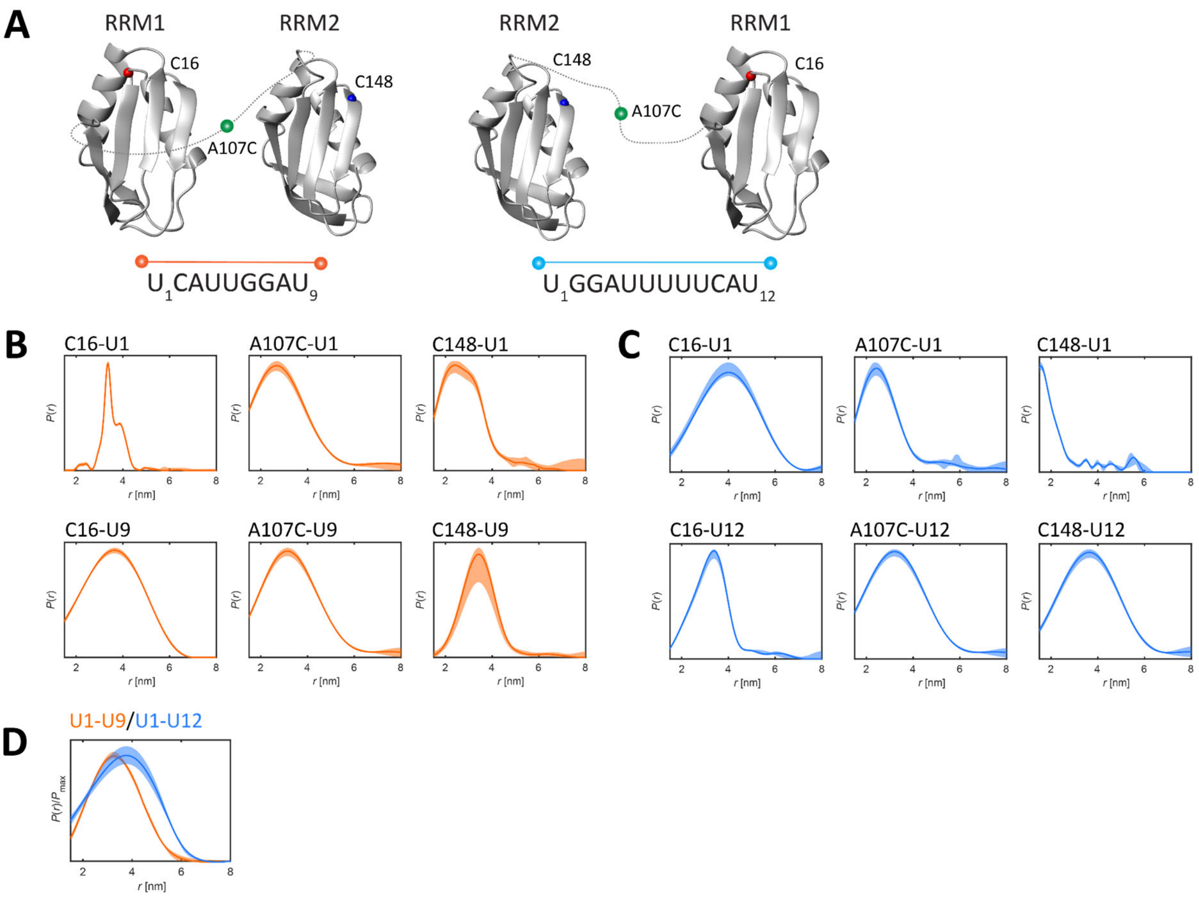


Figure 2.6 Protein-RNA distance distribution restraints. *(A) Schematic visualization of the sites involved in the acquisition of protein-RNA distance restraints (RRM1 = red spheres, linker = green sphere, RRM2 = blue spheres; 5'-UCAUUGGAU-3' RNA = orange spheres, 5'-UGGAUUUUUCAU-3' RNA = light blue spheres). (B) Protein-RNA distance restraints of SRSF1 RRM1+2 in complex with 5'-UCAUUGGAU-3' RNA. (C) Protein-RNA distance restraints of SRSF1 RRM1+2 in complex with 5'-UGGAUUUUUCAU-3' RNA. (D) Intra-RNA distance distributions acquired by labeling both RNA termini. This involves U1-U9 for the protein in complex with 5'-UCAUUGGAU-3' RNA (orange), and U1-U12 for the protein in complex with 5'-UGGAUUUUUCAU-3' RNA (light blue). The protein-RNA complexes are formed using unlabeled wild-type protein. (B-D) Solid lines are the medians and shaded areas the 95% confidence intervals obtained by bootstrapping of 1000 samples. Modified from (Esteban-Hofer 2022).*

Analyzing the complex formed by SRSF1 RRM1+2 with 5'-UCAUUGGAU-3' (Figure 2.6A), we observed, as expected, that the distance distribution of RRM1 and RRM2 to the RNA ends closest to their binding motifs (C16-U1 and C148-U9) are narrow, while the distributions from the RRMs to the RNA-ends further away (C16-U9 and C148-U1) are broad. The A107C-U1 distance distribution is slightly narrower and contains shorter distances than the A107C-U9 distance. However, it is important to note that these differences may be due to the varying flexibility and length of the RNA modifiers that were employed to label the RNA constructs. Similar results were obtained for SRSF1 RRM1+2 in complex with 5′-UGGAUUUUUCAU-3′ RNA (Figure 2.6C). The C148-U1 and C16-U12 distributions are narrow (the RRM2 binding motif is now upstream of the RRM1 binding motif, and hence U1 is closer to the RRM2 binding motif, while U12 is closer to the RRM1 binding motif); the C16-U1 and C148-U12 distributions are broad and longer than the distances to the RNA-ends closest to their binding motifs. The distance A107C-U1 is shorter than to A107C-U12. Finally, we measured the intra-RNA distance restraints for RNA in complex with unlabeled protein sample (U1-U9 and U1-U12, Figure 2.6D).

As expected, these measurements revealed that 5′-UGGAUUUUUCAU-3′ RNA, with more nucleotides, is more extended and flexible than 5′-UCAUUGGAU-3′ RNA.

2.2.3 Analysis of SRSF1 tandem RRMs using PRE

Concurrently with the DEER experiments, Paramagnetic Relaxation Enhancement (PRE) experiments were conducted to obtain information on distances ranging from approximately 12 to 25Å. To attach a paramagnetic probe (MTSL) to the protein, a disulfide bond was utilized, and single cysteine mutants were generated for this purpose. The first PRE samples analyzed included the labeling of the native cysteines (C16 and C148). Furthermore, Y37C and Y72C single cysteine mutants were generated to label RRM1, while S126C and T169C single cysteine mutants were generated to label RRM2 (Figure 2.7A). Due to the flexibility of the linker, experiments with a paramagnetic probe on the linker were not conducted, unlike the study by (De Silva et al. 2022), where only the residue E120C was labeled.

The correct folding of the protein was verified using ^{1}H-^{15}N heteronuclear single quantum spectroscopy (HSQC) in the presence and absence of the diamagnetic probe (Figure A2). Chemical shift differences were observed due to the mutations and the presence of the probe, and the paramagnetic sample labeling was confirmed using CW EPR experiments.

The transverse relaxation enhancement ($\Gamma_2 = R_{2,para} - R_{2,dia}$), or PRE rate, which represents the difference in the transverse relaxation rates of the paramagnetic ($R_{2,para}$) and diamagnetic ($R_{2,dia}$) states of the probe attached to the protein, was analyzed for each residue (Figure 2.7B). PRE rate is dependent on the average $\langle r^{-6} \rangle$ distance between the electron spin and the nuclear spin. The PRE rate values are in range from 0 Hz to 170 Hz, with low values corresponding to a weak PRE effect and high values to a strong PRE effect and a distance of 1.2 nm (Figure 2.7B). From the graphs, we observe that the Γ_2 values are not uniformly distributed; we find values that oscillate around a range of 0 Hz to 80 Hz and values close to 170 Hz, with the space in between largely unoccupied. This irregular distribution complicates comparing the free and the bound states and the spin label positions.

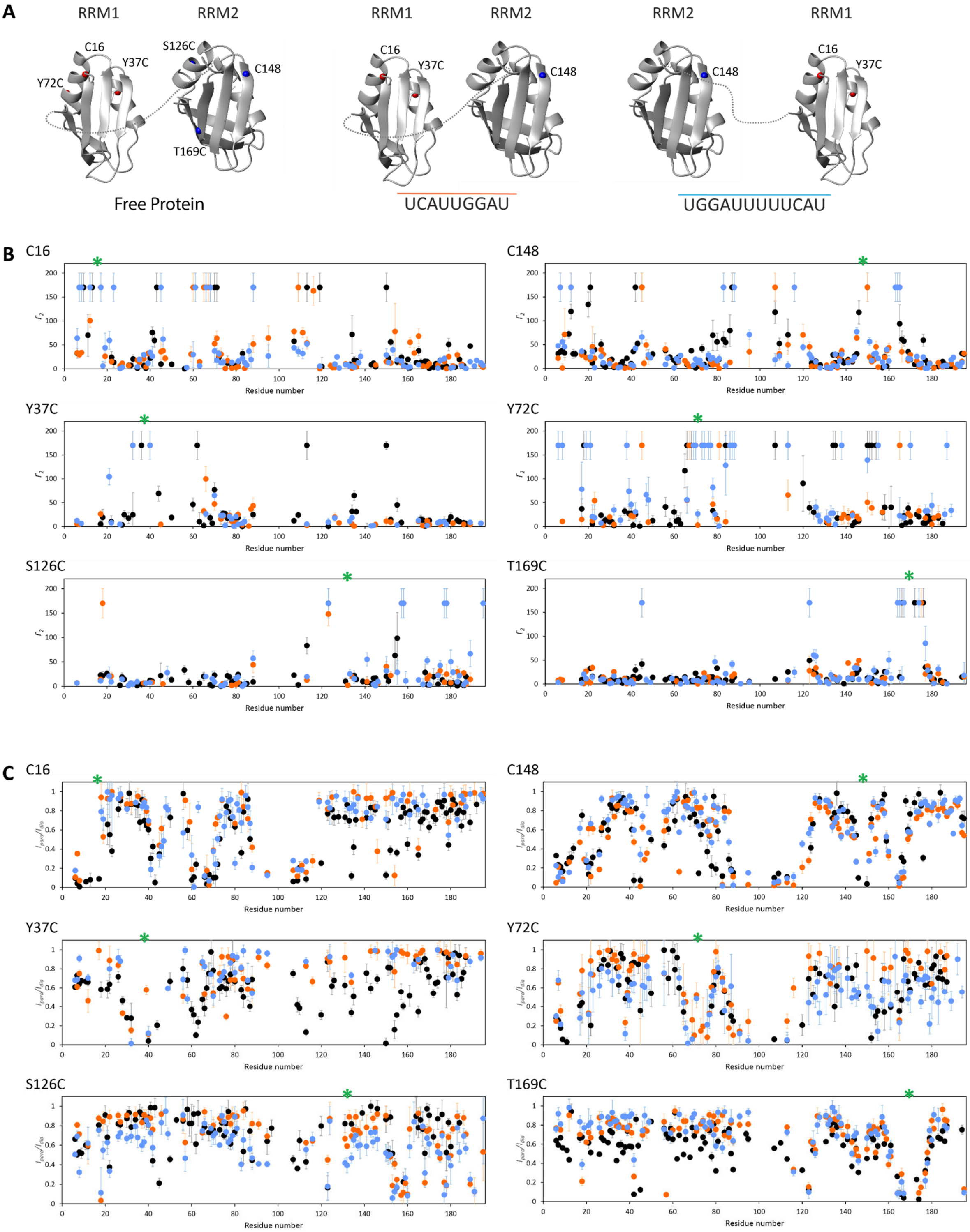


Figure 2.7 ***PRE effect on SRSF1 RRM1+2 on the free form and in complex with RNA.*** *(A) Schematic visualization of the sites involved in the acquisition of PRE restraints in the free from and in complex with RNA (RRM1 = red spheres and RRM2 = blue spheres); (B) PRE rate effect ($\Gamma_2 = R_{2,para} - R_{2,dia}$). (C) PRE ratio effect (I_{para}/I_{dia}). Black denotes free protein, orange SRSF1 RRM1+2 bound to 5'-UCAUUGGAU-3' RNA and in light blue SRSF1 RRM1+2 bound to 5'-UGGAUUUUUCAU-3' RNA. The spin label position is marked a green star on the top of each graph.*

We then analyzed also the PRE ratio which distributes values from 0 to 1, with low values corresponding to a strong PRE effect and high values to a weak PRE effect (Figure 2.7C) allowing us to compare free and bound states. The PRE ratio (I_{para}/I_{dia}) is calculated from the normalized peak integrals of the paramagnetic I_{para} and the diamagnetic I_{dia} samples; the normalization was performed using GB1, which does not show a PRE effect.

Analysis of the PRE results for the free protein (Figure 2.7C, Figure A3) revealed PRE effects both within the domain where the spin label was attached and on the other RRM. A stronger intra-RRM PRE effect was observed on residues close to the spin label, and a PRE effect was also observed in the linker, indicating a close arrangement to one of the two RRMs.

When analyzing the bound states, it was observed that the PRE effects were generally weaker than those observed for the protein alone, indicating that the two RRMs are in a closer conformation when they are in the free form. Comparing the two bound states, we observe difference depending on the mutant (Figure A4-5). When the spin label is attached to C16, Y37C, C148 and T169C we observe a stronger PRE effect on the other RRMs when the protein binds to 5′-U<u>CA</u>UU<u>GGA</u>U-3′ RNA. On the other hand, when the spin label is attached to Y72C or S126C we observed a stronger PRE effect on the other domain when the protein binds 5′-U<u>GGA</u>UUUU<u>CA</u>U-3′ RNA. This results as for the DEER experiments indicate that depending on the labelled residue we can observe different PRE effect and even in the bound state the protein can adopt different conformations.

The inter-RRMs PRE data can be used for structure calculation by determining the distance between the paramagnetic center and other residues from the PRE rate. The distance was calculated using the equation

$$r = \left[\frac{K}{\Gamma_2}\left(4\tau_c + \frac{3\tau_c}{1 + {\omega_h}^2{\tau_c}^2}\right)\right]^{1/6}$$

where τ_c is the correlation time from the experiments mentioned in Session 2.2.1. Further details on the implementation of the PRE data in structure calculation, including *RigiFlex* modeling and CYANA, are provided in Sections 2.2.4 and 2.2.5.

2.2.4 *RigiFlex* modeling of SRSF1 tandem RRMs in the free form and in complex with RNA

Dr. Laura Esteban-Hofer from the group of Prof. Gunnar Jeschke at ETH Zürich performed calculations and analysis on ensembles obtained using *RigiFlex* modeling. The methodology and detailed analysis are described in the PhD thesis (Esteban-Hofer 2022).

This method calculates a first ensemble by using DEER distance distributions as main restraints, while data from additional techniques, such as PRE, are included in the ensemble fit step(Jeschke and Esteban-Hofer 2022). For the calculation of the first ensembles with around 1200 conformers, 10 DEER distance distributions were used as main restraints for the free protein: seven inter-RRM and three RRM-linker. In addition, for protein-RNA complexes, six protein-RNA and an intra-RNA distance distributions were added (Section 2.2.2.1). The ensemble fitting step was then performed on the three initial ensembles, combining all DEER restraints with PRE data obtained from the first experiments conducted (data from C16 and C148 single-cysteine mutants).

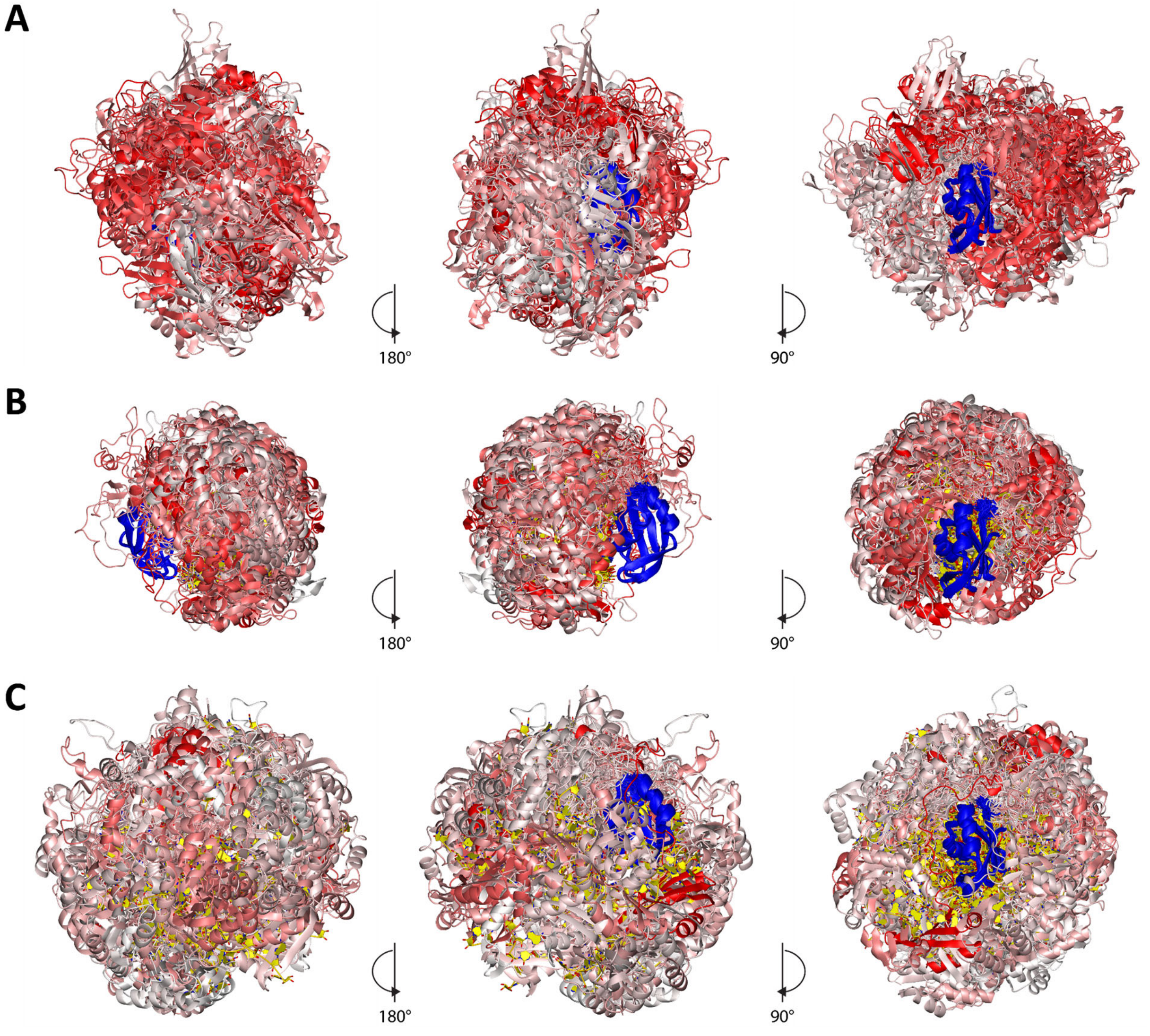


Figure 2.8 Ensembles of SRSF1 RRM1+2 in the free form and in complex with RNA calculated using the RigiFlex modeling. *(A) Free Form (97 conformers), (B) in complex with 5'-UCAUUGGAU-3' RNA (88 conformers) and (C) 5'-UGGAUUUUUCAU-3' RNA (116 conformers). Ensembles are superimposed on RRM2 depicted in blue, RRM1 and linker are in different intensity of red depending on the probability of each conformer (white = low probability, red = high probability); N-terminal (1-15) is not shown. All the three ensembles were calculated by Laura Dr. Esteban-Hofer.*

The final ensembles for the three conditions are shown in Figure 2.8, which are displayed superimposed on RRM2 (blue). Although the ensembles obtained using DEER combined with PRE rate as restraints are shown in this report, a detailed comparison when PRE ratios were used is available in (Esteban-Hofer 2022). The ensemble of the free protein superimposed on RRM2 shows an interesting feature: RRM1 shows preferred conformations as it does not take up the entire available space (Figure 2.8A).

A reduced occupied space was observed for the complex involving the 5′-UCAUUGGAU-3′ RNA (Figure 2.8B), while the complex involving the 5′-UGGAUUUUUCAU-3′ RNA occupies a larger space (Figure 2.8C). After analyzing the ensembles using the PairCorrelation module of MMMx, which calculates the C_α distance between a residue pair for each conformer, a correlation was observed among different regions of the protein in the three ensembles (data not shown). In the case of the free form, a strong correlation was found between the RRM2 β-sheet surface and all the secondary structure elements of RRM1. RRM1 forms a hemisphere around the RRM2 β-sheet surface adopting different conformations with respect to these secondary structure elements. There is a clear correlation between the two rigid domains, which appears to involve specific secondary structure elements and the inter-domain linker (data not shown).

In the case of the complex involving the 5′-UCAUUGGAU-3′ RNA, the pair correlation analysis showed a strong correlation between RRM2 β_2-strand and β_3-strand to all secondary structure elements of RRM1, especially to its β-sheet surface. In addition, the C-terminal region of RRM2 spanned between α_2-helix β_4-strand is correlated to all β-sheet elements in RRM1. In the complex involving the 5′-UGGAUUUUUCAU-3′ RNA, correlations are centered around the RRM1 β-sheet surface and α_1-helix, β_2-strand, β_3-strand, and the β_3'-β_3'' loop of RRM2, which can be rationalized by the binding mode of the protein. The GGA motif is recognized by the heptapeptide in α_1-helix of RRM2, placing the 5′-end towards its β_2-strand and β_3-strand, and its 3′-end towards its β3′-β3″ loop. The CA motif is recognized by the β-sheet surface of RRM1. Hence, these structural elements are correlated when SRSF1 RRM1+2 binds both motifs. The five uracil-linker between the RNA-binding motifs gives to the protein some flexibility, but the conformational arrangement is still limited to optimize the binding to RNA.

It was hypothesized that in the free form the interactions observed stabilize the preferred arrangement of the protein facilitating the binding to RNA once the protein comes in contact with it. Part of these interactions are then likely replaced or complemented by interactions involving the RNA. For each ensemble we also performed the backcalculation for both DEER and PRE restraints to evaluate the agreement with the initial restraints (Figure 2.9 for the free protein, Figure A6-7 for protein-RNA complexes): for the distance distribution restraints the overlap deficiency was calculated in the range of 0 and 1 ($1-\bar{o}$, with 1 meaning no overlap for at least one restraint and 0 meaning perfect overlap for all restraints), whereas for PRE it was used the χ^2. The results of all backcalculation analyses demonstrated very good agreement for the DEER restraints, with a $1-\bar{o}$ value consistently below 0.08. The χ^2 value for the PRE calculation was also good, but upon closer examination of the graphs, we observed that the residues with the highest PRE effect did not influence the DEER ensembles. In the case of the free protein, the fitting of the DEER data resulted in a 1−ō value of 0.065 (Figure 2.9 A), while the χ^2 value for the PRE rate was 1.57 (Figure 2.9 B). As an additional measure of control and to enhance visual clarity, we also performed an analysis of the backcalculation for the PRE rate, which yielded a χ^2 value of 4.06 (Figure 2.9 C).

Although the ensembles exhibited good agreement for both DEER and PRE restraints, they were unable to fully elucidate the molecular interactions underlying the preferred conformations in the free form. This could be due to the use of DEER as the main restraints, which is biasing the entire calculation on long distances. Consequently, we needed an alternative approach to understand the molecular interactions responsible for the various conformations that the RRMs can adopt both in the free form and in complex with RNA.

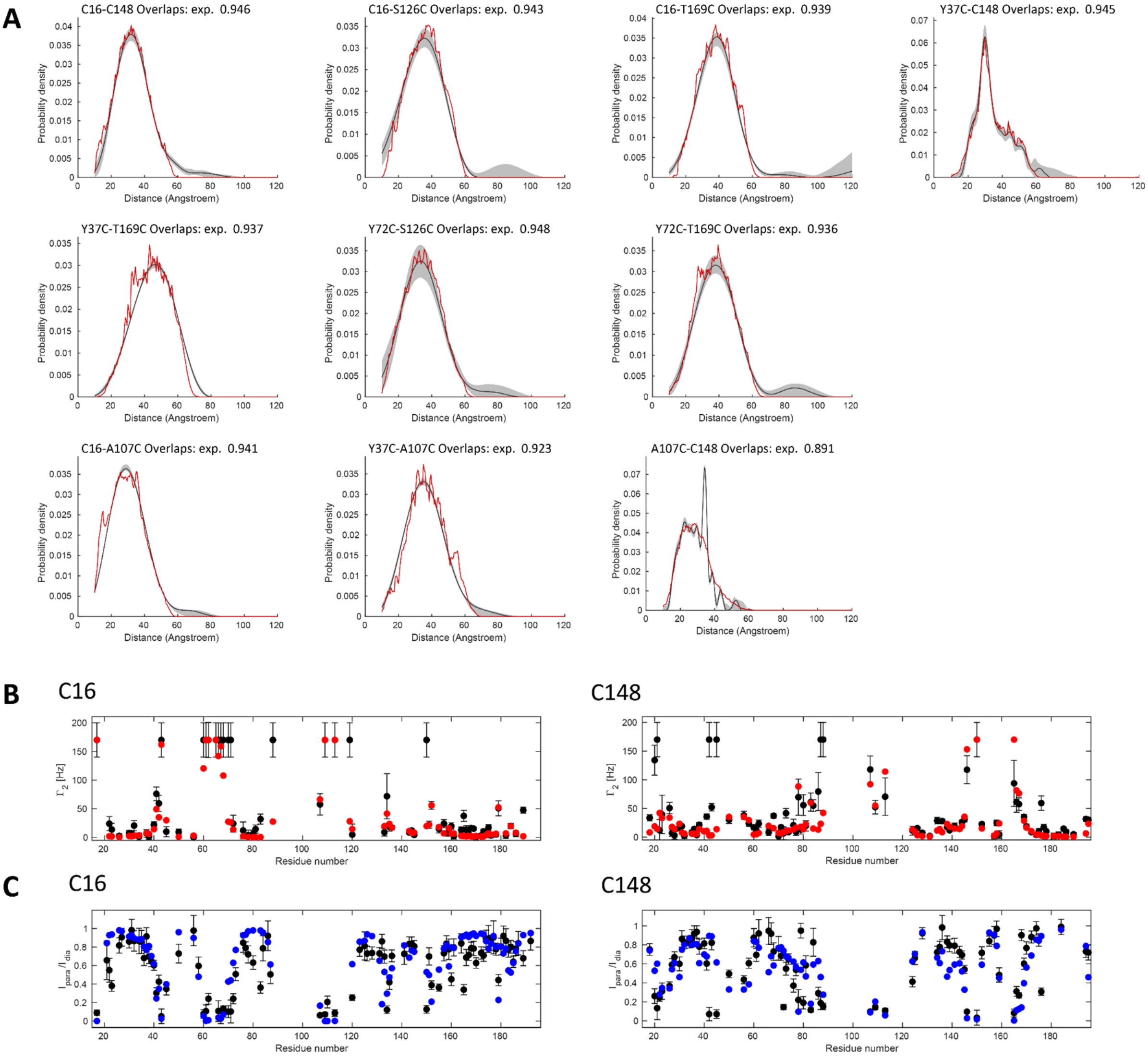


Figure 2.9 Restraint fit of the ensemble of SRSF1 RRM1+2 in the free form calculated using the RigiFlex modeling. *(A) DEER distance distribution (B) PRE rate restraints (C) PRE ratio restraints. (A-C) Experimentally determined distance distribution (black) and predicted for the entire ensemble (red or blue). Final ensemble = 97 conformers. 1−ō = 0.065, χ^2 (PRE rate) = 1.57, χ^2 (PRE ratio) = 4.06.*

2.2.5 Structure determination of SRSF1 tandem RRMs in the free form and in complex with RNA using the hybrid modeling

This section outlines the methodology used to generate ensembles of large systems by combining the CYANA Multistate calculation (Strotz et al. 2017; Vogeli, Guntert, and Riek 2013) with the ensemble fit from the MMMx toolbox (Figure 2.10). Our approach aims to address the limitations of the *RigiFlex* modeling method, which is based on DEER data as the main restraints and is not suitable for studying molecular interactions responsible for specific orientations of two RRMs. Instead, we employ short distances obtained from PRE experiments and incorporate the DEER distance distributions in the last step of the ensemble fit, combining the two complementary methodologies.

To account the presence of the spin label and its mobility in the CYANA calculation, we create dummy atoms that have the coordinates of the center of the cloud of the spin label, rather than creating a pseudo-atom for each residue mutated in cysteine, following the approach described in (Dorn et al. 2022). More details are mentioned in the Section 2.4.8.1 of **Material and Methods**. The distances between the Q8 dummy-atom (which represents the center of the cloud of the spin label) and the NH of the other residues are then recorded in .upl and .lol files.

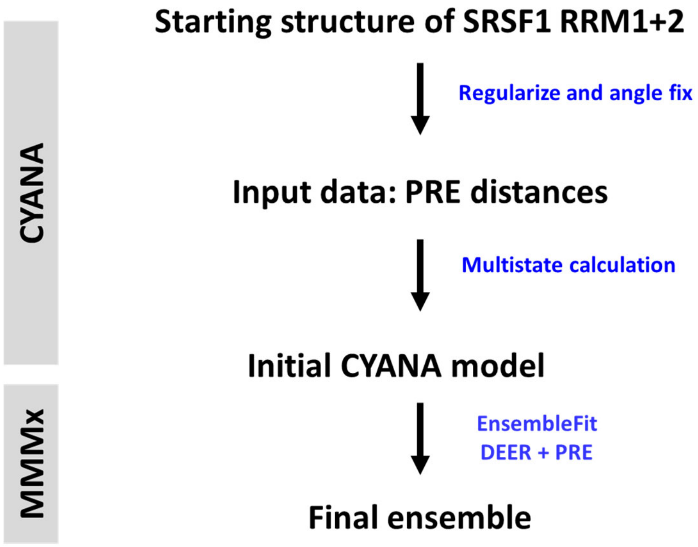


Figure 2.10 Scheme of the hybrid structural modeling approach.

Initially, the classical CYANA approach was employed to perform the calculations which aims to determine a final single structure that satisfies all the input data. The input data used in this study comprised only the PRE data acquired from labeling the native cysteines (C16 and C148).

However, this approach proved to be ineffective for both the free protein and protein-RNA complexes, as the obtained structures exhibited a high value of target function (TF) and a large number of violated restraints (# viol), rendering the structures inaccurate and incorrect (Figure 2.11, Table 2.3). From this outcome, we concluded that the input data utilized in the calculations were not supported by a single conformation, that CYANA failed to identify. Therefore, it was inferred that both the free and bound states were dynamic systems and required more than one solution.

Consequently, the novel Multistate CYANA calculation method, developed by Peter Güntert (ETH Zürich) was employed, using PRE data as ambiguous restraints that did not need to be satisfied by all the calculated states (Strotz et al. 2017; Vogeli, Guntert, and Riek 2013).

Here we report the main steps, details will be then mentioned in the following sections (Figure 2.10). Starting from a SRSF1 RRM1+2 structure, free or in complex with RNA, calculated using CYANA with data from the individual RRMs (Free: He et al., 2005, Tintaru et al., 2007; Bound: Clery et al., 2013; Clery et al., 2017), we regularize and fix the residues of the two RRMs (16-88 for RRM1 and 120-196 for RRM2) and initiate the multistate CYANA calculation. We performed several tests to determine the number of states needed to satisfy all the PRE data used as restraints. This step confirmed our observation from the initial classical CYANA approach calculation that a single structure could not satisfy all the PRE restraints. The ensemble obtained from this first step of multistate calculation constitutes the raw ensemble. As with the *RigiFlex* modeling, we perform ensemble fitting by combining PRE data with DEER restraints. These first ensembles are contracted by fitting populations and discarding conformers with zero or very low population (or probability). We analyze the agreement between the final ensembles and the restraints used for the calculation (both PRE and DEER) for quality control. By utilizing this novel approach that combines CYANA multistate calculation and EnsembleFit step from the MMMx toolbox based on PRE as the primary restraints, we obtained ensembles that enabled us to investigate and comprehend the molecular interactions responsible for the specific orientation of the two RRMs, which was not possible when we applied only the *RigiFlex* modeling.

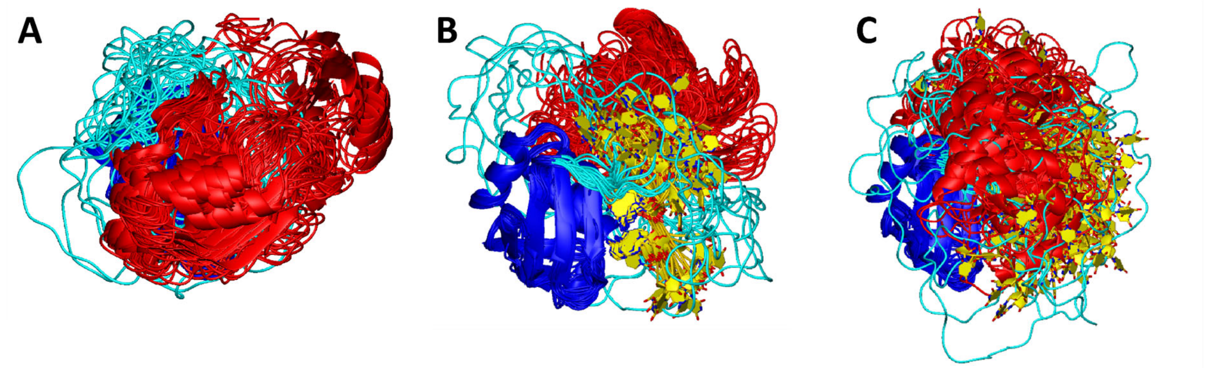


Figure 2.11 Structure calculation of SRSF1 RRM1+2 in the free form and in complex with RNA using standard CYANA. *(A) Free Form, (B) in complex with 5'-U<u>CA</u>UU<u>GGA</u>U-3' RNA and (C) in complex with 5'-U<u>GGA</u>UUUUU<u>CA</u>U-3' RNA; 300 structures were calculated and the best 20 selected.* ***.*** *Ensembles are superimposed on RRM2 depicted in blue, with RRM1 in red, the linker in cyan and RNA in yellow.*

# states	TF	# viol	max
Free protein	25.48	59	1.28
bound to 5′-UCAUUGGAU-3′	18.41	30	0.63
bound to 5′-UGGAUUUUUCAU-3′	31.39	55	0.85

Table 2.3. Statistic values of structure calculation of SRSF1 RRM1+2 in the free form and in complex with RNA using standard CYANA. *TF = target function; #viol = number of restraints that are violated, max = the maximal violation*

2.2.5.1 Ensemble of SRSF1 tandem RRMs in the free form

In this section, we will examine the ensemble obtained for the free SRSF1 RRM1+2 protein, which is shown in Figure 2.12. The starting structure of the free protein was generated using the standard CYANA software and the input files from the individual RRMs (He et al. 2005; Tintaru et al. 2007). We conducted several calculation tests to determine the number of states required for the final calculation, ranging from 1 to 10 states, using 213 PRE restraints from the 6 mutants (Table A3). For each calculation, we generated 2000 structures, 200000 annealing steps, and selected the top 20 structures (i.e., number of conformers = 20 x number of calculated states) based on the lowest target function and number of violations. We found that increasing the number of states led to decreasing values of the target function and the number of violated restraints. However, upon examining the structures, we observed that the ensembles obtained from 8, 9, and 10 states displayed non-informative states with a fully extended linker (Figure A8); this indicates that CYANA placed the last states at the maximum distance between the two RRMs when more states were calculated than necessary, even if the target function was close to zero with only a single restraint violated.

The best outcome was achieved by calculating 7 states. The ensemble for this state was recalculated using 5000 structures, 200000 annealing steps, and selecting the best 20 structures, resulting in a total of 140 conformers (Figure 2.12A). Upon analyzing the initial ensemble, we observed that despite the high percentage of glycines present in the linker domain, the two RRMs occupied specific space relative to each other. Furthermore, as a control measure, we conducted a backcalculation analysis for the DEER measurements, which surprisingly exhibited a remarkable agreement between the backcalculated and experimental values (1−ō = 0.195, Figure 2.13A), even though they were not included in this specific step. This unexpected and reassuring initial result demonstrated the efficacy of the Multistate CYANA calculation approach for our system, as we could obtain a good agreement for the DEER data. This result is remarkable as the CYANA method does not generally utilize any electrostatic information or force field, and the initial ensemble generated was chemically correct.

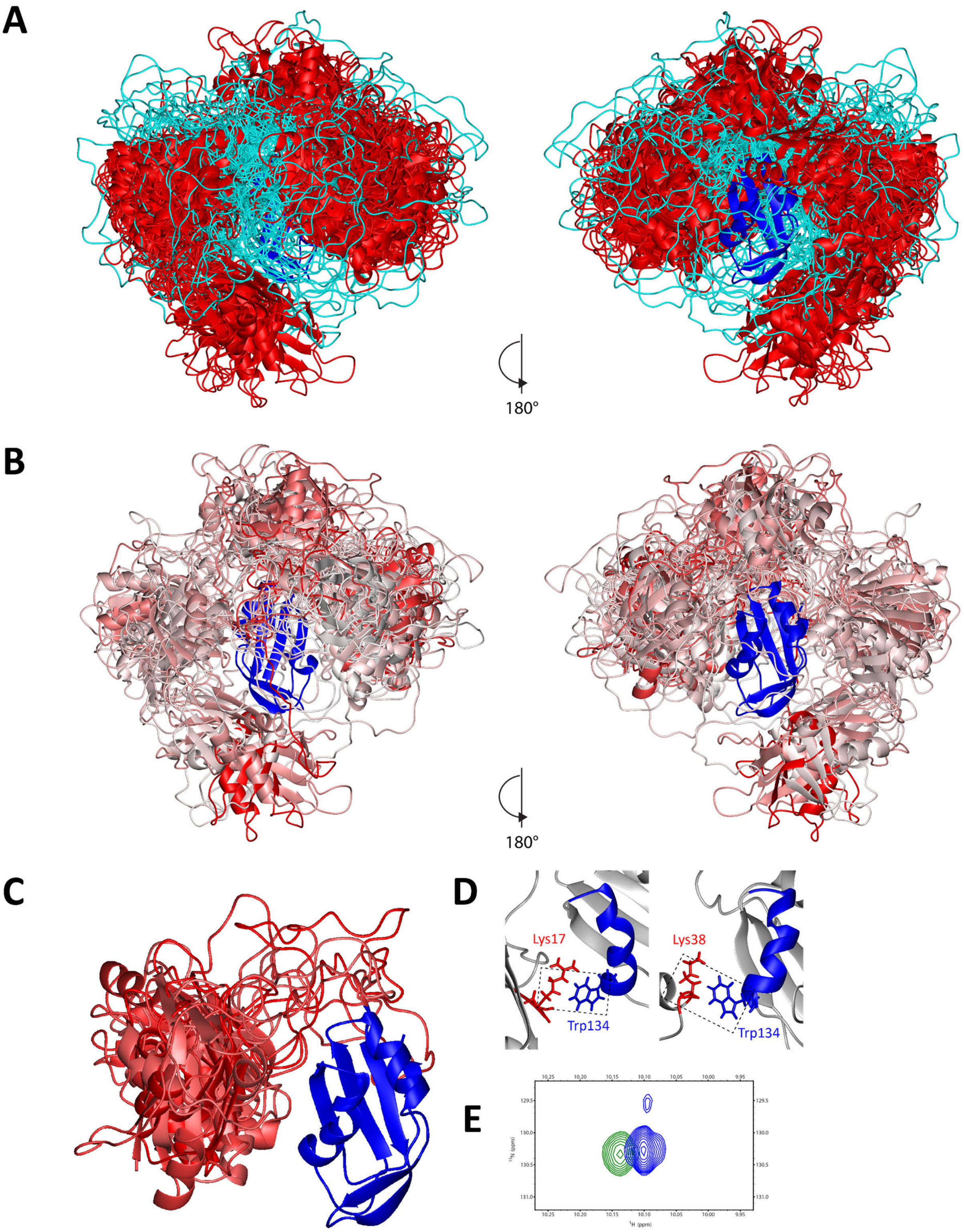


Figure 2.12 Ensembles of SRSF1 RRM1+2 in the free form calculated using the hybrid modeling. *(A) Raw ensemble obtained after the Multistate CYANA calculation using 213 PRE restraints (7 states, 140 conformers). Conformers are superimposed on RRM2 depicted in blue, with RRM1 in red and the linker in cyan. 5000 structures calculated with 200000 steps, 20 with lowest target function selected; TF = 1.09, # viol = 2; max =0.37. (B) Ensemble refined after the MMMx EnsembleFit step (49 conformers). Ensembles are superimposed on RRM2 depicted in blue, RRM1 and linker are in different intensity of red depending on the probability of each conformer (white = low probability, red = high probability); N-terminal (1-15) is not shown. (C) 7 conformers with the highest probability forming 40% of the total population. (D) Details of some interactions involving the Trp134 on the RRM2* a_1*-helix with RRM1. (E) Detail of the* ^{1}*H-*^{15}*N HSQC spectra overlay between SRSF1 RRM1+2 (residue 1-196, green) and RRM2 (residues 107-203, blue) showing the Trp134 side-chain.*

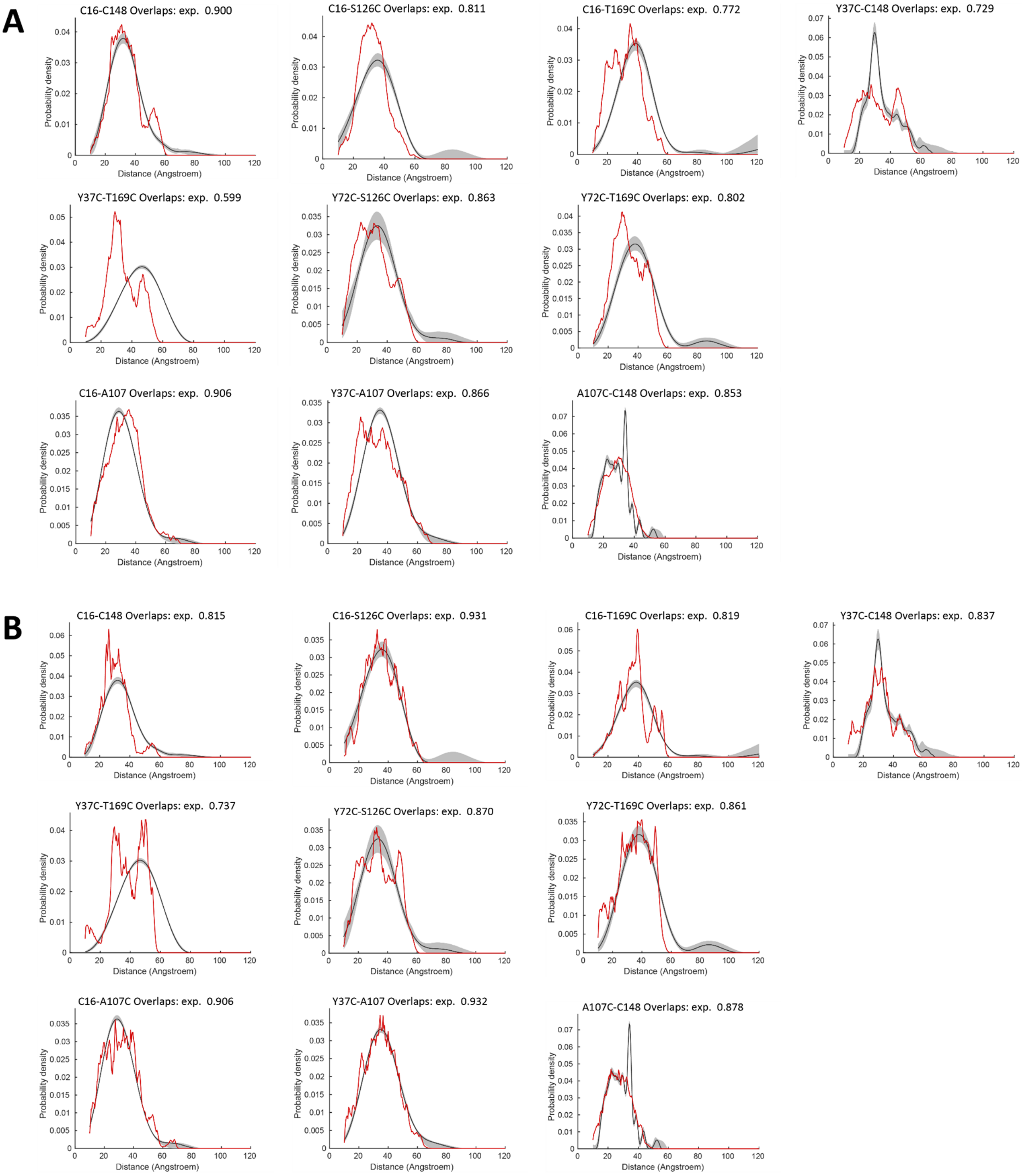


Figure 2.13 ***Distance distribution restraint fulfillment of SRSF1 RRM1+2 in the free state****. (A) After the Multistate CYANA calculation (140 conformers), $1-\bar{o} = 0.195$ (B) After the MMMx EnsembleFit step (49 conformers) $1-\bar{o} = 0.143$. Black = distance distribution determined experimentally, red = predicted distance distribution.*

Furthermore, two independent experiments were employed (PRE and DEER), and both demonstrated a high level of consistency with each other (even if they are performed at different temperatures). As a result, this new approach is reasoned to be valid and accurate.

The EnsembleFit step was then performed using MMMx, resulting in the selection of 49 conformers with different probabilities (Figure 2.12B). The back-calculation resulted in an improvement in the fitting of the DEER restraints ($1-\bar{o} = 0.143$, Figure 2.13B).

Furthermore, also the PRE restraints exhibited good agreement between experimental and backcalculated values (Figure A9-10).

All these results suggest that the approach we developed is applicable to systems comprising folded domains linked by unfolded and disordered regions and can effectively combine different data sets from various techniques, such as PRE for shorter distances and EPR for longer distance distributions. The ensemble of the free protein was then analyzed to determine the molecular interactions responsible for the specific orientation of the two RRMs.

Initially, we conducted an analysis of the conformers with the highest probability, and upon observation, we noted that a majority of these conformers displayed an interaction between the α_1-helix of RRM2 and various regions of RRM1 (Figure 2.12C-D). Notably, Trp134 from RRM2 was frequently involved in these interactions with residues present in the β_1-strand, the connecting loop between β_1-strand and α_1-helix, and the α_1-helix itself. Furthermore, we observed an interaction between Lys138 from RRM2 and α_2-helix of RRM1. Regarding conformers with lower probability, we noted interactions between the α_2-helix and β_4-strand of RRM2 with RRM1.

The most remarkable finding was the contribution of Trp134 in maintaining the proximity of the RRMs. Specifically, the comparison between the ^{1}H-^{15}N HSQC spectra of SRSF1 RRM2 and SRSF1 RRM1+2 showed chemical shift differences in the side chain of Trp134 (Figure 2.3A and Figure 2.12 E). Both the ensemble and previous NMR experiments highlight the significance of Trp134 in interacting with RRM1. The combination of results from these different and independent experiments supports the crucial role of Trp134 in facilitating interactions and inducing the preformed conformations of the two RRMs. This, further confirms the robustness of our approach, as the experiments were conducted independently. In addition, the interactions involving Trp134 suggest that the free form of the protein may adopt conformations that have the tandem RRMs in close proximity, potentially preventing RNA binding.

2.2.5.2 Ensemble of SRSF1 tandem RRMs in complex with 5′-U<u>CA</u>UU<u>GGA</u>U-3′ RNA and 5′-U<u>GGA</u>UUUU<u>CA</u>U-3′ RNA

Our aim was to apply the same methodology that was utilized for the free protein to obtain the two protein-RNA complexes. Following the procedure used for the free protein, we utilized the standard CYANA software to calculate the starting structures of the protein when bound to two distinct RNA molecules, specifically 5′-U<u>CA</u>UU<u>GGA</u>U-3′ or 5′-U<u>GGA</u>UUUU<u>CA</u>U-3′. We used distance constraints from previous studies on the individual RRMs in complex with CA or GGA motifs (Clery et al. 2013; Clery et al. 2021). After obtaining the starting structures, we performed the calculations to determine the number of states that were required for the protein-RNA complexes ranging from 1 to 6 states (we expected more compacted structures than the free form). We used 226 restraints for the complex with 5′-UCAUUGGAU-3′ RNA and 256 restraints for the complex with 5′-UGGAUUUUUCAU-3′ RNA. It is important to keep in mind that the target function is higher due to protein-RNA constraints (Table A4-5).

As for the free protein, the calculation was performed generating 2000 structures, 200000 annealing steps, and selected the top 20 structures (i.e., number of conformers = 20 x number of calculated states) based on the lowest target function and number of violations.

Unlike the free protein, where calculating more states helped to satisfy all the PRE input data, we observed that for both protein-RNA complexes, even when we increased the number of states, CYANA could not find a solution that satisfied all input data (Table 2.4 and Table 2.5).

# states	TF	# viol	max
1	2199.79	61	11.36
2	242.23	54	4.79
3	162.07	62	4.25
4	151.85	70	3.02
5	162.78	78	2.62
6	156.8	86	2.53

Table 2.4 Test of Multistate CYANA calculation for SRSF1 RRM1+2 in complex with 5′-UCAUUGGAU-3′ RNA (6 PRE mutants, 226 restraints). *Calculations were performed for 1-6 states, calculating 2000 structures, 200000 steps, and the best 20 structures were selected. TF = target function; #viol = number of restraints that are violated, max = the maximal violation.*

# states	TF	# viol	max
1	2169.67	67	13.6
2	213.99	42	3.89
3	105	49	3.07
4	77.26	49	2.19
5	88.15	88	2.32
6	108.92	67	2.41

Table 2.5 Test Multistate CYANA calculation for SRSF1 RRM1+2 in complex with 5′-UGGAUUUUUCAU-3′ RNA (6 PRE mutants, 252 restraints). *Calculations were performed for 1-6 states, calculating 2000 structures, 200000 steps, and the best 20 structures were selected. TF = target function; #viol = number of restraints that are violated, max = the maximal violation.*

The observed phenomenon could be explained by the variation in stoichiometry of the protein mutants and partial binding of the protein to RNA. In fact, the presence of unbound protein could contribute to the observed signal. To address this issue, we integrated PRE distances obtained from the native cysteine residues (C16 and C148) and conducted calculation tests incorporating other PRE mutants. After careful evaluation, we identified Y37C mutant as the most promising candidate, which exhibited excellent agreement and compatibility with CYANA calculations (Table 2.6 and Table 2.7). Based on the analysis of the statistical parameters (TF and # viol), the 2 states model was initially considered the most suitable candidate for both the protein-

RNA complexes. However, further analysis were performed by comparing the structures obtained from the 2 states and 3 states models with the DEER data through backcalculation. Interestingly, the results demonstrated that the 3 states model was better than the 2 states model in terms of accuracy and agreement with the experimental data. Subsequently, we performed the EnsembleFit step using MMMx on the raw ensemble calculated with 3 states. Results are presented in Figure 2.14 for the complex with 5′-UCAUUGGAU-3′ RNA and Figure 2.15 for the complex with 5′-UGGAUUUUUCAU-3′ RNA.

# states	TF	# viol	max
1	72.24	13	5.67
2	11.58	13	0.91
3	13.07	18	0.69
4	18.96	23	0.97
5	29.38	30	1.11
6	35.82	37	1.21

Table 2.6 Test Multistate CYANA calculation for SRSF1 RRM1+2 in complex with 5′-UCAUUGGAU-3′ RNA (3 PRE mutants = C16, Y37C, C148; 122 restraints). *Calculations were performed for 1-6 states, calculating 2000 structures, 200000 steps, and the best 20 structures were selected. TF = target function; #viol = number of restraints that are violated, max = the maximal violation.*

# states	TF	# viol	max
1	133.91	20	4.26
2	6.38	10	0.57
3	14.37	16	1.12
4	27.08	35	1.56
5	40.93	35	1.44
6	56.28	42	1.77

Table 2.7 Test Multistate CYANA calculation for SRSF1 RRM1+2 in complex with 5′-UGGAUUUUUCAU-3′ RNA (3 PRE mutants = C16, Y37C, C148; 122 restraints). *Calculations were performed for 1-6 states, calculating 2000 structures, 200000 steps, and the best 20 structures were selected. TF = target function; #viol = number of restraints that are violated, max = the maximal violation.*

The initial ensemble for the protein-RNA complex involving the 5′-UCAUUGGAU-3′ RNA exhibited already a compact conformation (Figure 2.14A). Out of the 60 conformers in the raw ensemble, 26 were selected following the EnsembleFit step (Figure 2.14B). Superimposing all the conformers on RRM2 (in blue) revealed that RRM1 occupies a highly specific space close to the α_1-helix of RRM2. The space between the CA and GGA binding site contained only two uridines, providing an explanation for why the two RRMs are in close proximity in the final ensemble. Comparing the final ensembles of the free and protein-RNA complex with 5′-UCAUUGGAU-3′ RNA (Figure 2.14C), we observed that the space occupied by RRM1 in the bound state was already present in the free ensemble (for all the 26 conformers). This suggests that what was observed in the free form was not a prevention of RNA binding (i.e., maintaining the two RRMs in a close conformation), but rather a case of conformational selection. In fact, the free form displays conformations that are very similar to some observed in the bound form (Figure 2.14D), and these conformations can facilitate RNA binding. In the free form, we observed that Trp134 plays a crucial role in maintaining the two RRMs in proximity to each other. Previous studies have shown that Trp134 also participates in the interaction between RRM2 and RNA (Clery et al. 2013). Therefore, in order for Trp134 to be involved in binding RRM2 to the GGA motifs, the interactions present in the unbound form needs to be disrupted. However, the domains are already in close proximity and in an appropriate orientation to bind RNA. In summary, the system appears to be pre-formed through interactions that must be disassembled upon RNA binding, allowing the RRMs to be in close proximity.

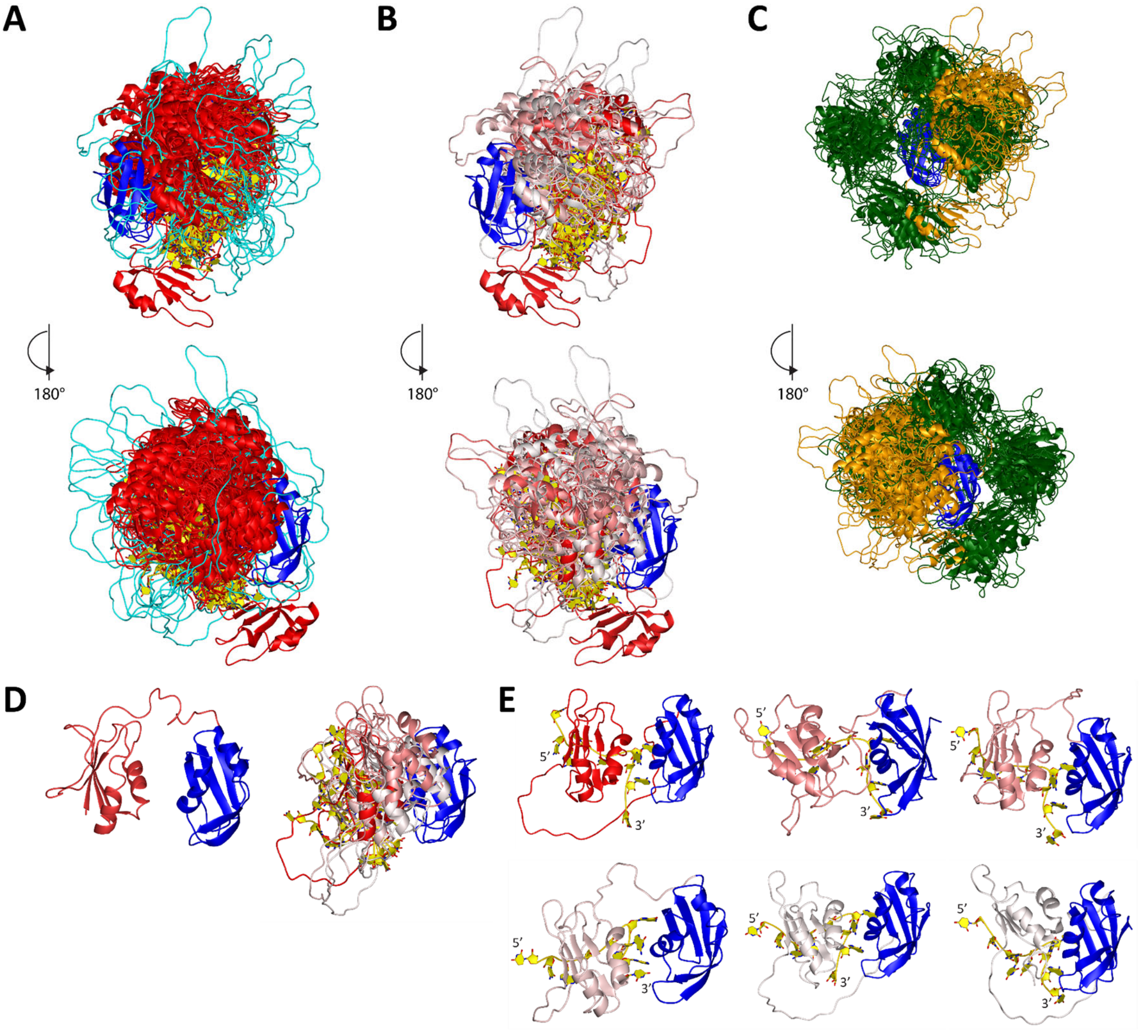


Figure 2.14 Ensembles of SRSF1 RRM1+2 in complex with 5'-UCAUUGGAU-3' RNA calculated using the hybrid modeling. *(A) Raw ensemble obtained after the CYANA Multistate calculation (3 states, 60 conformers). Conformers are superimposed on RRM2 depicted in blue, with RRM1 in red, the linker in cyan and the RNA in yellow. 5000 structures calculated with 200000 steps, 20 with lowest target function selected; TF = 8.79, # viol = 14; max =0.46 (From the RNA: TF = 7.35, # viol = 12; max =0.40) (B) Ensemble refined after the MMMx EnsembleFit step (26 conformers). Ensembles are superimposed on RRM2. (C) Comparison ensembles of SRSF1 RRM1+2 free and in complex with 5'-UCAUUGGAU-3' RNA: ensembles are superimposed on RRM2 depicted in blue, RRM1 and linker are green for the free protein and in orange for the protein-RNA complex. (D-E) Example of conformational selection. (D) Conformer of the free form and conformers of the bound form that show conformational selection (E). Overview of the conformers of the bound form showed in (D). (B, D, E) RRM2 is in blue, RRM1 and linker are in different intensity of red depending on the probability of each conformer (white = low probability, red = high probability) and RNA is in yellow; N-terminal (1-15) is not shown.*

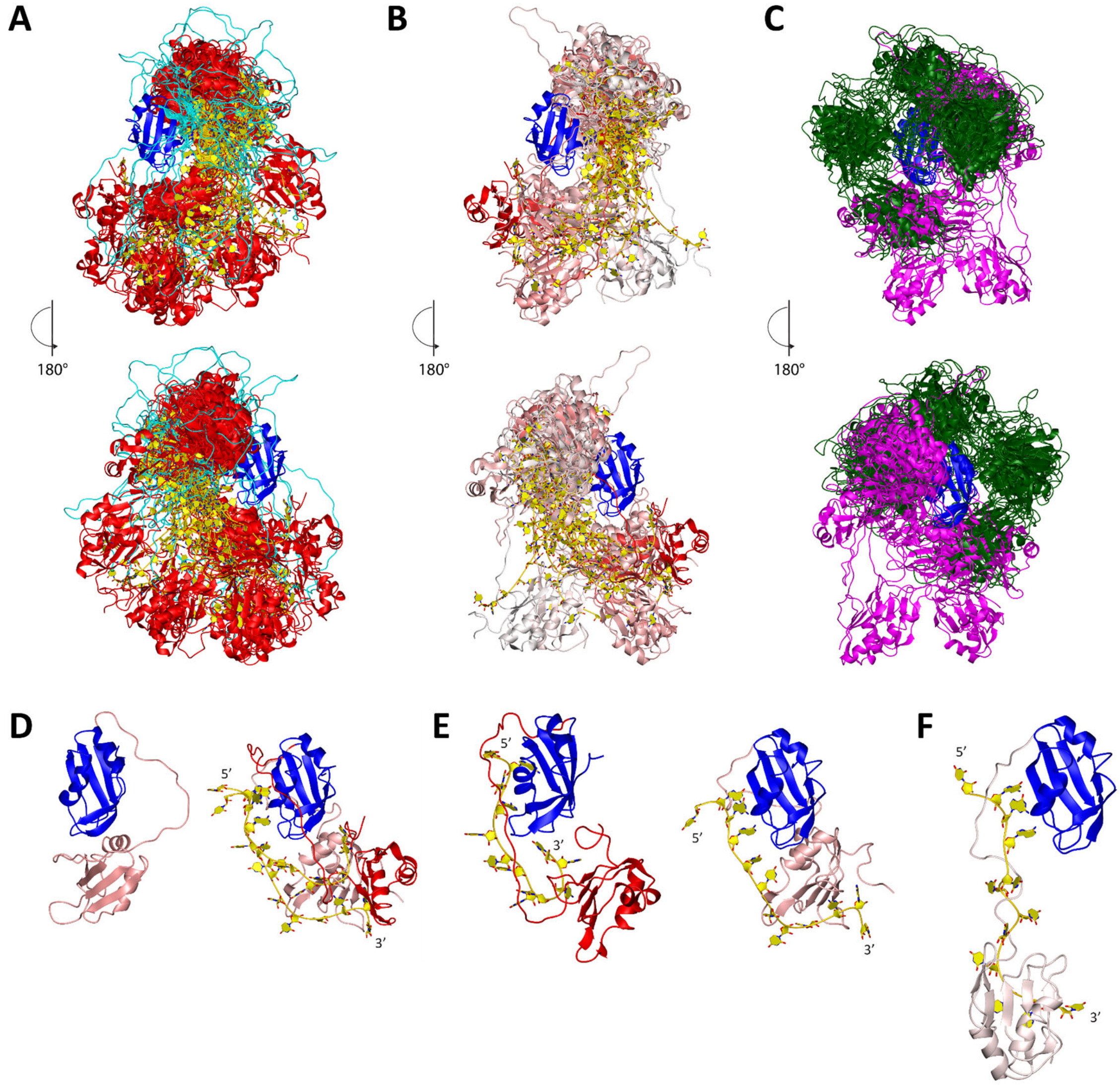


Figure 2.15 Ensembles of SRSF1 RRM1+2 in complex with 5'-UGGAUUUUUCAU-3' RNA calculated using the hybrid modeling. *(A) Raw ensemble obtained after the CYANA Multistate calculation (3 states, 60 conformers). Conformers are superimposed on RRM2 depicted in blue, with RRM1 in red, the linker in cyan and the RNA in yellow. 5000 structures calculated with 200000 steps, 20 with lowest target function selected; TF = 13.15, # viol = 15; max =0.98 (From the RNA: TF = 8.14, # viol = 12; max =0.41) (B) Ensemble refined after the MMMx EnsembleFit step (23 conformers). Ensembles are superimposed on RRM2. (C) Comparison ensembles of SRSF1 RRM1+2 free and in complex with 5'-UGGAUUUUUCAU-3' RNA. (D-E) Example of conformational selection. (D) Conformer of the free form and conformers of the bound form that show conformational selection (E). Overview of the conformers of the bound form showed in (D). (F) Example of conformer that showed induced fit. (B, D, E, F) RRM2 is in blue, RRM1 and linker are in different intensity of red depending on the probability of each conformer (white = low probability, red = high probability) and RNA is in yellow; N-terminal (1-15) is not shown.*

The ensemble of conformations for the protein-RNA complex involving the 5'-UGGAUUUUUCAU-3' RNA appears to be less compact than the other complex. The raw ensemble comprises 60 conformers, and upon superimposition of all conformers onto RRM2 (Figure 2.15A, in blue), shows that RRM1 occupies the region close to the RRM2 α_1-helix (as for the other ensembles) and conformers that have the RNA in an extended form are also observed. After the EnsembleFit step, 23 conformers were selected (Figure 2.15B). The final ensemble was then compared to the one obtained for the free protein (Figure 2.15C). Only 17 conformers showed similarity to the free form, which can be explained as conformational selection (Figure 2.17D). The remaining 6 conformers lacked an equivalent in the free form and were thus explained as induced fit. In addition to the Trp134-involved conformations observed in the free state, other conformations were found to prepare the domains for RNA binding using the 5'-UGGAUUUUUCAU-3' RNA. This observation is consistent with the findings of (De Silva et al. 2022) who reported that the linker assumes a more stretched conformation when the RRM1-binding motif CA is located downstream in simulations.

We conducted also the backcalculation analysis for both protein-RNA complexes (Figure A11-12). Our results indicate that the agreement with the input data was not as good as that observed for the free protein, but still within an acceptable range. It is noteworthy that among the DEER data, the best fitting was obtained for the protein-RNA distances, as opposed to the protein-protein restraints.

Taken together, these results demonstrate the development of a new method to study multidomain proteins that are connected by disordered regions, utilizing PRE and DEER experiments. This method enabled the elucidation of the molecular interactions responsible for the orientation of the two RRMs in the free form, which could explain the binding to RNA through both conformational selection and induced fit. Furthermore, the observation of distinct conformers in the bound states, which are not identical to each other, suggests that although the presence of RNA may reduce the degree of freedom, the protein still exhibits some flexibility through the inter-RRMs linker domain, enabling it to assume different conformations.

2.2.5.3 Comparison: MMMx *versus* Hybrid ensembles

The present study compared two different methods for generating protein ensembles, *RigiFlex* modeling and the hybrid approach that combines Multistate CYANA calculation and MMMx EnsembleFit step. Our analysis of the ensembles we obtained (Figure 2.16 and Figure A13), revealed significant differences between the two methods, both for the free protein and for the protein-RNA complexes; in fact, the ensembles generated via *RigiFlex* exhibit a larger number of conformers and appear less compact when compared to the ensembles obtained through the hybrid method. These differences could be attributed to variations in the manner in which the input data were utilized. In the *RigiFlex* method, DEER data were primarily used as restraints, which cover longer distances, while the PRE distances were only used in the final step of the EnsembleFit. Consequently, the two RRMs were positioned far apart from each other. Additionally, the DEER data consist of distributions and not single distances, which influenced all possible conformations that could have been calculated, leading to the formation of multiple conformers (around 100) in the final ensemble. Conversely, the Multistate CYANA approach utilized PRE data as the primary restraints. The shorter distances had the majority of influence on the calculation, and the number of structures calculated and selected were determined at the beginning of the process. Furthermore, the second method allowed for a more detailed study of the three ensembles obtained, particularly the molecular interactions underlying the different conformations.

The observed differences are particularly evident in the unbound state. Specifically, when using *RigiFlex* modeling alone to calculate the ensemble, a strong correlation was observed between the β-sheet surface of RRM2 and all secondary structure elements of RRM1. RRM1 surrounds the β-sheet of RRM2, and only a small number of conformers (and not with high probability) accounts for the interactions between Trp134 and RRM1. Both methods demonstrated that the SRSF1 tandem RRMs are not independent and can adopt preferred conformations. However, the new hybrid method allowed us to gain insight into the interactions that may be responsible for these specific orientations.

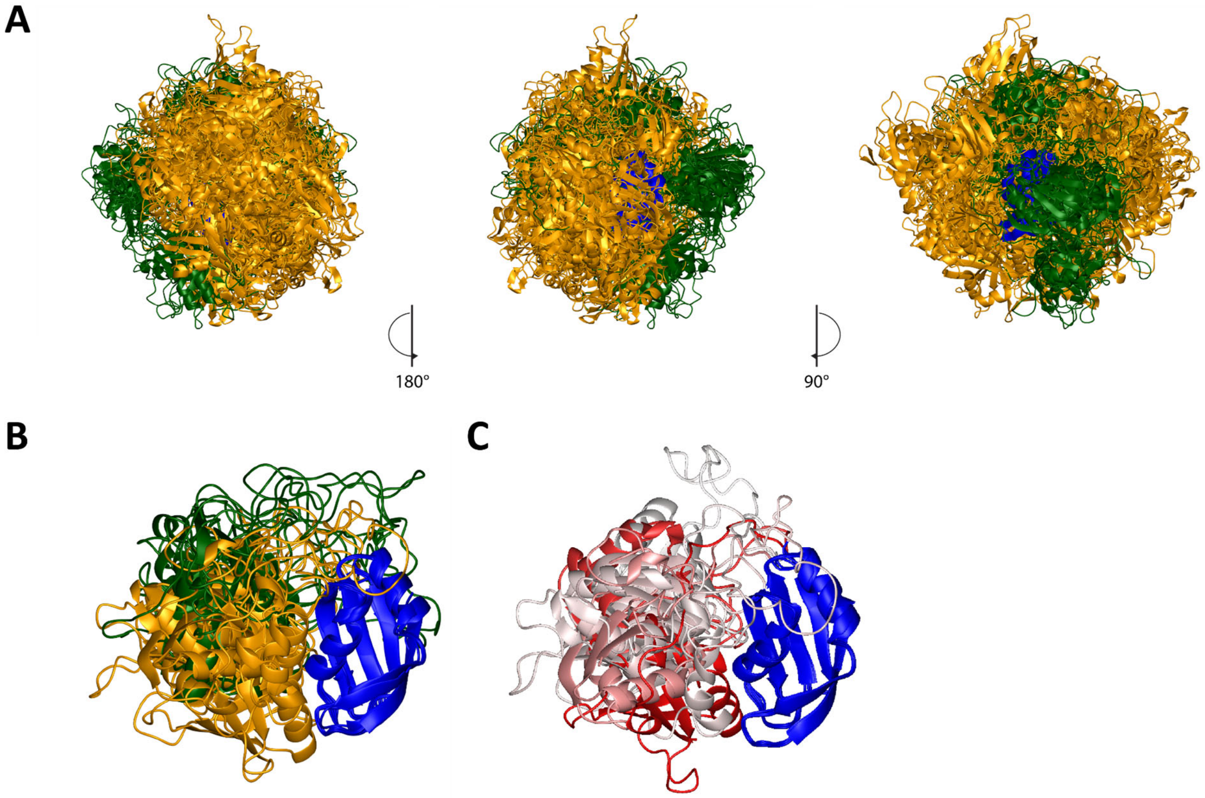


Figure 2.16 Comparison between ensembles obtained with RigiFlex and the hybrid modeling in the free state. *(A) SRSF1 RRM1+2 free protein. (B) Conformers that show interaction between Trp134 and RRM1. Ensembles are superimposed on RRM2 depicted in blue, RRM1 and linker are orange for the ensemble calculates using the RigiFlex modeling and in green for the ensemble obtained using the hybrid method. (C) Conformers of the ensemble calculated using RigiFlex modeling that show interaction between Trp134 and RRM1. Conformers are superimposed on RRM2 depicted in blue, RRM1 and linker are in different intensity of red depending on the probability of each conformer (white = low probability, red = high probability); N-terminal (1-15) is not shown.*

2.3 Discussion

In this study a novel approach for determining the structure of proteins with folded domains connected by disordered regions has been presented. The specific case study involved the SRSF1 tandem RRMs (lacking the RS domain), in which the linker domain that connects the two RRMs is 30 residues long and rich in glycines, suggesting high flexibility. Previous simulations also indicated that the linker was flexible even in its bound form (De Silva et al. 2022). The present study confirmed this finding using calculations that generated ensembles rather than a single conformation, but also demonstrating that certain conformations observed in the unbound protein are also detected upon RNA binding.

To achieve this, our study combined NMR (PRE experiments) and EPR (DEER experiments) techniques. Initially, the *RigiFlex* modeling developed by Prof. Gunnar Jeschke was used, which primarily used DEER distance distributions as main restraints while additional restraints from other techniques were included in the final step of the EnsembleFit (Jeschke and Esteban-Hofer 2022). This method generated positive results, with the obtained ensemble showing good agreement between the experimental data and backcalculated values. However, since the DEER restraints cover large distances (20-100Å), the domains were placed far apart from each other, making it challenging to understand the molecular interactions that facilitated the specific orientation of the two RRMs, particularly in the free protein.

Due to this limitation of the *RigiFlex* modeling, we chose to use PRE distances as primary restraints for the CYANA structure calculation. The PRE restraints cover a shorter distance range of 12-25 Å, which should allow us to better understand the molecular interactions between the two folded domains of SRSF1 tandem RRMs. However, the standard CYANA method did not adequately satisfy our PRE restraints. Therefore, we utilized the new Multistate CYANA calculation to generate multiple solutions (Strotz et al. 2017; Vogeli, Guntert, and Riek 2013; Ashkinadze et al. 2022).

The present study describes the first application on a multidomain and dynamic system of the Multistate CYANA calculation method, which was originally used only for single domain systems (Strotz et al. 2017; Vogeli, Guntert, and Riek 2013; Ashkinadze et al. 2022). The dynamic nature of our system was demonstrated through initial results on the free form using correlation time and DEER experiments, as well as the observation that the protein can bind RNA through a bimodal mode. The Multistate CYANA calculation method allowed the identification of different conformations that the protein could adopt, rather than calculating only a single final conformation. The initial calculation was subsequently refined by combining it with independent DEER experiments, which confirmed the agreement between the backcalculated DEER distance distributions and the experimental values, particularly for the free form. This approach (recording PRE data and performing multistate

CYANA calculations) enables the generation of an initial ensemble of dynamic protein conformations, which can be further refined using data from other techniques.

It has been demonstrated that the disordered region that connects different RNA binding domains serves not only as a connector but also plays a critical role in maintaining protein structure and interacting with other proteins or RNA (Ottoz and Berchowitz 2020). For instance, studies on SRSF1 have shown that residues at the end of the linker sequence are involved in RNA recognition. In addition, experiments examining the correlation time of the protein revealed that the two RRMs can interact with each other, but this interaction only occurs when the domains are connected rather than in isolation. Using NMR and EPR methods, the structure of the two RRMs was determined, revealing their specific localization and orientation relative to each other.

As reviewed in (Afroz et al. 2015), tandem RRMs recognize RNA primarily in three main ways. In the first case, the tandem RRMs are independent of each other in their free state and adopt a rigid structure upon binding to RNA. In the second case, the tandem RRMs maintain a fixed orientation in their free state, which is preserved upon RNA binding. In the last case, a conformational change in the structure of the tandem RRMs occurs upon RNA binding.

PTB RRM34 adopts a fixed conformation in its free form, which remains unchanged upon RNA binding (Vitali et al. 2006). This fixed conformation in the free form was also observed in the case of nPTB (Joshi et al. 2014). The U2AF protein presents a case where a change in conformation occurs between the free and bound states (Mackereth et al. 2011). Specifically, it has been observed that the tandem RRMs of U2AF65 can adopt two distinct structures in solution depending on the presence or absence of a high-affinity RNA ligand. In the free state, the tandem RRMs adopt a closed conformation, which occludes the RNA binding β-sheet surface of RRM1 due to its interaction with the α-helices of RRM2. However, in the presence of a high-affinity RNA ligand, the tandem RRMs adopt an open conformation, where the β-sheet surfaces of the two RRMs lie side-by-side, forming an extended RNA-binding surface.

The structures of Npl3 tandem RRMs both in the free form and in complex with RNA have been recently solved (Keil et al. 2023; Moursy A. et al. 2023). Npl3 is a SR-like protein present in yeast; similarly to SRSF1, Npl3 has a canonical and a pseudo RRMs; the linker region that connects the two RRMs is only 8 residues (SKLPAKRY). The short length of the linker suggests that the system is less dynamic than the one studied in this project. Structures of the two isolated RRMs of Npl3 were previously determined in their free form (Deka et al. 2008; Skrisovska and Allain 2008) and more recently also the structure of the tandem RRMs (Keil et al. 2023), which was obtained with NMR-PRE and small angle X-ray scattering SAXS data. In the free form the β-sheet surfaces of the two RRM domains face towards each other, forming a positively charged surface. The two domains adopt a rather fixed orientation, and the linker connecting the two RRM domains has reduced flexibility (Figure 2.17A). In our laboratory, using NMR spectroscopy, we determined the structure of Npl3 tandem RRMs bound to 5´-AUCCAGUGGAA-3´ RNA (Figure 2.17B): in general, RRM1 binds preferentially upstream of the RRM2 binding site, with the linker domain contributing to the RNA binding. In addition, structure-guided studies revealed that mutations within RRM1, but not RRM2, negatively impact Npl3 function (Moursy A. et al. 2023). Upon comparing the structures of Npl3 in the free form and in complex with RNA, it is clear that the two RRMs are pre-organized to facilitate RNA binding, and only a minor conformational change is required to accommodate the RNA.

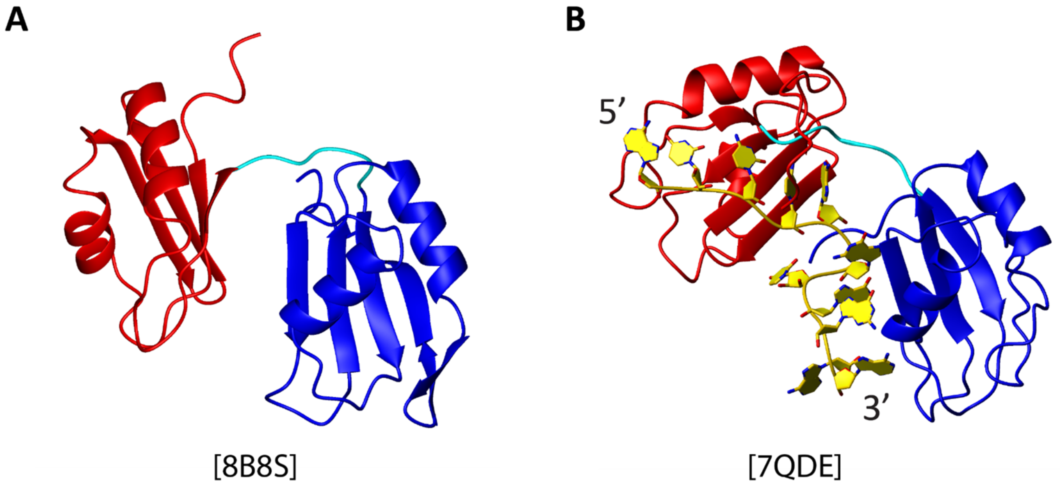


Figure 2.17 Npl3 tandem RRMs in the free form and in complex with RNA. *(A) Structure of Npl3 tandem RRMs in the free form (Keil et al. 2023). (B) Structure of Npl3 tandem RRMs bound to 5´-AUCCAGUGGAA-3´ RNA (Moursy A. et al. 2023). RRM2 is depicted in blue, with RRM1 in red, the linker in cyan and RNA in yellow. The PDB accession codes are indicated at the bottom of each structure.*

From the length and the sequence of the linker we initially expected that SRSF1 tandem RRMs was one of the cases in which the domains are totally independent. However, based on the correlation time experiments conducted on SRSF1, it was determined that this protein does not behave in the same manner. As already mentioned, the two RRMs in the free form showed already preferred conformations in which the two RRMs can bind the RNA both with conformational selection and induced fit.

Analyzing then our results, although initial observations suggested that the proximity of the two RRMs prevented the binding to RNA, this initial assumption changed when we compared the free protein ensemble with the structure ensembles of the protein-RNA complexes. Analysis of the conformations adopted by the free protein indicated that binding to the 5'-UCAUUGGAU-3' RNA is primarily driven by conformational selection, whereas binding to the 5'-UGGAUUUUUCAU-3' RNA can be driven by both conformational selection and induced fit.

Most of the conformations with high probability present in the free state showed interaction between Trp134 side chain and RRM1; Trp134 is an important residue involved also in the RRM2-RNA binding. This means that the interactions present in the free form need to be removed to allow the protein to binding the RNA. Our conclusion is that the protein has preformed conformations that facilitate the binding to RNA.

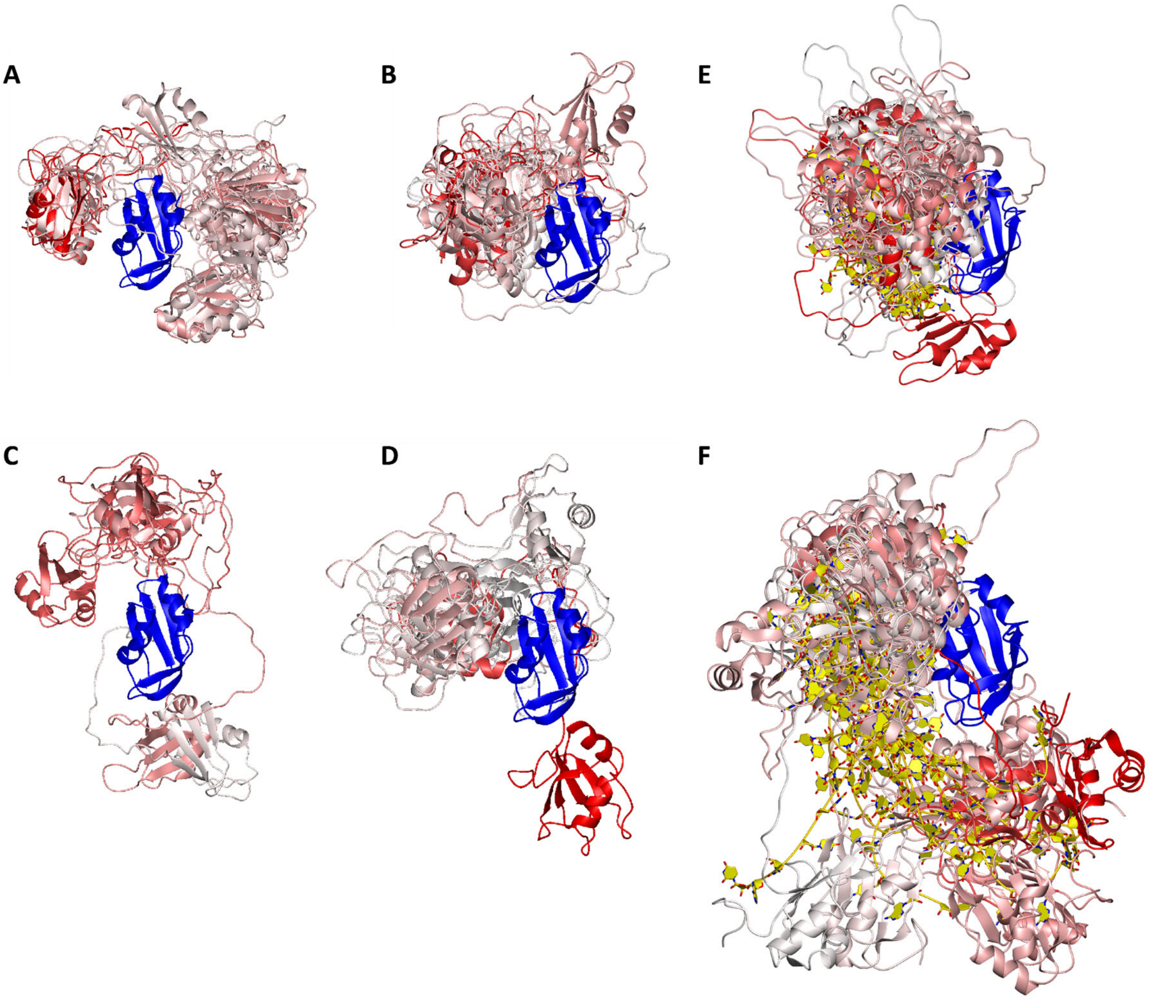


Figure 2.18 Comparison of the ensembles of SRSF1 RRM1+2 free form and in complex with RNA. *(A-D) Conformers of SRSF1 tandem RRMs in the free form. (A) Conformers not found in any protein-RNA complexes. (B) Conformers found only in the protein-RNA complex in presence of 5'-UCAUUGGAU-3' RNA. (C) Conformers found only in the protein-RNA complex in presence of 5'-UGGAUUUUUCAU-3' RNA. (D) Conformers not found in any protein-RNA complexes. (E) Ensembles of SRSF1 RRM1+2 in complex with 5'-UCAUUGGAU-3' RNA. (F) Ensembles of SRSF1 RRM1+2 in complex with 5'-UGGAUUUUUCAU-3' RNA. Ensembles are superimposed on RRM2 in blue, RRM1 and linker are in different intensity of red depending on the probability of each conformer (white = low probability, red = high probability), and RNA is in yellow; N-terminal (1-15) is not shown.*

We conducted a detailed comparison of the conformers shared between the free form and the two protein-RNA complexes (Figure 2.18). The final ensemble of the free form comprised 49 conformers with different probabilities (Figure 2.12B). Among these conformers, 16 were exclusively found in the complex with 5'-UCAUUGGAU-3' RNA, 6 were exclusively found in the complex with 5'-UGGAUUUUUCAU-3' RNA, and 11 were present in both protein-RNA complexes. Furthermore, 11 conformers were not detected in any of the protein-RNA complexes we calculated. Additionally, as mentioned earlier, all the conformers present in the complex with 5'-UCAUUGGAU-3' RNA were observed in the free ensemble, while 17 out of 23 conformers involving the 5'-UGGAUUUUUCAU-3' RNA were identified in the free form ensemble.

Based on a closer examination of the conformers, we found that SRSF1 tandem RRMs in the presence of 5'-UCAUUGGAU-3' RNA behaved similarly to Npl3 tandem RRMs. The RRMs are pre-oriented in the free form to accommodate this RNA sequence, with RRM1 binding upstream of RRM2, resulting in a low energy cost. Although there are similarities with Npl3, the main difference between the two systems is that for SRSF1, a single structure was not obtained in either the free form or the protein-RNA complex, unlike in the case of Npl3, which confirms the dynamic nature of SRSF1 tandem RRMs. In contrast, for the 5'-UGGAUUUUUCAU-3' RNA, some conformers showed that the RRMs were pre-oriented in the free form, while in others, a conformational change was required, which incurred a higher energy cost for the binding. This is similar to U2AF, where a change in the orientation of the RRMs was necessary for the binding to the RNA.

It has been previously reported that the two RRMs of other SR proteins show distinct mechanisms of RNA binding and display varying binding specificities when studied independently. The presence of two RRMs may enable SR proteins to interact with a wider range of RNAs compared to those with only one RRM. Alternatively, the two RRMs may enhance the specificity of SR proteins if both RRM recognition sequences are required for RNA targeting. Finally, one of the two RRMs may have evolved towards additional functions that are unrelated to RNA binding *in vivo* (Anko 2014). In addition, it is important to note that all our studies were conducted in the absence of the RS domain, which can interact with RRM2 inhibiting RNA binding when it is not phosphorylated (Serrano et al. 2016). This may promote certain conformations and hinder others. Our hypothesis is that SRSF1 can adopt any of the observed conformations, but one conformation may be preferred over others depending on the interaction with the RS domain, the cellular localization, and the RNA that SRSF1 can bind. This may also explain why we observe both conformational selection and induced fit.

2.4 Material and methods

2.4.1 Site-directed mutagenesis

Positions for site-directed spin labeling on the protein were chosen based on simulating the spin label attachment to existing NMR solution structures of individual RRMs of SRSF1 (He et al. 2005; Tintaru et al. 2007; Clery et al. 2013; Clery et al. 2021) using the software package Multiscale Modelling of Macromolecular systems (MMM) (Jeschke 2018). The following criteria were considered while scanning for potential labeling sites within the RRMs: potential mutation sites need to be accessible and in a folded region (α-helix or β-strand), the mutation to cysteine should not disrupt the structure of the protein, and the mutation should not affect the binding to RNA. Ideally, the three reference sites are positioned in a nearly equilateral triangle with the largest possible distance between the vertices. Alanine 107 was selected in the linker region since it does not show any influence in RNA binding and to not induce minimal perturbation on the linker due to the chemical similarity of the residues. Point mutations were introduced by a three-step polymerase chain reaction (PCR) protocol using primers carrying the desired base change. Phusion (NEB) or PFU (Promega) polymerases were used for the reaction. Subsequently, the DpnI-treated DNA was amplified in *E. coli* cells (strain TOP10) and isolated using a MiniPrep kit (MACHEREY-NAGEL). The incorporation of the correct mutation was verified by Sanger sequencing (Microsynth).

2.4.2 Expression and purification of the recombinant proteins

The coding sequences for SRSF1 RRM1 (amino acids 1-90 of SRSF1), RRM2 (amino acids 107-203 of SRSF1), and RRM1+2 (amino acids 1-196 of SRSF1) were cloned into the bacterial vector pET24 (Clery et al. 2013; Clery et al. 2021). All the recombinant proteins were fused to an N-terminal GB1 solubility tag followed by a 6xHis tag for the purification. In addition to increase the solubility, Y37S and Y72S point mutations were performed. All the recombinant proteins were overexpressed at 37°C in *E. coli* BL21 (DE3) codon plus cells in LB-broth medium (BD Difco) for non-isotopically labeled protein, or in M9 minimal medium supplemented with 1 g/L $^{15}NH_4Cl$ (Sigma-Aldrich or Cambridge Isotope Laboratories) and 4 g/L glucose for ^{15}N isotopically labeled protein. In addition, 50 µg/mL kanamycin and 50 µg/mL chloramphenicol were added to the media. The cell cultures were collected by centrifugation at 4000 rpm and 4°C for 15 to 20 minutes. The cell pellet was then resuspended in 20 mL of lysis buffer (50 mM Na2HPO4, 1 M NaCl, pH 8) and lysed using a microfluidizer. The resulting cell lysate was clarified by centrifugation at 17,000 rpm and 4°C for 45 minutes, and the protein of interest was purified using nickel affinity chromatography (QIAGEN). The protein solution was loaded onto a Nickel column and washed first with wash buffer A (20 mM Na_2HPO_4, 3 M NaCl, pH 8) to remove nucleic acids and then with 15 to 20 mL of wash buffer B (50 mM Na_2HPO_4, 50 mM L-Arg, 50 mM L-Glu, 1 M NaCl, 40 mM imidazole, pH 8) to remove nonspecifically bound proteins. The protein was then eluted with 10 to 15 mL of elution buffer (50 mM Na_2HPO_4, 1 M NaCl, 200 mM imidazole, pH 8) and dialyzed in dialysis buffer (50 mM Na_2HPO_4, 1 M NaCl, 200 mM imidazole, pH 8). The purity of the protein sample was checked by running it on a 12% SDS gel. The concentration of the recombinant proteins was carried out using 10-kDa molecular mass cutoff Centricon device (Vivascience). Finally, the protein samples were stored at −20°C until use.

2.4.3 Site-directed spin labeling of the protein samples

The reducing agent, 3 mM DTT, was eliminated from the dialysis buffer through the use of a desalting column filled with Sephadex G-25 resin, using PD10, PD MidiTrap, or PD MiniTrap column (GE Healthcare). Subsequently, the protein sample was incubated overnight at room temperature (RT) with a ten-fold excess of either MTSL ((1-oxyl-2,2,5,5-tetramethylpyrroline-3-methyl) methanethiosulfonate; Toronto Research Chemicals) or Gd(III)-maleimido-DOTA (1,4,7,10-tetraazacyclododecane-1,4,7-tris-acetic acid-10-maleimiodethylacetamide). Following this, the unreacted spin label was removed using a PD10 column (GE Healthcare) and the labeled protein was subsequently concentrated up to a maximum concentration of 200 µM via a 10-kDa molecular mass cutoff Centricon device (Vivascience). The labeling efficiency was determined using continuous-wave (CW) EPR spectroscopy for MTSL and Mass Spectrometry for Gd(III). For DEER samples, the final step was carried out in EPR buffer (20 mM NaPi, 50 mM L-Arg, 50 mM L-Glu, pH 6 with 10% glycerol), while for the PRE sample, the buffer was changed with PRE buffer (20 mM NaPi, 50 mM L-Arg, 50 mM L-Glu, pH 6).

2.4.4 Site-directed spin labeling of the RNA samples

The labelling of the RNA samples was performed by Dr. Laura Esteban-Hofer (PhD student in the group of Gunnar Jeschke, ETH Zürich). The details of the procedures are reported in the PhD Thesis (Esteban-Hofer 2022). Here we report the main steps. Site-directed spin labeling of RNA the two RNAs 5′-UCAUUGGAU-3′ and 5′-UGGAUUUUUCAU-3′ required to introduce of modifiers prior to spin labeling. The 5′-end was phosphorylated while the RNAs containing a 3′-end thiol modifier (C3 S-S) were purchased from Integrated DNA Technologies. The modified and deprotected RNAs were then spin labeled by addition of 3-(2-Iodoacetamido)-PROXYL (IAP). The labeling efficiency was determined by Continuous-Wave (CW) EPR experiments.

2.4.5 NMR experiments

All NMR experiments were conducted with the sample solubilized in NMR buffer comprising of 20 mM NaPi, 50 mM L-Arg, 50 mM L-Glu, and 3 mM DTT at pH 6. However, for PRE experiments, samples were prepared in the absence of DTT in 20 mM NaPi, 50 mM L-Arg, and 50 mM L-Glu at pH 6 to prevent spin-label reduction. In the case of protein-RNA complexes, a 1:1.2 ratio of protein and RNA was mixed to form the complex.

NMR spectra were acquired at 313 K on Bruker AVIII-500 MHz, AVIII-600 MHz, AVIII-700 MHz, and Avance-900 MHz spectrometers equipped with a cryoprobe. TopSpin 3.x or 4.x (Bruker) was used to process the spectra, and the analysis was performed using DynamicsCenter (Bruker), Sparky (http://www.cgl.ucsf.edu/home/sparky) and CARA (http://cara.nmr.ch/doku.php).

2.4.5.1 Chemical shift differences analysis

^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) experiments were performed to characterize and compare ^{15}N isotope-labeled samples of SRSF1 RRM1, SRSF1 RRM2, and SRSF1 RRM1+2. Backbone assignments of the isolated RRMs in the free and in complex with RNA were previously performed in a different buffer (Clery et al. 2013; Clery et al. 2021) and it was possible to transfer the. To account for differences between the spectra obtained in different conditions, peak positions were adapted for each set of spectra. Chemical shift differences between the different constructs were then calculated using the formula

$$\Delta CS = \sqrt{(\delta\mathrm{HN})^2 + (\delta\mathrm{N} / 6.51)^2}$$

where δHN is the chemical shift difference of the amide proton and δN represents the chemical shift difference of nitrogen atom.

2.4.5.2 ^{15}N-{^{1}H} Heteronuclear NOE experiments

Heteronuclear ^{15}N-{^{1}H} nuclear Overhauser effect (NOE) experiments were performed to measure the dynamics of the backbone of SRSF1 RRM1+2 construct. The experiment was conducted at a proton frequency of 700 MHz with a relaxation delay of 2 s and 3.5 s saturation period in the saturation experiment. The NOE values were then determined by calculating the ratios of peak intensities in the saturated spectrum to those in the unsaturated spectrum.

2.4.5.3 Protein dynamics study: ^{15}N spin-relaxation experiments

To investigate the correlation time of the isolated RRMs and SRSF1 RRM1+2 (free state and in complex with RNA), ^{15}N longitudinal relaxation times T_1 and transverse relaxation times T_2 were recorded at a ^{1}H frequency of 600 MHz using established and previously described protocols (Kay, Torchia, and Bax 1989; Skelton et al. 1993; Barraud and Allain 2013). For T_1, seven relaxation delays plus one repetition were measured in an interleaved manner per increment using times of 0.01/0.15(x2)/0.3/0.6/1/1.5 and 2 s in the order 6/(1,8)/4/2/3/7/5. For T_2, seven relaxation delays plus one repetition were measured in an interleaved manner per increment using times of 0/17/34/68(x2)/102/204 and 340 ms in the order 6/2/3/(1,8)/5/4/7. Peak lists were carefully adjusted, and data was extracted and fitted using DynamicsCenter (Bruker) using the intensity of the peak. The correlation time was then calculated from the obtained T_1 and T_2 times using equation

$$\tau_c = \frac{1}{4\pi\nu_N}\sqrt{6\frac{T_1}{T_2} - 7}$$

where ν_N is the ^{15}N resonance frequency (60.82 MHz).

2.4.5.4 PRE experiments

Transverse relaxation rates were extracted from ^1H_N-T_2 measurements of all the six single-cystein mutants labeled with diamagnetic and paramagnetic probes, both in the free form and in complex with RNA.

For the paramagnetic samples, seven relaxation delays plus one repetition were measured in an interleaved manner per increment using times of 0/7/8/10(x2)/12/14 and 20 ms in the order 1/6/3/(5,7)/8/4/2. For the diamagnetic samples, seven relaxation delays plus one repetition were measured in an interleaved manner per increment using times of 0//8/10(x2)/12/14/16 and 20 ms in the order 1/3/(5,7)/8/6/4/2. Peak lists were carefully adjusted, and data was extracted and fitted using DynamicsCenter (Bruker) using the intensity of the peak.

The PRE rate constants were determined by subtracting the transverse relaxation rates of the diamagnetic $R_{2,dia}$ and paramagnetic $R_{2,para}$ samples. This difference between the two rates is also represented as $\Gamma 2 = R_{2,para} - R_{2,dia}$. In instances, where amide peaks were excessively broadened and not detectable during the acquisition of $R_{2,para}$, we assigned a value of $\Gamma_2 = 170$ Hz.

The spin label-residues (inter-RRMs) distances were calculated using the formula

$$r = \left[\frac{K}{\Gamma_2}\left(4\tau_c + \frac{3\tau_c}{1 + {\omega_h}^2{\tau_c}^2}\right)\right]^{1/6}$$

where K is the magnetic susceptibility difference between the paramagnetic center and the surrounding medium, Γ_2 is the transverse relaxation rate of the nucleus, τ_c is the correlation time of the protein, ω_h is the Larmor frequency of the nucleus.

For the CYANA calculation the inter-RRM distances were calculated between the DUMMY atom Q8 of the labeled residue (see session 2.4.8.1) and the NH atom of the other residue.

To obtain the PRE ratios (I_{para}/I_{dia}), the integrated peaks of the plane obtained with the first timing point (6 ms) were normalized by the integral of the reference peak (chosen as the most intense peak in the GB1-tag). The PRE ratio I_{para}/I_{dia} was then calculated with the normalized integrals of the paramagnetic I_{para} and the diamagnetic sample I_{dia}.

2.4.6 EPR experiments

All the EPR experiments (CW and DEER) were performed and analysed by Dr. Laura Esteban-Hofer (group of Prof. Gunnar Jeschke, ETH Zürich). The details of the procedures are reported in the PhD Thesis (Esteban-Hofer 2022). Here we report the main steps.

2.4.6.1 Continuous-Wave (CW) experiments

CW experiments were performed to confirm the correct labelling of all the sample labelled with MTSL (both samples for PRE and DEER experiments). CW spectra were obtained at room temperature using a Bruker Elexsys E500 spectrometer equipped with a super high Q resonator ER4122SHQ at X-band frequencies (9.5 GHz). The protein samples were placed in a microcapillary tube (BLAUBRAND®) and the measurements were conducted using 100 kHz field modulation with a 1.5 G modulation amplitude. The labeling efficiency was determined by double integration of the spectrum and compared with a control sample of known concentration.

2.4.6.2 DEER sample preparation and measurements

All SRSF1 mutants were measured in a deuterated matrix. The nitroxide-nitroxide DEER measurements were performed on a homebuilt Q-band spectrometer (≈ 34 GHz) (Polyhach et al. 2012) on the shock-frozen sample at a temperature of 50 K. This was achieved by liquid-helium cooling and was controlled by a He flow cryostat (ER 4118CF, Oxford Instruments) and a temperature control unit (ITC 503, Oxford Instruments). A Hahn-echo sequence was utilized to obtain an echo-detected field-swept EPR spectrum. For the four-pulse DEER experiment, the pulse sequence followed $\pi/2_{obs}$-τ_1-π_{obs}-t_1-π_{pump}-$(\tau_1 + \tau_2 - t_1)$-π_{obs}-τ_2 (Pannier et al. 2000), where the pump pulse was applied at the spectral maximum and the observer pulses were applied at a frequency offset of 100 MHz. All measurements employed a 16 ns pump pulse, and either a 12 ns or 16 ns observer pulse, with a pulse delay τ_1 of 400 ns and a dead-time delay t_1 of 280 ns. An eight-step nuclear modulation averaging with a 16 ns averaging time step was employed, and the pulse delay τ_2 and the time scale increment were adjusted based on the expected distance.

For the Gd(III)-nitroxide DEER measurements, the experiments were conducted at 10 K with some adjustments made to the pulse sequence. The pump pulse was applied to the spectral maximum of the nitroxide spectrum, while the observer pulses were applied at the maximum of the Gd(III) spectrum, corresponding to a 300 MHz frequency offset. The pulse duration was set to 12 ns, and a $\pi/2$ pump pulse was used (Garbuio et al. 2013; Yulikov 2015).

2.4.6.3 DEER data analysis

The DeerLab software, which is a Matlab-based program (version 0.9.2) available at github.com/JeschkeLab/DeerLab-Matlab (Fabregas Ibanez, Jeschke, and Stoll 2020), was employed to analyze the DEER traces. The primary DEER data was subjected to zero-time and phase corrections, and the trace was cropped to remove the 2+1 artifact at the end. To obtain the distance distributions, a one-step analysis was carried out, wherein both the background and non-parametric distributions were fitted simultaneously. The background correction was based on a stretched exponential function, and the regularization parameter was determined using one of the following methods: the Bayesian information criterion, the residual method, or the L-curve minimum-radius method. Bootstrapped confidence intervals were estimated based on 1000 bootstrap samples drawn from a Gaussian distribution, using the noise in the measurements for all fitted signals, distributions, and parameters.

2.4.7 Modeling with *RigiFlex*

The ensembles obtained using the *RigiFlex* modelling were calculated and analysed by Dr. Laura Esteban-Hofer (group of Prof. Gunnar Jeschke, ETH Zürich). Detailed methodology is described in the PhD Thesis (Esteban-Hofer 2022). Here we report the main steps.

The ensembles were generated with the software package MMMx (version from 21 December 2021, available at https://github.com/gjeschke/MMMx; (Jeschke and Esteban-Hofer 2022) with dependencies on YASARA (Krieger et al. 2009), and SCWRL4 (Krivov, Shapovalov, and Dunbrack 2009). The ensemble generation, refinement, and EnsembleFit calculations were performed on the ETH Euler cluster. Visualization of the protein ensembles was done using VMD (Humphrey, Dalke, and Schulten 1996) or MOLMOL (MOLecular analysis and MOLecular display) (Koradi, Billeter, and Wuthrich 1996). In the ensemble generation, rigid bodies were first arranged relative to each other based on inter-RRMs distance restraints using the *Rigi* module. The number of rigid body arrangements used for free and protein-RNA complexes varied depending on the number of successfully generated conformers in smaller trials. For the protein-RNA complexes, the two RNA-binding motifs were connected using the *FlexRNA* module with either 5′-UU-3′ for the 5′-UCAUUGGAU-3′ RNA or 5′-UUUUU-3′ for the 5′-UGGAUUUUUCAU-3′ RNA. The RRMs were then connected by the inter-domain linker using the *Flex* module with all restraints involving the A107C mutant. The raw ensembles were refined using YASARA and contracted to a representative ensemble using the EnsembleFit module, which utilized both DEER and PRE restraints.

2.4.8 Modeling with Hybrid method

All the CYANA calculation were performed with the version CYANA 3.98.15.

2.4.8.1 DUMMY atoms to reproduce the MTSL spin label and calculation of the starting structure

The primary issue at hand involves the incorporation of the spin label in the CYANA calculation. To tackle this challenge, we built upon the approach described in a previous study (Dorn et al. 2022) and introduced additional steps. Specifically, we began by recalculating the domains of interest using constraints derived from NOEs defined in the structures of the individual RRMs (Free: He et al., 2005, Tintaru et al., Bound: 2007 Clery et al., 2013; Clery et al., 2017), while replacing the native cysteines and mutated cysteine residues with the new residue, `CYSM` (Appendix A.1.1). `CYSM` contains additional coordinates for the MTSL attached to the cysteine. The `CYSM` spin label atoms were unconstrained and sampled the available conformational space in the generated ensembles of 150 conformers. We then computed the average distance of atom `N1`, which contains partially the unpaired electron, since the unpaired electron is distributed between the N and O (`O1` or `Q1` could have also been used) and seven residues (`CA`) within the folded regions of the domain to define restraints from these atoms to the center of the cloud of the spin label (upl and lol). Unlike the approach taken in (Dorn et al. 2022), which involved the use of a long chain of pseudo-atoms for each spin label in the CYANA calculation, we incorporated additional residues (`CYSL`, `TYRL`, `SERL`, and `THRL`, Appendix A.1.2) into the CYANA library that include dummy-atoms (from `Q1` to `Q8` with Q1 attached to the residues CA atom) to be used for defining the position of the center of the cloud of the spin-label. We then recalculated the individual RRMs using including the atoms Q1-Q8 for each spin label (Appendix A.1.3). Similar to the approach described in Dorn et al. 2022) this method accounts for the average position of the unpaired electron of the spin label but does not consider the real dynamics of the spinlabel. The new approach using Q1-Q8 to position the center of the spin clouds dramatically reduces the number of coordinates required (only 8 additional coordinates per spinlabel which are fixed during calculations) compared to the previous method which for each spinlabel required long flexible linkers of approximately 50 residues containing dummy atoms followed by a GLY residue whose CA atom was constrained to place it at the center of each spin cloud. This was critical for allowing convergence of multistate calculations which included up to 10 copies of the protein and RNA as described below. After using the above method to define the location of each spinlabel center, we performed an additional calculation using the entire sequence of the SRSF1 RRM1+2 and the two different RNAs (5′-UCAUUGGAU-3′ and 5′-UGGAUUUUUCAU-3′). For the free protein, we used the original residue numbering, while for the protein-RNA complexes, we added +200 to the protein residue numbers (the RNA numbering was 1-9 and 1-12 in the case of 5′-UCAUUGGAU-3′ and 5′-UGGAUUUUUCAU-3′ respectively). 300 structures were generated, and the best 20 were selected based on the lowest target function and the number of violations. The structure with the best target function and the lowest number of violations was the used as the starting structure for the CYANA Multistate calculation. All protein structures were visualized with MOLMOL (Koradi, Billeter, and Wuthrich 1996).

2.4.8.2 Multistate CYANA calculation: PREP, CALC and SPLIT

We adapted the `PREP.cya` macro from Peter Güntert to prepare all the input data for the CYANA Multistate calculation (Strotz et al. 2017; Vogeli, Guntert, and Riek 2013). The data were prepared according to the number of states that were to be calculated. Except for the PRE data, which were treated as ambiguous restraints, all other data were treated as unambiguous restraints. The raw ensemble generation was performed on the ETH Euler cluster by adapting the `CALC.cya` macro. The .pdb file of the Starting structure was read, and the RRMs coordinates were fixed (residues 16-88 for RRM1 and residues 120-196 for RRM2). NOE-based restraints derived from structure determinations of the individual RRM-RNA complexes published previously were included to position the CA and GGA motifs close to the binding residues on the respective RRMs for the protein-RNA complexes. PRE input data were then read: six PRE datasets were used for the free protein, and three PRE datasets were used for the protein-RNA complexes.

To determine the number of states required in CYANA multistate calculations, tests were performed, calculating from 1 to 10 states in all three conditions (2000 structures with 200000 annealing steps, best 20 structures selected). The number of states that exhibited the best structural statistics and the structure without elongated states was selected.. In all three conditions, 5000 structures with 200000 annealing steps were generated, and the best 20 structures were selected. As all the 20 structures contained 7 or 3 states, we used the adapted `SPLIT.cya` macro (Appendix A.1.4.3) to obtain the initial CYANA model with the correct residue numbering (140 conformers for the free protein, 60 conformers for the protein-RNA complexes). Before running the EnsembleFit step, we removed all the dummy atoms.

2.4.8.3 MMMx EnsembleFit

The EnsembleFit step was performed using the EnsembleFit modeling function from the MMMx toolbox and run on MATLAB (Jeschke and Esteban-Hofer 2022). In this step, DEER restraints were also included. For the protein-RNA complexes, we had to change the residue numbering back to the original numbering. All the final ensembles were visualized with MOLMOL (MOLecular analysis and MOLecular display) (Koradi, Billeter, and Wuthrich 1996). Other analyses of the ensembles such as those showing ensembles pairwise distance statistics were performed with MMMx.

Chapter 3: SRSF1 and Phase Separation

SRSF1 and phase separation

Abstract

Nuclear speckles (NSs) are membraneless organelles located in the nucleus that act as central hub to coordinate various steps of nuclear gene expression regulation. These steps include chromosome localization, chromatin modification, transcription, splicing, 3'-end processing, mRNA modification, mRNA coating with proteins, and messenger ribonucleoprotein (mRNP) export. While NS proteins have been studied extensively to understand their functional characterization, the precise role of NSs requires further clarification. The assembly and maintenance of NSs depend on interactions among their different components, many of which contain flexible low-complexity regions (LCRs). LCRs play a critical role in protein-protein and protein-RNA interactions and can alter the properties of a protein permanently or transiently upon post-translational modification or protein partner binding. Proteins with LCRs are regulated to ensure that cellular processes can be adjusted, and depletion or mutation of LCRs can disrupt protein interactions, functions, and localization to NSs. High concentrations of macromolecules promote phase separation, resulting in the formation of liquid droplets with clear boundaries in aqueous solutions *in vivo* and *in vitro* when concentrated proteins with LCRs are present.

In this study, we focused on investigating SRSF1 as one of the main components of the nuclear speckles in the context of phase separation. Due to solubility challenges, we decided to analyze the behavior of the SRSF1 tandem RRMs using light microscopy, turbidity measurements, and NMR Diffusion-ordered spectroscopy (DOSY) experiments. Our results showed that even in the absence of the RS domain, the SRSF1 tandem RRMs form droplets *in vitro*, and RNA play an important role in dissolving the droplets.

Author contributions

Cristina K. X. Nguyen (C.K.X.N.) prepared protein samples, performed the turbidity measurements, microscopy analysis and NMR experiments. Antoine Cléry (A.C.) performed preparatory work. Leonidas Emmanouilidis (L.E) assisted with the design and analysis of the NMR experiments. Frédéric H.-T. Allain (F.H.-T.A.) provided financial support and overall supervision of the project. C.K.X.N. wrote the text with contributions from all the involved authors.

3.1 Introduction

The discovery of nuclear speckles (NSs) dates back to 1910 when Santiago Ramón y Cajal first observed them using light microscopy. Subsequently, in 1959, Hewson Swift identified NSs with the help of electron microscopy, but the term "speckles" was coined by J. Swanson Beck only two years later. Originally, NSs were believed to play a crucial role in regulating splicing factors at transcription sites, as changes in their function or composition resulted in alterations in alternative pre-mRNA splicing. However, as research on NSs progressed, additional functions were discovered. Recent studies have shown that NSs serve as a hub to coordinate all nuclear gene expression regulation steps, including chromosome localization, chromatin modification, transcription, splicing, 3'-end processing, mRNA modification, mRNA coating with proteins, and messenger ribonucleoprotein (mRNP) export. All these processes are all coupled with RNA polymerase II transcription, which occurs within perichromatin fibrils in close proximity to NSs (Biggiogera and Fakan 1998). While NSs proteins have been extensively studied to understand their functional characterization, the precise role of NSs requires further clarification. Additionally, most NS proteins can be found in other nuclear locations, and their specific roles in NSs, interacting partners, and post-translational modifications still need to be elucidated. NSs remain highly dynamic structures, and their components are constantly in flux. However, they remain clearly separated from the nucleoplasm, and when isolated from the nuclei of mouse liver cells, they are stable and resistant to subsequent purification procedures (Phair and Misteli 2000; Mintz et al. 1999; Saitoh et al. 2004). The assembly and maintenance of NSs depend on interactions among their different components, many of which contain flexible low-complexity regions (LCRs). LCRs play a critical role in protein-protein and protein-RNA interactions and can alter the properties of a protein permanently or transiently upon post-translational modification or protein partner binding. Proteins with LCRs are regulated to ensure that cellular processes can be adjusted, and depletion or mutation of LCRs can disrupt protein interactions, functions, and localization to NSs (Marzahn et al. 2016).

Another factor contributing to the spatial distinction of NSs is cellular crowding. High concentrations of macromolecules promote phase separation, resulting in the formation of liquid droplets with clear boundaries in aqueous solutions *in vivo* and *in vitro* when concentrated proteins with LCRs are present. Phase separation can be influenced by factors such as temperature, pH, ionic strength, and LCR modification (Li et al. 2012; Nott et al. 2015). Post-translational modifications within LCRs, such as phosphorylation and methylation, have been shown to regulate protein-protein interactions and recruitment to NSs while increasing structural order (Gui, Lane, and Fu 1994; Colwill, Pawson, et al. 1996; Colwill, Feng, et al. 1996; Misteli and Spector 1996; Xiang et al. 2013). Multiple LCRs, such as serine/arginine-rich (RS) motifs and folded domains, can coexist within individual NS proteins and enable interactions with multiple proteins simultaneously.

Regarding SRSF1, the phosphorylation of the RS domain by specific kinases plays a crucial role in regulating both its presence in the NSs and its localization within the cell. The RS domain is a characteristic domain of SRSF1, which is 50 residues long in this protein but varies in length among other members of the family. SRPK1 and cdc2-like kinase (Clk) are involved in the phosphorylation of the RS domain, and the subcellular localizations and substrate specificities of these two kinase families are distinct (Figure B1) (Gui, Lane, and Fu 1994; Colwill, Pawson, et al. 1996; Colwill, Feng, et al. 1996; Sanford and Bruzik 1999). SRPK1 is detected in both the cytoplasm and the nucleus, whereas Clk is constitutively located in the nucleus and co-localizes with SR proteins in nuclear speckles. SRPK1 binds to SRSF1 with unusually high affinity and rapidly modifies about 10-12 serines in the N-terminal portion of the RS domain (RS1) using a mechanism that incorporates sequential, C-to-N phosphorylation in several processive steps. In contrast, Clk/Sty phosphorylates the entire RS domain (RS1 and RS2). Nuclear CLKs are activated by osmotic and heat-shock stresses, so SR proteins are re-phosphorylated by CLKs during the recovery phase of stress (Ninomiya et al. 2020). The two SR kinase systems appear to act in symbiosis for proper phosphorylation of SR proteins and splicing regulation (Aubol et al. 2016), and their roles are interrelated, contributing in a coordinated way toward protein phosphorylation and localization in response to different stimuli (Corkery et al. 2015; Ghosh and Adams 2011). Additionally, de-phosphorylation and re-phosphorylation of SR proteins seem to be important for cytoplasmic functions (Huang, Yario, and Steitz 2004).

In a separate investigation, NMR spectroscopy was employed to analyze two isolated domains of SRSF1, revealing that certain residues in the RRM2 region interact with the N-terminal segment of the RS domain (RS1) *in trans*. Disturbing this intramolecular RRM2-RS domain interaction impedes both the directional phosphorylation mechanism and the nuclear translocation of SRSF1, emphasizing the necessity of the inherent phosphorylation bias for the biological function of SR proteins (Serrano et al. 2016).

3.2 Results

3.2.1 Phase separation of SRSF1 tandem RRMs and influence of the RNA

In our laboratory, all SRSF1 constructs, including individual RNA recognition motifs (RRMs), tandem RRMs, and full-length protein, were designed containing a GB1-tag and a His6-tag for the purpose of enhancing solubility and enabling purification. The presence of the GB1-tag is known to improve solubility and may affect artificially the behavior of the protein during phase separation studies. To investigate phase separation, we aimed to obtain the full-length SRSF1 protein, including its arginine-serine (RS) domain, which is known to be rich in RS residues and prone to phase separation. However, due to solubility issues, we began our experiments using constructs containing only the two RRMs, specifically SRSF1 RRM1+2 with Tyr37Ser (Y37S) and Tyr72Ser (Y72S) point mutations, which were used in previous experiments.

To study the SRSF1 tandem RRMs in the context of phase separation, we needed to obtain the protein in a soluble form without the GB1-tag. To achieve this, we cloned the constructs with a Tobacco Etch Virus (TEV) cleavage site between the His6-tag and the N-terminus of the protein (Figure 3.1 A). Obtaining the protein in a soluble form was a significant challenge, requiring adjustments of the salt conditions and protocols. Initial experiments were conducted under the same conditions as all the previous NMR experiments (Chapter 2) to evaluate the correct folding of the protein (Figure 3.1 B). We performed ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) experiments to compare the construct purified without the GB1-tag and the construct used in previous experiments, which was fused with the GB1-tag. The results indicated that the protein folded properly even in the absence of GB1.

To study the phase separation mechanism, we changed the buffer conditions to use a more physiological buffer. The protein was concentrated and stored in 50 mM Tris pH 8, 1 M NaCl, 3 mM DTT (250 µM) and then diluted 1:5 to a final concentration of 50 µM, with the salt concentration reduced to 200 mM.

The initial results are shown in Figure 3.1 C-D. The protein lacking the GB1-tag had the tendency to form droplets at low concentrations (20 µM), with round and dynamic droplets that fused with each other. Additionally, we measured the turbidity of the protein at different concentrations and we found that the turbidity increased linearly with protein concentration.

These results indicated that even though we initially believed that the RS domain was primarily responsible for the phase separation of SRSF1, *in vitro*, the two tandem RRMs were sufficient to form droplets at low concentration.

Once we demonstrated that the SRSF1 tandem RRMs construct forms droplets *in vitro*, we aimed to investigate whether RNA can influence this mechanism. Specifically, we tested the 5′-UCAUUGGAU-3′ and 5′-UGGAUUUUUCAU-3′ RNAs that we used to obtain structures of the protein-RNA complexes, designed based on the results of the SELEX experiments; in addition, we used 9-uridine (5′-UUUUUUUUU-3′) as control RNA that lacks binding sites for both RRMs.

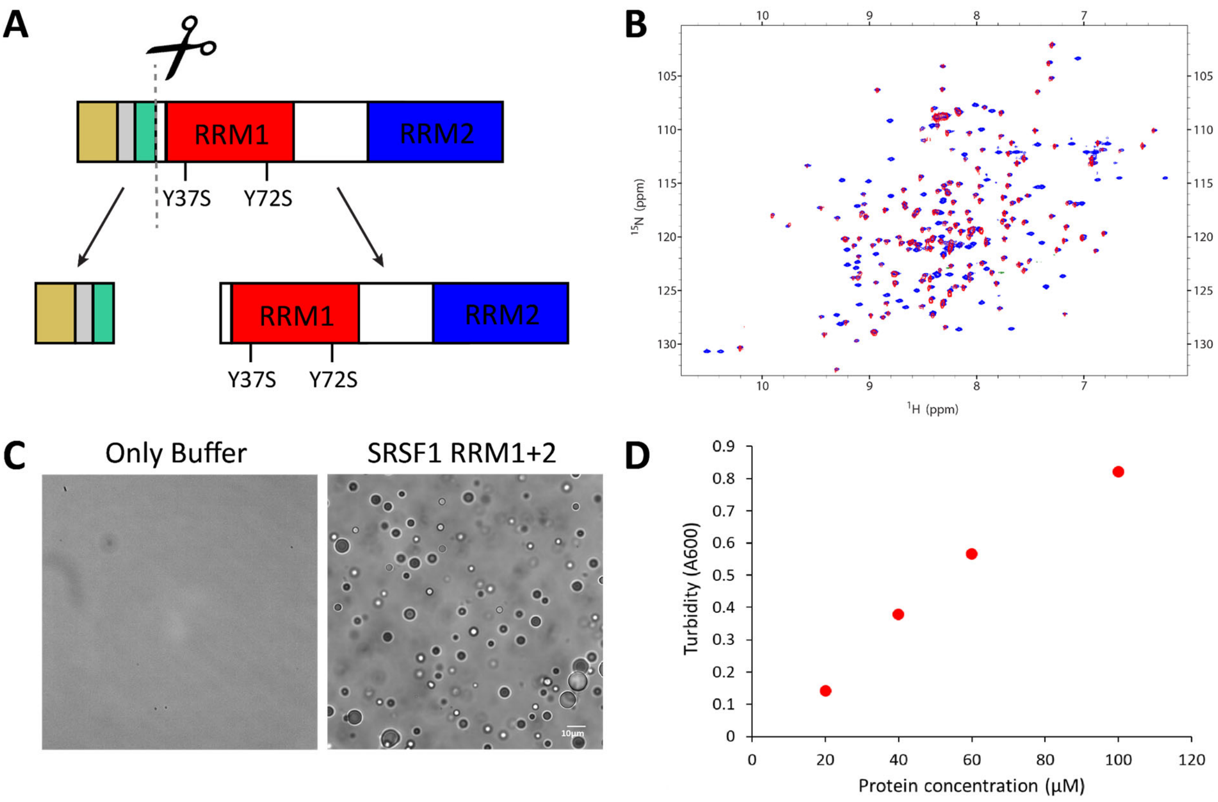


Figure 3.1 Characterization of SRSF1 RRM1+2 construct in absence of GB1. *(A) SRSF1 RRM1+2 (1-196) construct used in our experiments: the protein contains two point mutations in RRM1 (Tyr37 and Tyr72); the N-terminal GB1-tag (yellow) for solubility purposes and His6-tag (grey) for purification purposes are then removed after the TEV cleavage (TEV cleavage sequence is shown in green). (B) Overlay of* ^{1}*H-*^{15}*N HSQC spectra of SRSF1 RRM1+2 (1-196) fused with GB1 (blue) and in absence of GB1 (red). Buffer: 20 mM NaPi, 50 mM L-Arg, 50 mM L-Glu, 3 mM DTT, pH 6; Temperature = 313K. (C) Light microscopy images of the buffer alone as control on the left and SRSF1 RRM1+2 (50 μM). Buffer: 50 mM Tris, 200 mM NaCl, 3 mM DTT, pH 8; Temperature = RT. (D) Turbidity measurements (absorbance at 600 nm) of SRSF1 RRM1+2 at increasing concentrations. Buffer: 50 mM Tris, 200 mM NaCl, 3 mM DTT, pH 8; Temperature = RT.*

To test the protein-RNA binding, we performed ^{1}H-^{15}N HSQC experiments (Figure B2) and confirmed that the tandem RRMs bind both 5′-UCAUUGGAU-3′ RNA and 5′-UGGAUUUUUCAU-3′ RNA. In the absence of CA or GGA motifs, the polyU did not result in any chemical shift perturbation. To investigate the effect of RNA on droplet formation *in vitro*, we used both light microscopy and turbidity measurements (Figure 3.2 A-B). We gradually added RNA keeping the concentration of the protein fixed and observed the same result with both techniques. Addition of the 9-uridine RNA did not show any significant effect on droplet formation, as observed through microscopy and consistent turbidity measurements, indicating that this RNA does not affect droplet formation. A totally different result was observed upon addition of RNA containing both binding sequences for the two RRMs (5′-UCAUUGGAU-3′ and 5′-

UGGAUUUUUCAU-3'). In both cases, the droplets were less and smaller than those observed in the absence of RNA, even with a small amount of RNA (protein:RNA ratio 1:0.3) using microcopy. At the end of the titration (protein:RNA ratio 1:1), the droplets were entirely dissolved. This outcome was confirmed by the turbidity measurements (absorbance at 600 nm), showing a significant decrease in turbidity with increasing RNA concentration.

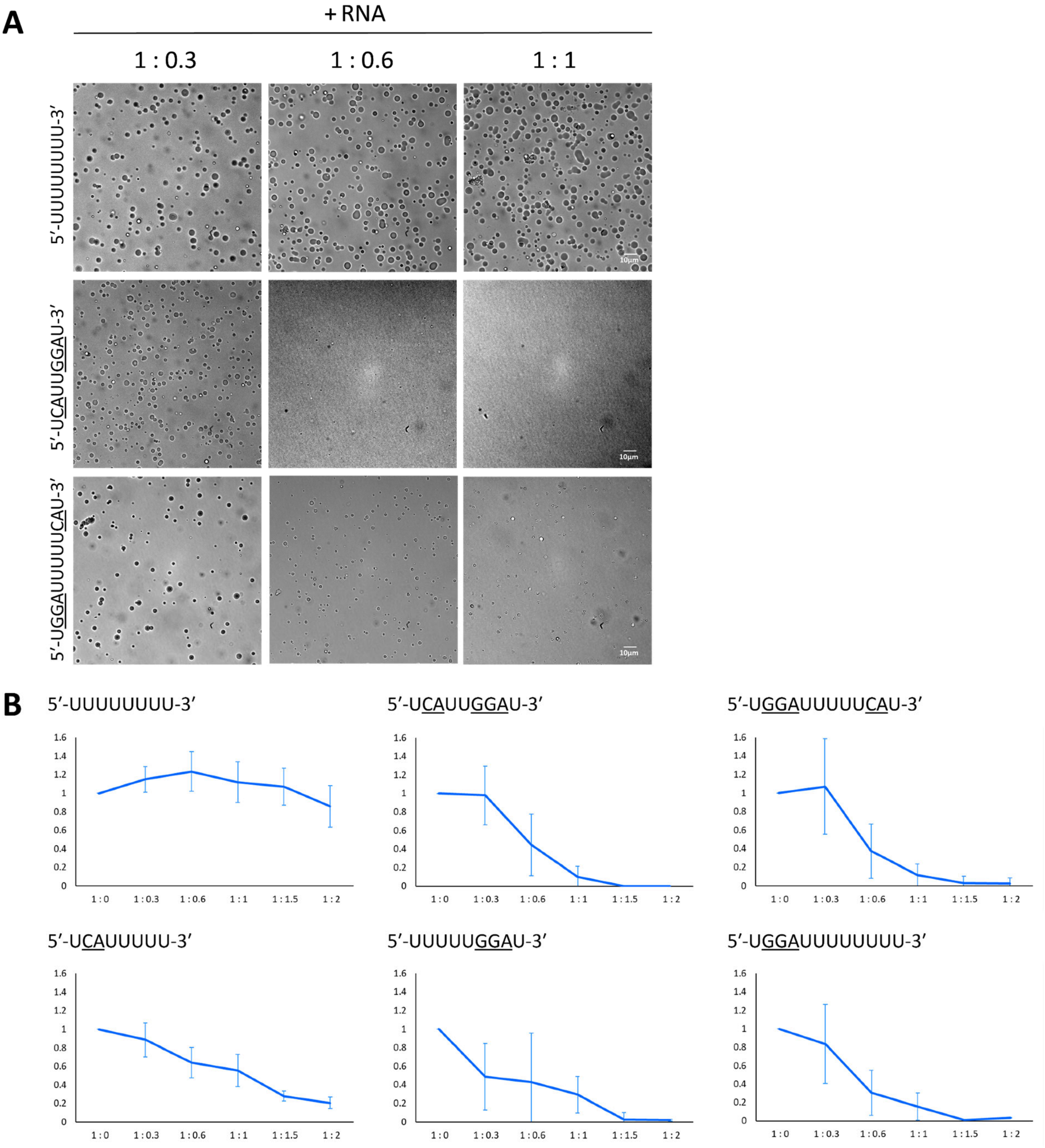


Figure 3.2 Phase separation of SRSF1 RRM1+2 in presence of different RNAs in vitro. *(A) Light microscopy images of SRSF1 RRM1+2 (50 µM) with increasing concentration of different RNAs: 5'-UUUUUUUUU-3', 5'-UCAUUGGAU-3' and 5'-UGGAUUUUUCAU-3'. (B) Turbidity measurements (absorbance at 600 nm) of SRSF1 RRM1+2 at fix concentration (50 µM) and increasing concentrations of different RNAs: 5'-UUUUUUUUU-3', 5'-UCAUUGGAU-3', 5'-UGGAUUUUUCAU-3', 5'-UCAUUUUUU-3', 5'-UUUUUGGAU-3' and 5'-UGGAUUUUUUUU-3'. All measurements were done in triplicate. Buffer: 50 mM Tris, 200 mM NaCl, 3 mM DTT, pH 8; Temperature = RT.*

To further control, we measured turbidity while adding RNAs containing only one binding sequence for a single RRM: either CA or GGA, with the other binding site replaced with UU or UUU. The RNAs tested were: 5′-UCAUUUUUU-3′, 5′-UUUUUGGAU-3′, and 5′-UGGAUUUUUUUU-3′.

Results indicated that when adding RNA containing only GGA, the turbidity decreased to 0 at a protein:RNA ratio of 1:1.5 in both cases (GGA at the 3′ or 5′). However, when adding RNA containing only the CA motif, the turbidity did not reach 0 even at a protein:RNA ratio of 1:2, unlike the previous cases.

These findings can be interpreted in various ways. Firstly, RNA lacking specific binding sequences for the RRMs does not impact droplet behavior *in vitro*, as indicated by both light microscopy and turbidity measurements. Secondly, RNA containing at least one binding site for either of the two RRMs has a dissolving effect on the droplets. However, the degree of this effect varies among the different RNAs tested, with the "detergent effect" being more pronounced when the RNA contains the GGA motif (with or without the CA motif). Conversely, RNA containing only the CA motif causes droplets to dissolve, but the turbidity does not drop to zero. This result may be attributed to the different binding affinities of the RRMs for their respective binding sequences. As mentioned in Section 2.1.2, the K_d values obtained were approximately 20 µM for RRM1 and 0.7 µM for RRM2; the droplets dissolve more rapidly in the presence of the GGA sequence due to the higher affinity of RRM2 for RNA.

These results were obtained for the SRSF1 tandem RRM constructs (SRSF1 RRM1+2). Further research is necessary to determine how the presence of the RS domain and its phosphorylation, which significantly affect protein solubility, may also impact the phase separation of SRSF1 in the presence or absence of RNA. This is currently being investigated in our laboratory as part of a new project.

3.2.2 NMR Diffusion experiments on stabilized droplets *in vitro*

The results presented in this paragraph are a subset of the research project published in (Emmanouilidis et al. 2021). Specifically, we focus on the results obtained for SRSF1 in this work. The authors, Dr. Leonidas Emmanouilidis and Dr. Laura Esteban-Hofer, investigate the *in vitro* droplet formation of the FUS N-terminal domain (NTD) using NMR and EPR techniques. To stabilize the droplets *in vitro*, the authors introduced 0.5% agarose in the buffer, which prevented droplet fusion and adhesion to the glass surface of the tube. The idea behind this approach was to mimic the cytoskeleton in cells to prevent the fusion of the droplets. This method of stabilizing droplets was also applied to other proteins, including PTPB1 and SRSF1 tandem RRMs, which contain folded domains. Figure 3.3 shows that droplets were still present in the sample after stabilizing with agarose for several hours. In addition, SRSF1 remained properly folded even in the presence of agarose-hydrogel (Figure B.3).

The conditions of droplets formation varied among the three samples studied: FUS was concentrated in a buffer containing 6M urea and decreasing the urea concentration facilitated phase separation. Similarly, SRSF1 RRM1+2 was concentrated in a buffer containing 1M NaCl to obtain droplets, and the buffer needed to be diluted to decrease the salt concentration. In contrast, PTPB1 in the free form was soluble, but droplets were formed *in vitro* after adding RNA.

To determine the quantity of protein responsible for the phase separation mechanism and included in the droplets, diffusion-based NMR experiments (DOSY) were performed on the FUS NTD sample. The experiments revealed that we can quantify the partitioning of intrinsically disordered proteins between the dispersed and condensed phases. However, for SRSF1 and PTBP1, which contain mainly folded regions, the approach was unable to detect the protein in the condensed phase. It was hypothesized that the fast backbone dynamics of the disordered domains keep the lines sufficiently narrow for NMR observation of FUS.

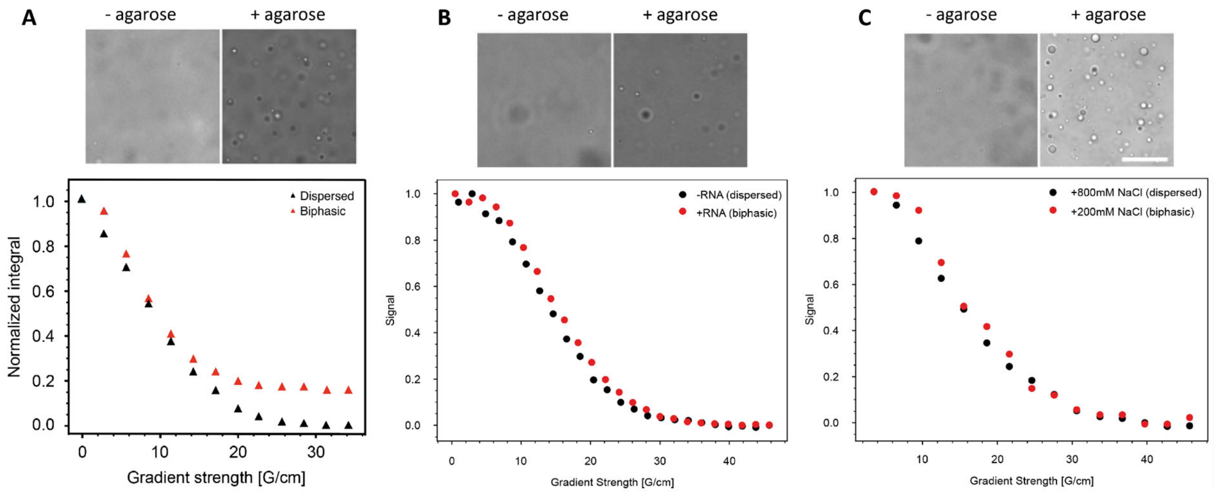


Figure 3.3 Agarose stabilization and DOSY experiments. *(A) FUS NTD, (B) PTB1 and (C) SRSF1 RRM1+2. Top: Effect of agarose hydrogel on stability of protein liquid droplets; Bottom: Integrated normalized spectral region plotted vs. gradient strength for dispersed (black) and biphasic (red). Modified from (Emmanouilidis et al. 2021).*

3.3 Discussion

In this study, our objective was to investigate the behavior of SRSF1 in the context of phase separation. Our goal was to understand the interactions of SRSF1 with other proteins within nuclear speckles. Due to the difficulty in obtaining the full-length protein in a soluble form, we decided to focus on analyzing the behavior of SRSF1 tandem RRMs in the context of phase separation. Based on the protein sequence, we initially hypothesized that the RS domain was primarily responsible for the phase separation phenomenon.

Our results revealed that the SRSF1 construct lacking the RS domain was able to form droplets *in vitro* at low concentrations. However, the most interesting finding was observed when we added RNA to the system. We tested RNAs containing one or both binding sequences for the two RRMs, as well as RNAs that did not contain any binding sequence. Surprisingly, we observed that even a single binding sequence was sufficient to induce droplet dissolution *in vitro*. We also found that the presence of GGA, which is specifically recognized by RRM2, led to faster dissolution of the droplets, likely due to the higher affinity of RRM2 for RNA.

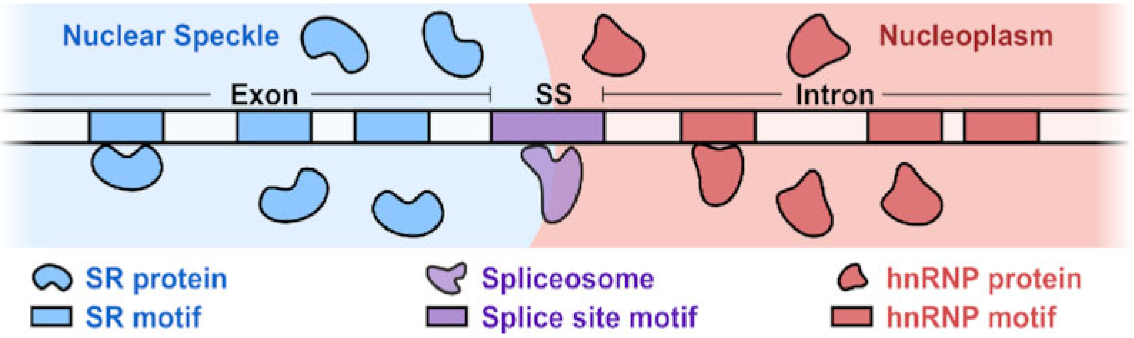


Figure 3.4 Interfacial splicing model. *The SR motif-rich exon is positioned inside the NS, whereas the hnRNP motif-rich intron is held outside in the nucleoplasm. The 3' or 5' splice site motif at an exon-intron boundary is positioned at the interface (Liao and Regev 2021).*

Our current research focus is to investigate the impact of the RS domain on the phase separation mechanism of SRSF1. We hypothesize that besides RNA binding, the RS domain may interact with other disordered regions of different SR proteins or other proteins. To explore this hypothesis, we plan to study the role of the RS domain in phase separation in both isolated form and in the context of the full-length protein. Furthermore, we aim to elucidate the influence of RS domain phosphorylation on the mechanism and binding with RNA and other proteins. By understanding the behavior of the RS domain alone or in the context of the full-length protein, and the SRSF1 tandem RRMs, we aim to understand the behavior of the protein in the context of phase separation and the role of the different domains.

In addition, a recent study (Liao and Regev 2021) proposed a model for the formation of nuclear speckles, where exons are preferentially sequestered into NSs through binding by SR proteins, while introns are excluded through binding by nucleoplasmic hnRNP proteins (Figure 3.4). This model suggests that splice sites at exon-intron boundaries are positioned at NS interfaces, which exposes the splice sites to interface-localized spliceosomes, enabling subsequent splicing reaction. In our lab, we have initiated a project to confirm this model.

3.4 Material and methods

3.4.1 Protein expression and purification

The coding sequence for SRSF1 RRM1+2 (amino acids 1-196 of SRSF1) were cloned into the bacterial vector pET24 with N-terminal GB1 solubility tag followed by a 6xHis tag and TEV cleavage sequence (ENLYFQG) for the purification. In addition to increase the solubility, Y37S and Y72S point mutations were performed. All the recombinant proteins were overexpressed at 37°C in E. coli BL21 (DE3) codon plus cells in LB-broth medium (BD Difco) for non-isotopically labeled protein, or in M9 minimal medium supplemented with 1 g/L $^{15}NH_4Cl$ (Sigma-Aldrich or Cambridge Isotope Laboratories) and 4 g/L glucose for ^{15}N isotopically labeled protein. In addition, 50 µg/mL kanamycin and 50 µg/mL chloramphenicol were added to the media.
The cell cultures were then collected by centrifugation (15 min to 20 min at 4000 rpm and 4 °C) and the resulting cell pellet was resuspended in 20 mL lysis buffer (50 mM Tris pH 8, 1 M NaCl, 3 mM DTT, 10 mM Imidazole).
After the cell lysis using a microfluidizer, the cell lysate was cleared by centrifugation (45 min at 17 000 rpm and 4 °C), and the protein purified using nickel affinity chromatography (QIAGEN). After loading the protein solution onto the Nickel column, the protein was washed first with wash buffer A (50 mM Tris pH 8, 3 M NaCl, 10 mM Imidazole, 3 mM DTT) to remove nucleic acids and then with 15mL to 20mL wash buffer B (50 mM Tris pH 8, 1 M NaCl, 30 mM Imidazole, 3 mM DTT) to remove unspecifically bound protein. The protein was finally eluted with 10mL to 15mL of elution buffer (50 mM Tris pH 8, 1 M NaCl, 200 mM Imidazole, 3 mM DTT). Subsequently, imidazole was removed and His-tag TEV cleaved simultaneously against the dialysis buffer (50 mM Tris pH 8, 1 M NaCl, 3 mM DTT) overnight at RT. The sample was loaded onto a second nickel column to remove the cleaved tag and any non-specifically binding molecules. Protein was collected from the flow-through and subsequently concentrated. The purity of the samples was analyzed on a 12 % SDS gel. Protein samples were stored at −80 °C until further use.

3.4.2 Light microscopy

An Olympus CKX41 microscope and Eclipse Ti Nikon microscope with ×40 air objective or a Widefield Zeiss LifeCell Station (from ScopeM facility, ETH Zürich) microscope x 10 x 20 ×40 immersion objective were used to visually inspect samples. Samples were loaded in a 384-well glass-bottom plate (Corning 4581).

3.4.3 Turbidity measurements

The turbidity (light scattering at 600 nm) of the samples was measured using a UV-Vis spectrophotometer (ND-1000 Spectrophotometer, NanoDrop). For SRSF1, a defined volume of protein was diluted in SRSF1 buffer (50 mM Tris pH 8, 200 mM NaCl, 3 mM DTT) to obtain a specific protein concentration (50 µM) in a total reaction volume of 10 µL. The RNA was added to perform different concentration points: 1:0, 1:0.3, 1:0.6, 1:1, 1:1.5. 1:2. All samples were done in triplicate.

3.4.4 Sample preparation in agarose gel

The preparation of the samples in agarose followed the protocol mentioned in (Emmanouilidis et al. 2021) adapted with SRSF1 conditions.

To form stabilized liquid droplets, agarose buffer (50 mM Tris pH 8, 200mM NaCl, 0.5% (wt/vol) agarose (ThermoFisher), 3 mM DTT) was boiled to solubilize agarose powder and then cooled in a room-temperature water bath. Agarose buffer was still liquid at ~55 °C. To form protein droplets inside the hydrogel, protein stock was diluted in this warm agarose buffer in a 1.5-ml pre-warmed Eppendorf and quickly transferred to either the sample tube or a glass-bottom multi-well plate. Owing to the small volume used, the temperature dropped quickly, leading to liquid-droplet formation and agarose gelation.

SRSF1 protein was dispersed at 50 µM concentration in the presence of high salt concentration; the agarose mixture for the dispersed sample contained 1 M NaCl, while the biphasic sample contained 200 mM NaCl.

3.4.5 NMR experiments

3.4.5.1 ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) experiments

^{15}N isotope-labeled SRSF1 RRM1+2 sample was characterized in NMR buffer by ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) experiments. Backbone assignments of SRSF1 tandem RRMs was performed in a different buffer (Figure 2.3A, Session 2.2.1) and the assignment could be transferred.

3.4.5.2 Diffusion-ordered spectroscopy (DOSY) experiments

The preparation of the samples in agarose followed the protocol mentioned in (Emmanouilidis et al. 2021).

A standard pulse sequence (**stebpgp1s19** from Topspin 3.2, Bruker) was used for diffusion experiments on the 750-MHz instrument. In total, 4,096 points with 32 scans were recorded in the proton dimension for each one dimension with variable diffusion gradient strength ranging between 2 and 95% in various steps. The following parameters were used: diffusion time (Δ) 0.08 s, gradient pulse (δ) 12 ms, smoothed rectangular-shaped gradients SMSQ 10.100, relaxation delay (d1) 5 s.

Chapter 4: Conclusions and Outlook

The majority of mRBPs implicated in alternative splicing contain two or more RNA-binding domains (RBDs, Fugure 2.1) (Lunde, Moore, and Varani 2007). It has been established that combining multiple RBDs increases the specificity and affinity of a protein (Jankowsky and Harris 2015). Among the different RBDs, the RNA recognition motif (RRM) is the most abundant, and structural data is available for the recognition of RNA by single and tandem RRMs (Afroz et al. 2015). These structural analyses have revealed that the RRM can bind RNA in various modes and recognizes diverse RNA sequences.

In the present work, we studied SRSF1 tandem RRMs; the structures of the individual RRMs, both in their unbound and RNA-bound states, had already been published in previous studies (He et al. 2005; Tintaru et al. 2007; Clery et al. 2013; Clery et al. 2021). Our initial hypothesis was that the two RRMs acted as independent domains due to the peculiar sequence of glycine-rich linker that connects them. However, our experiments involving correlation time, PRE, and DEER methods revealed that the two domains were not entirely free, but rather appeared to be in closer proximity than in the bound state. However, they did not behave as a single rigid body. Additionally, we observed that there was still some dynamic behavior even in the bound form of the protein. Unlike most RNA-binding proteins studied thus far, SRSF1 does not bind RNA in a unique manner. SELEX experiments had previously shown that the protein could use a bimodal mode to bind RNA, arranging the RRMs in different orders upstream and downstream, depending on the RNA sequence (Clery et al. 2021).

Upon analyzing the structures of the protein in both its free and RNA-bound forms, we observed that even in its free state, the protein exhibited preferred conformations in which the two RRMs were in close proximity. This finding indicates that the free form protein has conformations that may facilitate the RNA binding via a combination of conformational selection and induced fit mechanisms.

As previously mentioned in Chapter 2, all experiments and analyses were conducted on a truncated variant of the protein that lacked the RS domain. Further investigations are required to elucidate the role of the RS domain and how it may impact RNA binding.

In order to determine the structure of SRSF1 tandem RRMs, the classical CYANA approach was not optimal, as it is typically used for single and folded domains or systems that are not dynamic and attempts to find a unique final structure (Herrmann, Guntert, and Wuthrich 2002). Since the individual domains did not behave as a unique rigid body, it was apparent that both in the free form and in complex with RNA, the RRMs did not adopt unique conformations and allowed for more conformations. As a result, alternative methods were employed to determine ensembles.

Initially, the *RigiFlex* tool from MMMx modeling developed by Gunnar Jeschke (ETH Zürich) was utilized, and even if the final ensemble structures agreed with input data, it was not the optimal method due to the long-range DEER distance distributions (20-

100Å) used as the main restraints which did not provide sufficient understanding of molecular interactions since the two RRMs were positioned too far apart (Jeschke 2016, 2018, 2021; Jeschke and Esteban-Hofer 2022). Consequently, the more recent Multistate CYANA calculation by Peter Güntert (ETH Zürich) was utilized, which primarily used NMR data (in this case PRE data) that covered a shorter range than EPR data (Strotz et al. 2017; Vogeli, Guntert, and Riek 2013). While this method had been previously used for single and globular domains in the past to show conformations of disordered regions such as loops, this study was the first to apply it to a system with different domains that were dynamic, demonstrating its efficacy in calculating ensembles with multiple states and conformations.

To integrate data from multiple techniques, the EnsembleFit step from MMMx was used to fit the ensembles. DEER data were included as distributions (which was not possible using only CYANA as it works with distances, and different conformers were weighted with probability values based on their agreement with the input data. As with previous work from this laboratory such as the last study published in (Dorn et al. 2022), the combination of NMR and EPR approaches proved complementary and demonstrated their utility in determining structures of dynamic and multidomain systems.

In addition, we studied SRSF1 tandem RRMs in the context of phase separation. It is well known that SRSF1 and other SR proteins are key members of the nuclear speckles (NSs) and of the splicing events. In addition, some members of the serine-arginine protein family, including SRSF1, SRSF2, SRSF3, SRSF7, and SRSF10, are known to be recruited to cytoplasmic stress granules (SGs) together with non-translated mRNAs (Twyffels, Gueydan, and Kruys 2011). SRSF1 is recruited to SGs by binding to its target transcripts, while SRSF3 appears to regulate SG and P-body assembly (Jeong 2017). Disassembly of SGs upon stress relief is facilitated by the small ubiquitin-related modifier (SUMO) pathway, and SRSF1 plays a role in promoting SUMOylation by either recruiting the SUMO conjugating enzyme UBC9 or regulating the activity of the SUMO E3 ligase PIAS1 (Keiten-Schmitz et al. 2020; Keiten-Schmitz et al. 2021). Additionally, SRSF1 influences the SUMOylation of specific RNA-binding proteins (RBPs) and spliceosomal components, which is mediated by its RRM2 domain (Sliskovic, Eich, and Muller-McNicoll 2022).

In this part of the project, we utilized the truncated version of the protein lacking the RS domain. It is well known that arginine-rich disordered regions have a tendency to undergo phase separation *in vitro*. Our aim was to investigate the behavior of the SRSF1 tandem RRMs in the absence of the RS domain in the context of phase separation. Our findings demonstrate that the RRMs have the ability to undergo phase separation, and specific RNA sequences that can be bound by the protein can dissolve the resulting droplets *in vitro*. The propensity of the tandem RRMs to form droplets in the free form can be attributed to the existence of multiple conformations influenced by weak interactions. These results serve as a foundation for ongoing projects in our laboratory that focus on the phosphorylation of the RS domain and the interaction between the full-length protein and other proteins such as hnRNPA1 or FUS.

In summary, this study presents several novel findings. Firstly, a new approach utilizing an updated version of the well-established CYANA method is proposed to solve structures of dynamic multidomain proteins. Moreover, an EnsembleFit technique from the MMMx toolbox is employed to incorporate data from various experimental techniques. From a biological perspective, it is confirmed that SRSF1 exhibits a distinct bimodal mechanism for binding RNA, unlike other RNA binding proteins. NMR and EPR experiments were utilized to elucidate this mechanism, and the structures of both free and bound states were resolved. The results indicate that SRSF1 free form has conformations that facilitate RNA binding via conformational selection and induced fit. Additionally, the study explored the behavior of SRSF1 in the context of phase separation, demonstrating that the tandem RRMs form droplets *in vitro* even without the RS domain, and this behavior can be prevented by specific RNA sequences.

Chapter 5: Appendix

Appendix Chapter 2

A.1 Supporting data for CYANA calculation

A.1.1 Additional residues for the Library file (cyanamtslq2.lib)

Residue with MTSL (CYSM)

```
RESIDUE   CYSM    12   49    3   48
   1 OMEGA    0    0    0.0000 -O   -C    N    H
   2 PHI      0    0    0.0000 -C    N    CA   C
   3 CHI1     0    0    0.0000  N    CA   CB   SG   QQ5
   4 CHI2     0    0    0.0000  CA   CB   SG   SD   QQ5
   5 CHI3     0    0    0.0000  CB   SG   SD   CE   QQ5
   6 CHI4     0    0    0.0000  SG   SD   CE   C3   QQ5
   7 CHI5     0    0    0.0000  SD   CE   C3   C2   QQ5
   8 CHI21    0    0    0.0000  C3   C2   C21  H211 H213
   9 CHI22    0    0    0.0000  C3   C2   C22  H221 H223
  10 CHI51    0    0    0.0000  C4   C5   C51  H511 H513
  11 CHI52    0    0    0.0000  C4   C5   C52  H521 H523
  12 PSI      0    0    0.0000  N    CA   C   +N
   1 C    C_BYL    0    0.0000    0.0000    0.0000    0.0000 -O    N
   2 O    O_BYL    0    0.0000   -0.6699    0.0000   -1.0316 -C
   3 N    N_AMI    0    0.0000    1.3290    0.0000    0.0000 -C    H    CA
   4 H    H_AMI    0    0.0000    1.8071    0.0000    0.8555  N
   5 CA   C_ALI    0    0.0000    2.0987    0.0000   -1.2511  N    CB   HA   C
   6 HA   H_ALI    0    0.0000    1.8516   -0.8900   -1.8302  CA
   7 CB   C_ALI    0    0.0000    1.7522    1.2498   -2.0636  CA   SG   HB2  HB3
   8 HB2  H_ALI    0    0.0000    1.9516    2.1386   -1.4649  CB   -    -    -    QB
   9 HB3  H_ALI    0    0.0000    2.3604    1.2755   -2.9676  CB   -    -    -    QB
  10 QB   PSEUD    0    0.0000    2.1560    1.7070   -2.2163
  11 SG   S_RED    0    0.0000   -0.0042    1.2075   -2.5170  CB   SD
  12 SD   S_RED    0    0.0000   -0.9438    1.7359   -0.7735  SG   CE
  13 CE   C_ALI    0    0.0000   -0.9126    3.5483   -0.8267  SD   C3   HE2  HE3
  14 HE2  H_ALI    0    0.0000   -1.2985    3.9465    0.1125  CE   -    -    -    QE
  15 HE3  H_ALI    0    0.0000    0.1117    3.8904   -0.9719  CE   -    -    -    QE
  16 QE   PSEUD    0    0.0000   -0.5934    3.9184   -0.4297
  17 C3   C_BYL    0    0.0000   -1.7700    4.0323   -1.9679  CE   C2   C4
  18 C2   C_ALI    0    0.0000   -3.2790    4.1197   -1.9494  C3   C21  C22  N1
  19 Q21  PSEUD    0    0.0000   -3.8350    5.6469   -0.9858  -    -    -    -    QQ2
  20 Q22  PSEUD    0    0.0000   -4.0153    2.5576   -1.1806  -    -    -    -    QQ2
  21 C21  C_ALI    0    0.0000   -3.7291    5.3559   -1.1694  C2   H211 H212 H213
  22 H211 H_ALI    0    0.0000   -3.5042    5.2220   -0.1216  C21  -    -    -    Q21
  23 H212 H_ALI    0    0.0000   -3.2079    6.2257   -1.5412  C21  -    -    -    Q21
  24 H213 H_ALI    0    0.0000   -4.7930    5.4929   -1.2946  C21  -    -    -    Q21
  25 C22  C_ALI    0    0.0000   -3.8751    2.8551   -1.3270  C2   H221 H222 H223
  26 H221 H_ALI    0    0.0000   -4.9530    2.9208   -1.3407  C22  -    -    -    Q22
  27 H222 H_ALI    0    0.0000   -3.5598    1.9917   -1.8941  C22  -    -    -    Q22
  28 H223 H_ALI    0    0.0000   -3.5331    2.7603   -0.3070  C22  -    -    -    Q22
  29 QQ2  PSEUD    0    0.0000   -3.9252    4.1022   -1.0832
  30 N1   N_AMO    0    0.0000   -3.6853    4.2330   -3.3703  C2   O1   C5
  31 O1   O_BYL    0    0.0000   -3.9650    2.9150   -3.9419  N1
  32 Q1   DUMMY    0    0.0000   -3.8252    3.5740   -3.6561
  33 C4   C_BYL    0    0.0000   -1.3259    4.4436   -3.1314  C3   C5   H4
  34 H4   H_ALI    0    0.0000   -0.2858    4.4880   -3.4207  C4
  35 C5   C_ALI    0    0.0000   -2.4931    4.8439   -4.0041  N1   C4   C51  C52
  36 Q51  PSEUD    0    0.0000   -2.6639    6.7258   -4.0440  -    -    -    -    QQ5
  37 Q52  PSEUD    0    0.0000   -2.2635    4.1717   -5.7558  -    -    -    -    QQ5
  38 C51  C_ALI    0    0.0000   -2.6314    6.3674   -4.0364  C5   H511 H512 H513
  39 H511 H_ALI    0    0.0000   -2.5999    6.7532   -3.0281  C51  -    -    -    Q51
  40 H512 H_ALI    0    0.0000   -1.8195    6.7905   -4.6093  C51  -    -    -    Q51
  41 H513 H_ALI    0    0.0000   -3.5724    6.6338   -4.4946  C51  -    -    -    Q51
  42 C52  C_ALI    0    0.0000   -2.3072    4.2997   -5.4222  C5   H521 H522 H523
  43 H521 H_ALI    0    0.0000   -3.1800    4.5322   -6.0143  C52  -    -    -    Q52
  44 H522 H_ALI    0    0.0000   -1.4363    4.7542   -5.8709  C52  -    -    -    Q52
  45 H523 H_ALI    0    0.0000   -2.1741    3.2287   -5.3822  C52  -    -    -    Q52
  46 QQ5  PSEUD    0    0.0000   -2.4637    5.4488   -4.8999
  47 C    C_BYL    0    0.0000    3.5719    0.0000   -0.9342  CA   O   +N
  48 O    O_BYL    0    0.0000    3.9655    0.0000    0.2312  C
  49 N    N_AMI    0    0.0000    4.3963    0.0000   -1.9766  C
```

A.1.2 Residues with Q1-Q8 DUMMY atoms

CYSL

```
RESIDUE   CYSL    12   23    3   22
   1 OMEGA     0    0     0.0000 -O   -C    N    H
   2 PHI       0    0     0.0000 -C    N    CA   C
   3 CHI1      0    0     0.0000  N    CA   CB   SG   HG
   4 CHI2      0    0     0.0000  CA   CB   SG   HG   HG
   5 L1        0    0     0.0000  N    CA   Q1   Q2   Q8
   6 L2        0    0     0.0000  CA   Q1   Q2   Q3   Q8
   7 L3        0    0     0.0000  Q1   Q2   Q3   Q4   Q8
   8 L4        0    0     0.0000  Q2   Q3   Q4   Q5   Q8
   9 L5        0    0     0.0000  Q3   Q4   Q5   Q6   Q8
  10 L6        0    0     0.0000  Q4   Q5   Q6   Q7   Q8
  11 L7        0    0     0.0000  Q5   Q6   Q7   Q8   Q8
  12 PSI       0    0     0.0000  N    CA   C   +N
   1 C    C_BYL    0     0.0000    0.0000    0.0000    0.0000 -O    N
   2 O    O_BYL    0     0.0000   -0.6709    0.0000   -1.0328 -C
   3 N    N_AMI    0     0.0000    1.3283   -0.0000    0.0000 -C    H    CA
   4 H    H_AMI    0     0.0000    1.8071    0.0000    0.8552  N
   5 CA   C_ALI    0     0.0000    2.0929    0.0010   -1.2422  N    HA   CB   C
   6 HA   H_ALI    0     0.0000    2.6904    0.8996   -1.2620  CA
   7 CB   C_ALI    0     0.0000    3.0195   -1.2145   -1.2960  CA   HB2  HB3  SG
   8 HB2  H_ALI    0     0.0000    2.7705   -1.8096   -2.1624  CB   -    -    -    QB
   9 HB3  H_ALI    0     0.0000    4.0411   -0.8753   -1.3824  CB   -    -    -    QB
  10 QB   PSEUD    0     0.0000    3.4058   -1.3425   -1.7724
  11 SG   S_RED    0     0.0000    2.9142   -2.2877    0.1554  CB   HG
  12 HG   H_SUL    0     0.0000    2.0215   -1.7678    0.9840  SG
  13 Q1   DUMMY    0     0.0000    4.0922    0.0010   -1.1860
  14 Q2   DUMMY    0     0.0000    5.1405    0.0024   -2.8892
  15 Q3   DUMMY    0     0.0000    7.1396    0.0023   -2.8329
  16 Q4   DUMMY    0     0.0000    8.1880    0.0038   -4.5362
  17 Q5   DUMMY    0     0.0000   10.1872    0.0037   -4.4799
  18 Q6   DUMMY    0     0.0000   11.2355    0.0052   -6.1831
  19 Q7   DUMMY    0     0.0000   13.2348    0.0051   -6.1269
  20 Q8   DUMMY    0     0.0000   14.2831    0.0065   -7.8301
  21 C    C_BYL    0     0.0000    1.1644    0.0018   -2.4522  CA   O   +N
  22 O    O_BYL    0     0.0000    1.6188    0.0012   -3.5958  C
  23 N    N_AMI    0     0.0000   -0.1388    0.0016   -2.1915  C
```

TYRL

```
RESIDUE   TYRL    13   36    3   35
   1 OMEGA    0    0    0.0000 -O   -C     N    H
   2 PHI      0    0    0.0000 -C     N    CA   C
   3 CHI1     0    0    0.0000  N    CA   CB   CG   HH
   4 CHI2     0    0    0.0000  CA   CB   CG   CD1  HH
   5 CHI6     0    0    0.0000  CE1  CZ   OH   HH   HH
   6 L1       0    0    0.0000  N    CA   Q1   Q2   Q8
   7 L2       0    0    0.0000  CA   Q1   Q2   Q3   Q8
   8 L3       0    0    0.0000  Q1   Q2   Q3   Q4   Q8
   9 L4       0    0    0.0000  Q2   Q3   Q4   Q5   Q8
  10 L5       0    0    0.0000  Q3   Q4   Q5   Q6   Q8
  11 L6       0    0    0.0000  Q4   Q5   Q6   Q7   Q8
  12 L7       0    0    0.0000  Q5   Q6   Q7   Q8   Q8
  13 PSI      0    0    0.0000  N    CA   C   +N
   1 C    C_BYL    0    0.0000    0.0000    0.0000    0.0000 -O    N
   2 O    O_BYL    0    0.0000   -0.6703    0.0000   -1.0329 -C
   3 N    N_AMI    0    0.0000    1.3283   -0.0000    0.0000 -C    H    CA
   4 H    H_AMI    0    0.0000    1.8067    0.0000    0.8551  N
   5 CA   C_ALI    0    0.0000    2.0926    0.0026   -1.2418  N    HA   CB   C
   6 HA   H_ALI    0    0.0000    2.6924    0.9008   -1.2602  CA
   7 CB   C_ALI    0    0.0000    3.0189   -1.2133   -1.2975  CA   HB2  HB3  CG
   8 HB2  H_ALI    0    0.0000    2.7704   -1.8072   -2.1634  CB   -    -    -    QB
   9 HB3  H_ALI    0    0.0000    4.0413   -0.8748   -1.3805  CB   -    -    -    QB
  10 QB   PSEUD    0    0.0000    3.4059   -1.3410   -1.7719
  11 QD   PSEUD    0    0.0000    2.9109   -2.2068    0.0602  -    -    -    -    QR
  12 QE   PSEUD    0    0.0000    2.7523   -3.6572    2.0402  -    -    -    -    QR
  13 QR   PSEUD    0    0.0000    2.8316   -2.9320    1.0502
  14 CG   C_VIN    0    0.0000    2.9217   -2.1050   -0.0802  CB   CD1  CD2
  15 CD1  C_ARO    0    0.0000    2.0545   -1.7991    0.9615  CG   HD1  CE1
  16 HD1  H_ARO    0    0.0000    1.4448   -0.9096    0.8930  CD1  -    -    -    QD
  17 CE1  C_ARO    0    0.0000    1.9627   -2.6102    2.0760  CD1  HE1  CZ
  18 HE1  H_ARO    0    0.0000    1.2826   -2.3561    2.8757  CE1  -    -    -    QE
  19 CZ   C_VIN    0    0.0000    2.7426   -3.7448    2.1597  CE1  CE2  OH
  20 CE2  C_ARO    0    0.0000    3.6111   -4.0701    1.1387  CZ   HE2  CD2
  21 HE2  H_ARO    0    0.0000    4.2221   -4.9583    1.2047  CE2  -    -    -    QE
  22 CD2  C_ARO    0    0.0000    3.6975   -3.2523    0.0288  CG   CE2  HD2
  23 HD2  H_ARO    0    0.0000    4.3770   -3.5040   -0.7726  CD2  -    -    -    QD
  24 OH   O_HYD    0    0.0000    2.6548   -4.5560    3.2677  CZ   HH
  25 HH   H_OXY    0    0.0000    2.0111   -4.1923    3.8803  OH
  26 Q1   DUMMY    0    0.0000    4.0919    0.0026   -1.1855
  27 Q2   DUMMY    0    0.0000    5.1402    0.0061   -2.8888
  28 Q3   DUMMY    0    0.0000    7.1393    0.0060   -2.8326
  29 Q4   DUMMY    0    0.0000    8.1876    0.0096   -4.5358
  30 Q5   DUMMY    0    0.0000   10.1869    0.0095   -4.4796
  31 Q6   DUMMY    0    0.0000   11.2351    0.0131   -6.1829
  32 Q7   DUMMY    0    0.0000   13.2344    0.0130   -6.1266
  33 Q8   DUMMY    0    0.0000   14.2826    0.0165   -7.8298
  34 C    C_BYL    0    0.0000    1.1646    0.0051   -2.4526  CA   O   +N
  35 O    O_BYL    0    0.0000    1.6185    0.0080   -3.5965  C
  36 N    N_AMI    0    0.0000   -0.1385    0.0045   -2.1912  C
```

SERL

```
RESIDUE   SERL    12   23    3   22
   1 OMEGA    0    0    0.0000 -O   -C    N    H
   2 PHI      0    0    0.0000 -C    N    CA   C
   3 CHI1     0    0    0.0000  N    CA   CB   OG   HG
   4 CHI2     0    0    0.0000  CA   CB   OG   HG   HG
   5 L1       0    0    0.0000  N    CA   Q1   Q2   Q8
   6 L2       0    0    0.0000  CA   Q1   Q2   Q3   Q8
   7 L3       0    0    0.0000  Q1   Q2   Q3   Q4   Q8
   8 L4       0    0    0.0000  Q2   Q3   Q4   Q5   Q8
   9 L5       0    0    0.0000  Q3   Q4   Q5   Q6   Q8
  10 L6       0    0    0.0000  Q4   Q5   Q6   Q7   Q8
  11 L7       0    0    0.0000  Q5   Q6   Q7   Q8   Q8
  12 PSI      0    0    0.0000  N    CA   C   +N
   1 C     C_BYL    0    0.0000    0.0000    0.0000    0.0000 -O    N
   2 O     O_BYL    0    0.0000   -0.6701    0.0000   -1.0326 -C
   3 N     N_AMI    0    0.0000    1.3292    0.0000    0.0000 -C    H    CA
   4 H     H_AMI    0    0.0000    1.8069    0.0000    0.8561  N
   5 CA    C_ALI    0    0.0000    2.0937    0.0040   -1.2417  N    HA   CB   C
   6 HA    H_ALI    0    0.0000    2.6924    0.9025   -1.2595  CA
   7 CB    C_ALI    0    0.0000    3.0195   -1.2125   -1.2993  CA   HB2  HB3  OG
   8 HB2   H_ALI    0    0.0000    2.7707   -1.8106   -2.1625  CB   -    -    -    QB
   9 HB3   H_ALI    0    0.0000    4.0442   -0.8782   -1.3759  CB   -    -    -    QB
  10 QB    PSEUD    0    0.0000    3.4074   -1.3444   -1.7692
  11 OG    O_HYD    0    0.0000    2.8852   -2.0110   -0.1363  CB   HG
  12 HG    H_OXY    0    0.0000    2.2348   -1.6171    0.4495  OG
  13 Q1    DUMMY    0    0.0000    4.0929    0.0039   -1.1853
  14 Q2    DUMMY    0    0.0000    5.1414    0.0094   -2.8885
  15 Q3    DUMMY    0    0.0000    7.1406    0.0092   -2.8320
  16 Q4    DUMMY    0    0.0000    8.1890    0.0147   -4.5352
  17 Q5    DUMMY    0    0.0000   10.1883    0.0145   -4.4787
  18 Q6    DUMMY    0    0.0000   11.2367    0.0201   -6.1819
  19 Q7    DUMMY    0    0.0000   13.2360    0.0199   -6.1254
  20 Q8    DUMMY    0    0.0000   14.2844    0.0254   -7.8285
  21 C     C_BYL    0    0.0000    1.1644    0.0078   -2.4517  CA   O   +N
  22 O     O_BYL    0    0.0000    1.6174    0.0114   -3.5959  C
  23 N     N_AMI    0    0.0000   -0.1384    0.0069   -2.1893  C
```

THRL

```
RESIDUE   THRL    13   26    3   25
   1 OMEGA    0    0    0.0000 -O   -C    N    H
   2 PHI      0    0    0.0000 -C    N    CA   C
   3 CHI1     0    0    0.0000  N    CA   CB   OG1  HG23
   4 CHI21    0    0    0.0000  CA   CB   OG1  HG1  HG1
   5 CHI22    0    0    0.0000  CA   CB   CG2  HG21 HG23
   6 L1       0    0    0.0000  N    CA   Q1   Q2   Q8
   7 L2       0    0    0.0000  CA   Q1   Q2   Q3   Q8
   8 L3       0    0    0.0000  Q1   Q2   Q3   Q4   Q8
   9 L4       0    0    0.0000  Q2   Q3   Q4   Q5   Q8
  10 L5       0    0    0.0000  Q3   Q4   Q5   Q6   Q8
  11 L6       0    0    0.0000  Q4   Q5   Q6   Q7   Q8
  12 L7       0    0    0.0000  Q5   Q6   Q7   Q8   Q8
  13 PSI      0    0    0.0000  N    CA   C   +N
   1 C    C_BYL    0    0.0000    0.0000    0.0000    0.0000 -O    N
   2 O    O_BYL    0    0.0000   -0.6693   -0.0000   -1.0339 -C
   3 N    N_AMI    0    0.0000    1.3296    0.0000    0.0000 -C    H    CA
   4 H    H_AMI    0    0.0000    1.8063    0.0000    0.8561  N
   5 CA   C_ALI    0    0.0000    2.0938    0.0013   -1.2409  N    HA   CB   C
   6 HA   H_ALI    0    0.0000    2.6985    0.8965   -1.2592  CA
   7 CB   C_ALI    0    0.0000    3.0318   -1.2176   -1.3242  CA   HB   OG1  CG2
   8 HB   H_ALI    0    0.0000    2.7578   -1.8050   -2.1889  CB
   9 QG2  PSEUD    0    0.0000    4.8227   -0.6725   -1.5108
  10 OG1  O_HYD    0    0.0000    2.8932   -2.0237   -0.1489  CB   HG1
  11 HG1  H_OXY    0    0.0000    2.2436   -1.6270    0.4367  OG1
  12 CG2  C_ALI    0    0.0000    4.4799   -0.7768   -1.4751  CB   HG21 HG22 HG23
  13 HG21 H_ALI    0    0.0000    4.5282    0.3020   -1.4903  CG2  -    -    -    QG2
  14 HG22 H_ALI    0    0.0000    5.0588   -1.1514   -0.6438  CG2  -    -    -    QG2
  15 HG23 H_ALI    0    0.0000    4.8810   -1.1682   -2.3983  CG2  -    -    -    QG2
  16 Q1   DUMMY    0    0.0000    4.0930    0.0013   -1.1841
  17 Q2   DUMMY    0    0.0000    5.1418    0.0031   -2.8871
  18 Q3   DUMMY    0    0.0000    7.1410    0.0030   -2.8303
  19 Q4   DUMMY    0    0.0000    8.1898    0.0049   -4.5331
  20 Q5   DUMMY    0    0.0000   10.1890    0.0048   -4.4763
  21 Q6   DUMMY    0    0.0000   11.2378    0.0067   -6.1793
  22 Q7   DUMMY    0    0.0000   13.2370    0.0066   -6.1225
  23 Q8   DUMMY    0    0.0000   14.2858    0.0085   -7.8254
  24 C    C_BYL    0    0.0000    1.1717    0.0026   -2.4547  CA   O   +N
  25 O    O_BYL    0    0.0000    1.6322    0.0044   -3.5958  C
  26 N    N_AMI    0    0.0000   -0.1328    0.0023   -2.2009  C
```

A.1.3 Restraints for DUMMY atoms representing center of spincloud (Q8)

A.1.3.1 Free Form (upl and lol)

```
# ------- CYS 16 --------
 16 CYSL CA     16 CYSL Q8      8.30         8.28
 16 CYSL Q8     20 VAL  CA     20.47        20.45
 16 CYSL Q8     30 LYS  CA     24.57        24.55
 16 CYSL Q8     44 ASP  CA     14.56        14.54
 16 CYSL Q8     58 PHE  CA     16.54        16.52
 16 CYSL Q8     69 ASP  CA     17.14        17.12
 16 CYSL Q8     86 VAL  CA     20.88        20.86

# ------- TYR 37 --------
 37 TYRL CA       37 TYRL Q8      8.52       8.50
 20 VAL  CA       37 TYRL Q8     21.43      21.41
 30 LYS  CA       37 TYRL Q8     11.99      11.97
 37 TYRL Q8       44 ASP  CA     15.54      15.52
 37 TYRL Q8       58 PHE  CA     20.50      20.48
 37 TYRL Q8       69 ASP  CA     19.30      19.28
 37 TYRL Q8       86 VAL  CA     22.43      22.41

# ------- TYR 72 -------
 72 TYRL CA     72 TYRL Q8     8.56         8.54
 20 VAL  CA     72 TYRL Q8    19.10        19.08
 30 LYS  CA     72 TYRL Q8    28.48        28.46
 44 ASP  CA     72 TYRL Q8    26.04        26.02
 58 PHE  CA     72 TYRL Q8    22.23        22.21
 69 ASP  CA     72 TYRL Q8    10.82        10.80
 72 TYRL Q8     86 VAL  CA    15.13        15.11

# ------- SER 126 --------
126 SERL CA    126 SERL Q8      8.42         8.40
126 SERL Q8    143 GLU  CA     24.52        24.50
126 SERL Q8    151 ASP  CA     19.57        19.55
126 SERL Q8    158 GLY  CA     13.97        13.95
126 SERL Q8    166 GLU  CA     26.51        26.49
126 SERL Q8    188 ALA  CA     13.57        13.55
126 SERL Q8    192 VAL  CA     12.26        12.24

# ------- CYS 148 --------
148 CYSL CA    148 CYSL Q8      8.20         8.18
143 GLU  CA    148 CYSL Q8     19.58        19.56
148 CYSL Q8    151 ASP  CA     15.74        15.72
148 CYSL Q8    158 GLY  CA     19.56        19.54
148 CYSL Q8    166 GLU  CA     17.02        17.00
148 CYSL Q8    188 ALA  CA     29.73        29.71
148 CYSL Q8    192 VAL  CA     22.23        22.21

# ------- THR 169 --------
169 THRL CA    169 THRL Q8      8.53         8.51
143 GLU  CA    169 THRL Q8     19.72        19.70
151 ASP  CA    169 THRL Q8     26.33        26.31
158 GLY  CA    169 THRL Q8     23.91        23.89
166 GLU  CA    169 THRL Q8     10.47        10.45
169 THRL Q8    188 ALA  CA     28.12        28.10
169 THRL Q8    192 VAL  CA     16.39        16.37
```

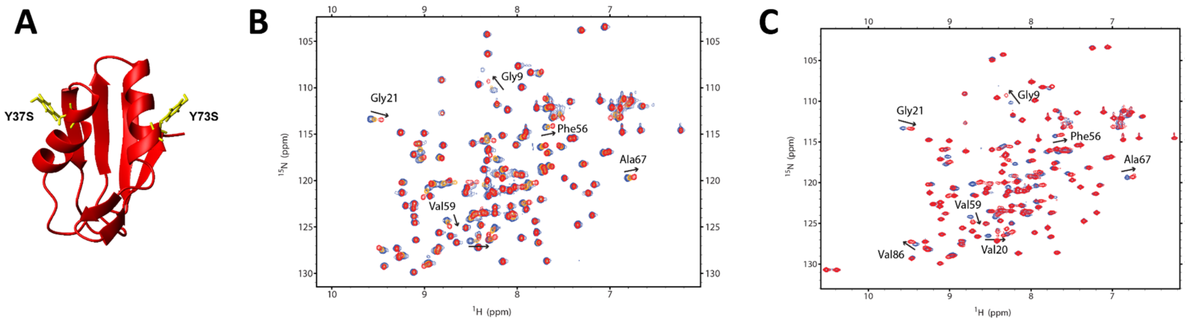
Residue name and number: original

A.1.3.2 Bound form (upl and lol)

```
# ------- CYS 16 --------
216 CYSL CA    216 CYSL Q8      8.66    8.64
216 CYSL Q8    220 VAL  CA     20.59   20.57
216 CYSL Q8    230 LYS  CA     24.12   24.10
216 CYSL Q8    244 ASP  CA     13.00   12.98
216 CYSL Q8    238 LYS  CA     19.01   18.99
216 CYSL Q8    269 ASP  CA     17.87   17.85
216 CYSL Q8    286 VAL  CA     20.96   20.94

# ------- TYR-SER 37 --------
237 SERL CA    237 SERL Q8      8.59    8.57
220 VAL  CA    237 SERL Q8     20.39   20.37
230 LYS  CA    237 SERL Q8     11.29   11.27
237 SERL Q8    244 ASP  CA     12.21   12.19
237 SERL Q8    238 LYS  CA      8.08    8.06
237 SERL Q8    269 ASP  CA     17.81   17.79
237 SERL Q8    286 VAL  CA     21.33   21.31

# ------- TYR-SER 72 --------
272 SERL CA    272 SERL Q8      8.34    8.32
220 VAL  CA    272 SERL Q8     18.58   18.56
230 LYS  CA    272 SERL Q8     28.44   28.42
244 ASP  CA    272 SERL Q8     25.53   25.51
238 LYS  CA    272 SERL Q8     21.15   21.13
269 ASP  CA    272 SERL Q8     11.40   11.38
272 SERL Q8    286 VAL  CA     14.64   14.62

# ------- SER 126 --------
326 SERL CA    326 SERL Q8      8.21    8.19
326 SERL Q8    343 GLU  CA     23.81   23.79
326 SERL Q8    361 GLU  CA     20.37   20.35
326 SERL Q8    358 GLY  CA     12.94   12.92
326 SERL Q8    366 GLU  CA     25.09   25.07
326 SERL Q8    388 ALA  CA     15.18   15.16
326 SERL Q8    392 VAL  CA     11.74   11.72

# ------- CYS 148 --------
348 CYSL CA    348 CYSL Q8      8.26    8.24
343 GLU  CA    348 CYSL Q8     17.14   17.12
348 CYSL Q8    361 GLU  CA     12.98   12.96
348 CYSL Q8    358 GLY  CA     21.06   21.04
348 CYSL Q8    366 GLU  CA     17.08   17.06
348 CYSL Q8    388 ALA  CA     29.14   29.12
348 CYSL Q8    392 VAL  CA     23.03   23.01

# ------- THR 169 --------
369 THRL CA    369 THRL Q8      8.59    8.57
343 GLU  CA    369 THRL Q8     20.22   20.20
361 GLU  CA    369 THRL Q8     17.28   17.26
358 GLY  CA    369 THRL Q8     22.92   22.90
366 GLU  CA    369 THRL Q8     10.10   10.08
369 THRL Q8    388 ALA  CA     27.77   27.75
369 THRL Q8    392 VAL  CA     16.22   16.20
```

Residue name: include Y37S and Y72S mutations (Clery et al. 2021); Residue number: +200

A.2 Supporting Figures



SRSF1 RRM1 YS + AACAAA

SRSF1 RRM1 WT + AACAAA

Figure A1 SRSF1-RRM1(YS) construct. *(A) Tyr37 and Tyr72 located in a-helices of RRM1 which are solvent accessible (B) Overlay of ^{1}H-^{15}N HSQC spectra measured with SRSF1 RRM1 Y37S+Y72S (YS) free form (in blue) and bound to AACAAA RNA at a 0.3:1 (orange) and 1:1 (red) RNA:protein ratio. (C) Overlay of ^{1}H-^{15}N HSQC spectra measured with SRSF1 RRM1 WT free form (in blue) and bound to AACAAA RNA at 0.3:1 (orange) and 1:1 (red) RNA:protein ratio. Similar chemical shift perturbations are observed at saturation (RNA:protein ratio of 1:1) showing that the mode of interaction with RNA and the affinity is similar for the WT and Y37S+Y72S versions of SRSF1 RRM1. Modified from (Clery et al. 2021).*

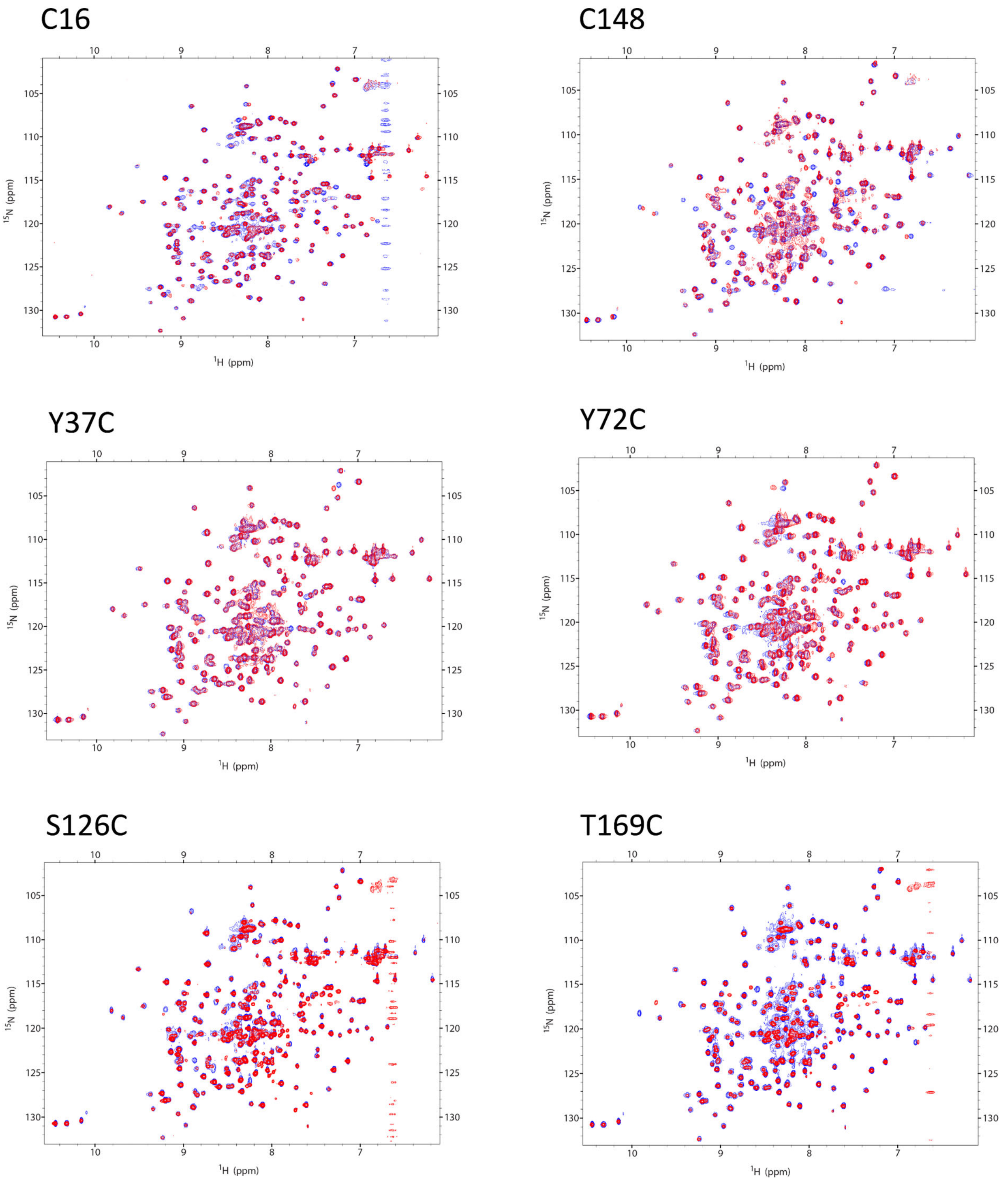


Figure A2 SRSF1 RRM1+2 single-cysteine mutants for PRE experiments. *Comparison of ^{1}H-^{15}N HSQC spectra of each SRSF1 RRM1+2 single-cysteine mutant (C16, Y37C, Y72C, S126C, C148 and T169C) in absence of MTSL (unlabeled sample = blue) and when the diamagnetic MTSL is attached (labeled sample = red).*

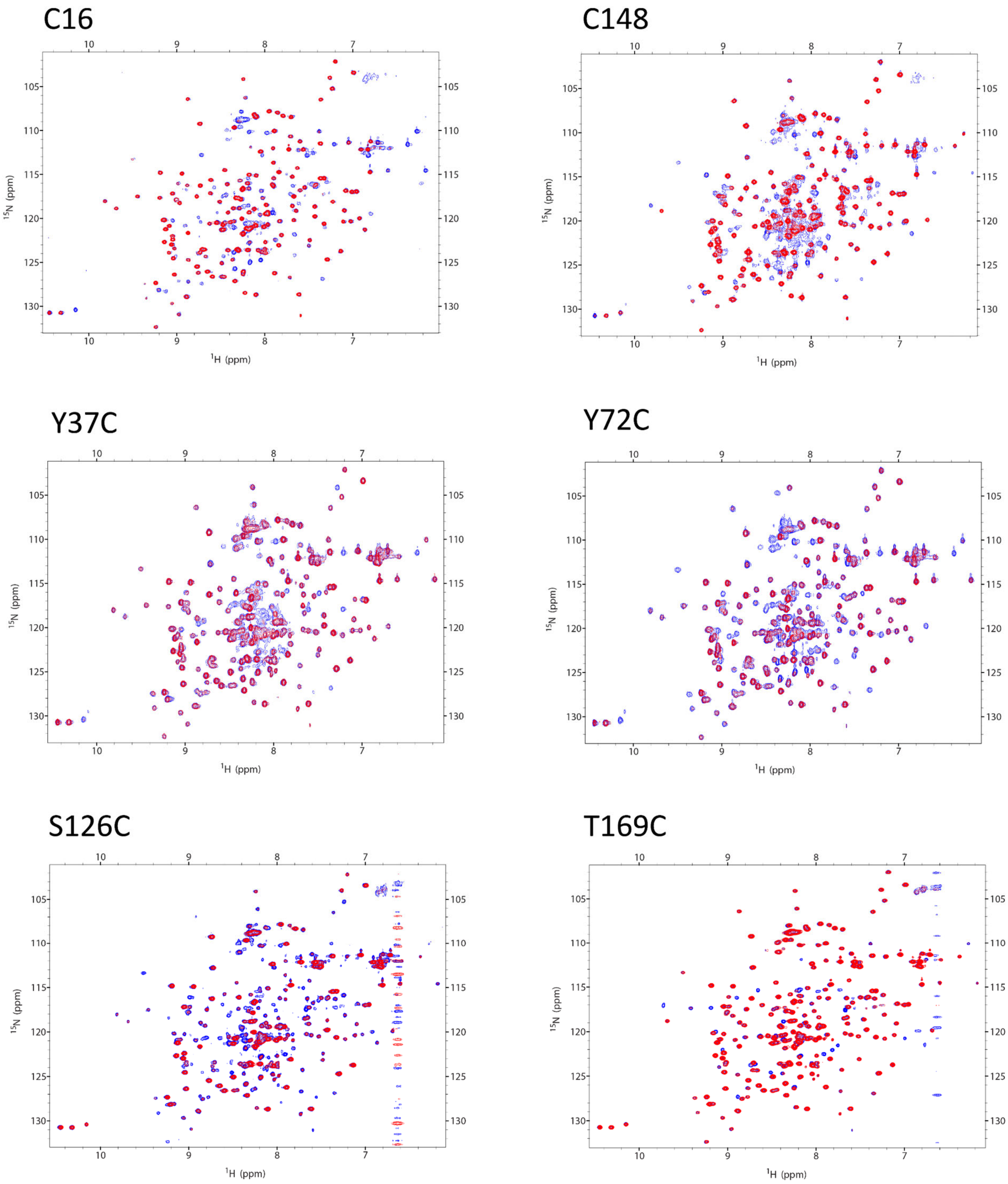
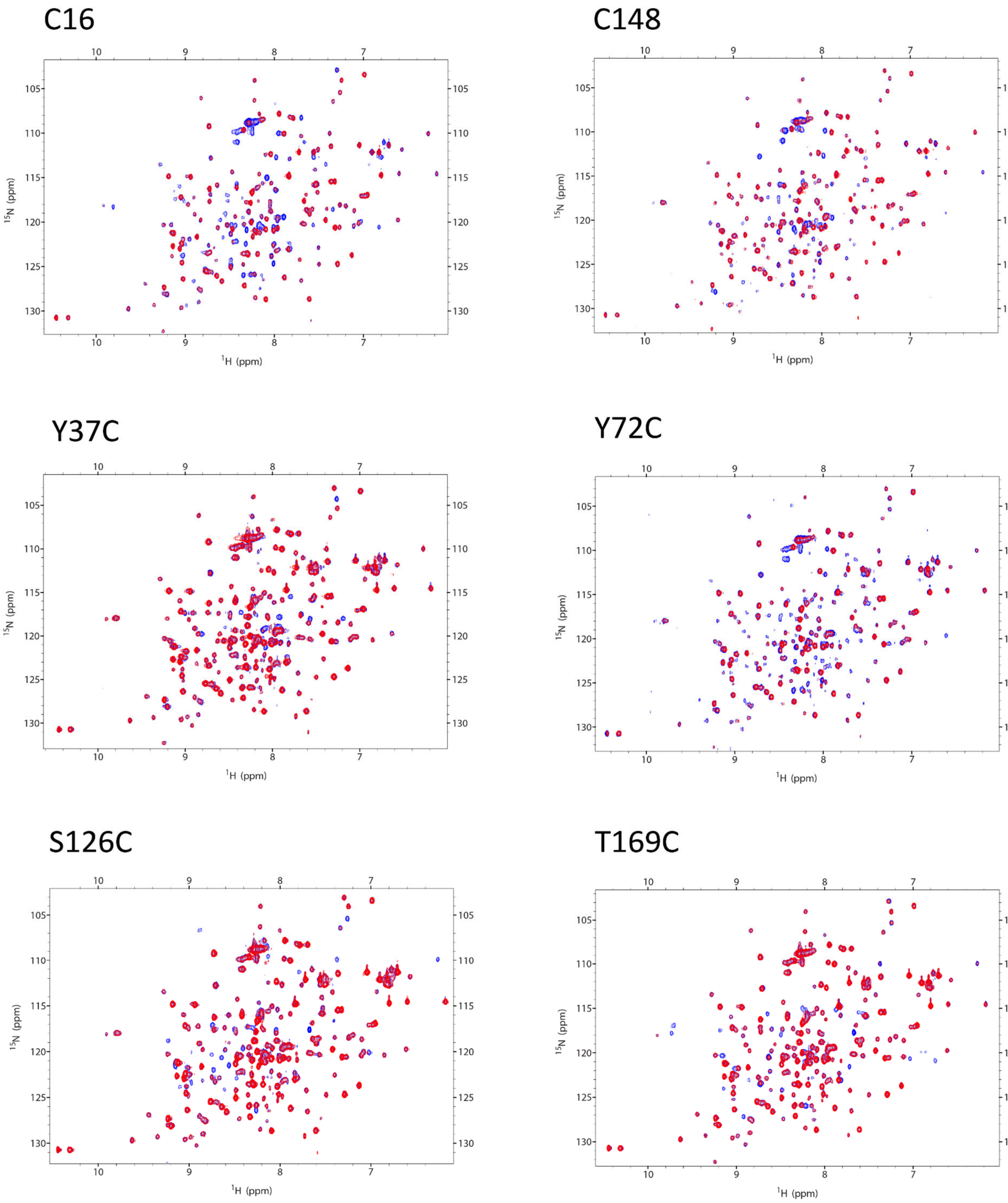


*Figure A3 **PRE experiments of SRSF1 RRM1+2 single-cysteine mutants in the free form.** Comparison of ^{1}H-^{15}N HSQC spectra of each SRSF1 RRM1+2 single cysteine mutant (C16, Y37C, Y72C, S126C, C148 and T169C) with the MTSL in the reduced state (diamagnetic sample = blue) and when the MTSL is in active state (paramagnetic sample = red).*


C16

C148

Y37C

Y72C

S126C

T169C

Figure A4 PRE experiments of SRSF1 RRM1+2 single-cysteine mutants in complex with 5'-UCAUUGGAU-3' RNA. *Comparison of ^{1}H-^{15}N HSQC spectra of each SRSF1 RRM1+2 single cysteine mutant (C16, Y37C, Y72C, S126C, C148 and T169C) with the MTSL in the reduced state (diamagnetic sample = blue) and when the MTSL is in active state (paramagnetic sample = red).*

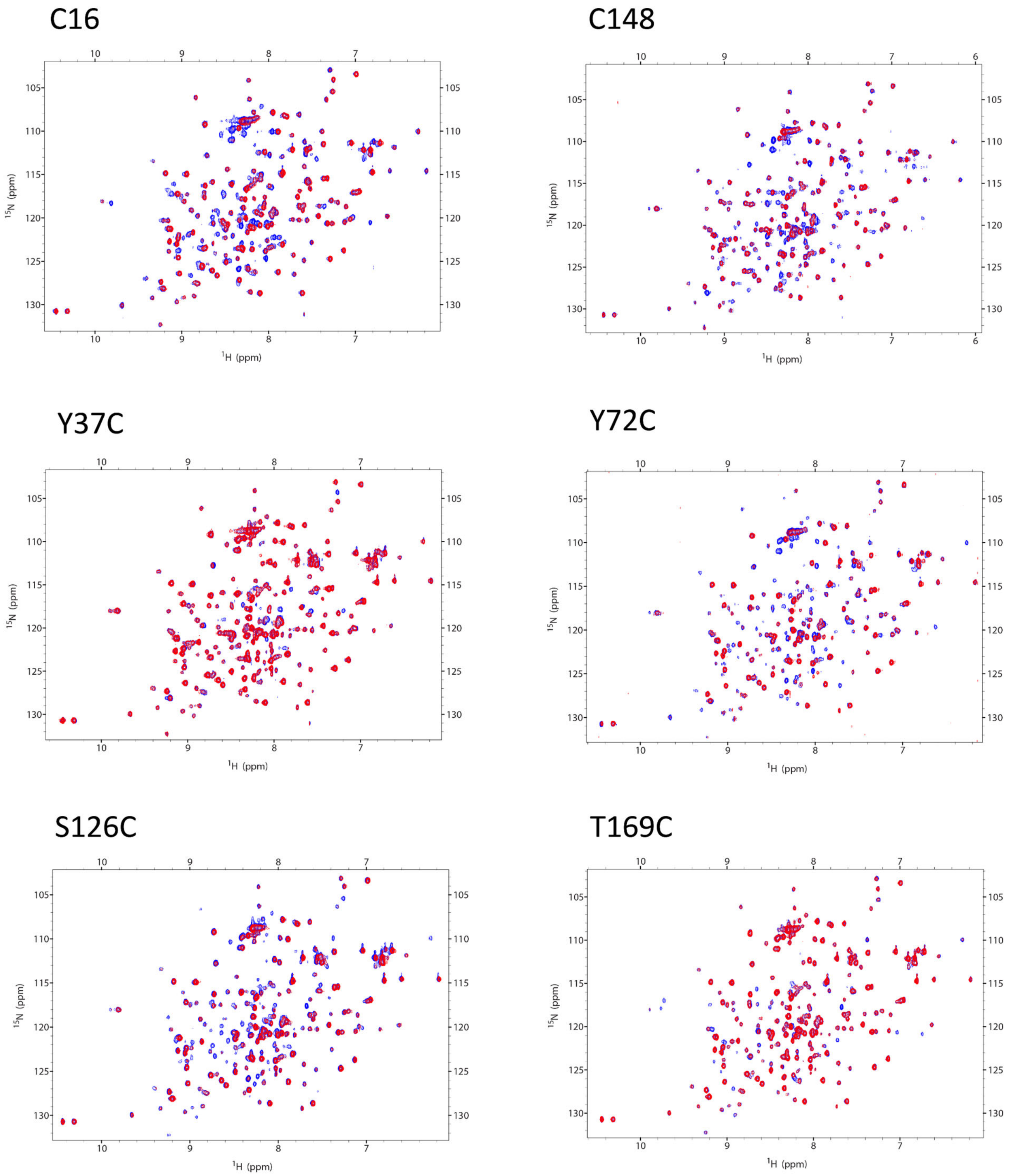


Figure A5 PRE experiments of SRSF1 RRM1+2 single-cysteine mutants in complex with 5'-UGGAUUUUUCAU-3' RNA. *Comparison of ^{1}H-^{15}N HSQC spectra of each SRSF1 RRM1+2 single cysteine mutant (C16, Y37C, Y72C, S126C, C148 and T169C) with the MTSL in the reduced state (diamagnetic sample = blue) and when the MTSL is in active state (paramagnetic sample = red).*

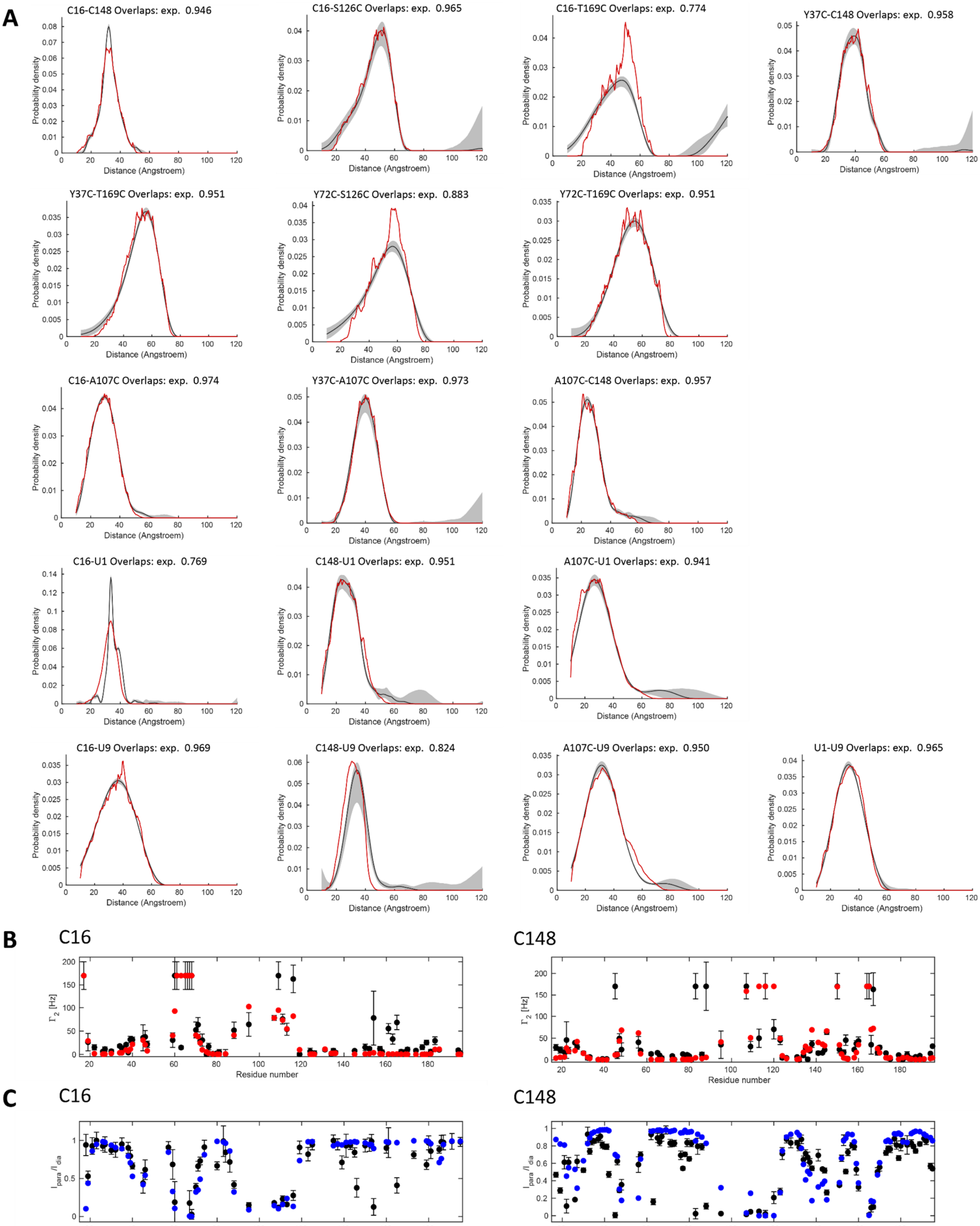


Figure A6 Restraint fit of the ensemble SRSF1 RRM1+2 in complex with 5'-UCAUUGGAU-3' RNA calculated using the RigiFlex modeling. *(A) DEER distance distribution (B) PRE rate restraints (C) PRE ratio restraints. (A-C) Experimentally determined distance distribution (black) and predicted for the entire ensemble (red or blue). Final ensemble = 88 conformers.* $1-\bar{o} = 0.079$, χ^2 *(PRE rate) = 5.05,* χ^2 *(PRE ratio) = 12.75*

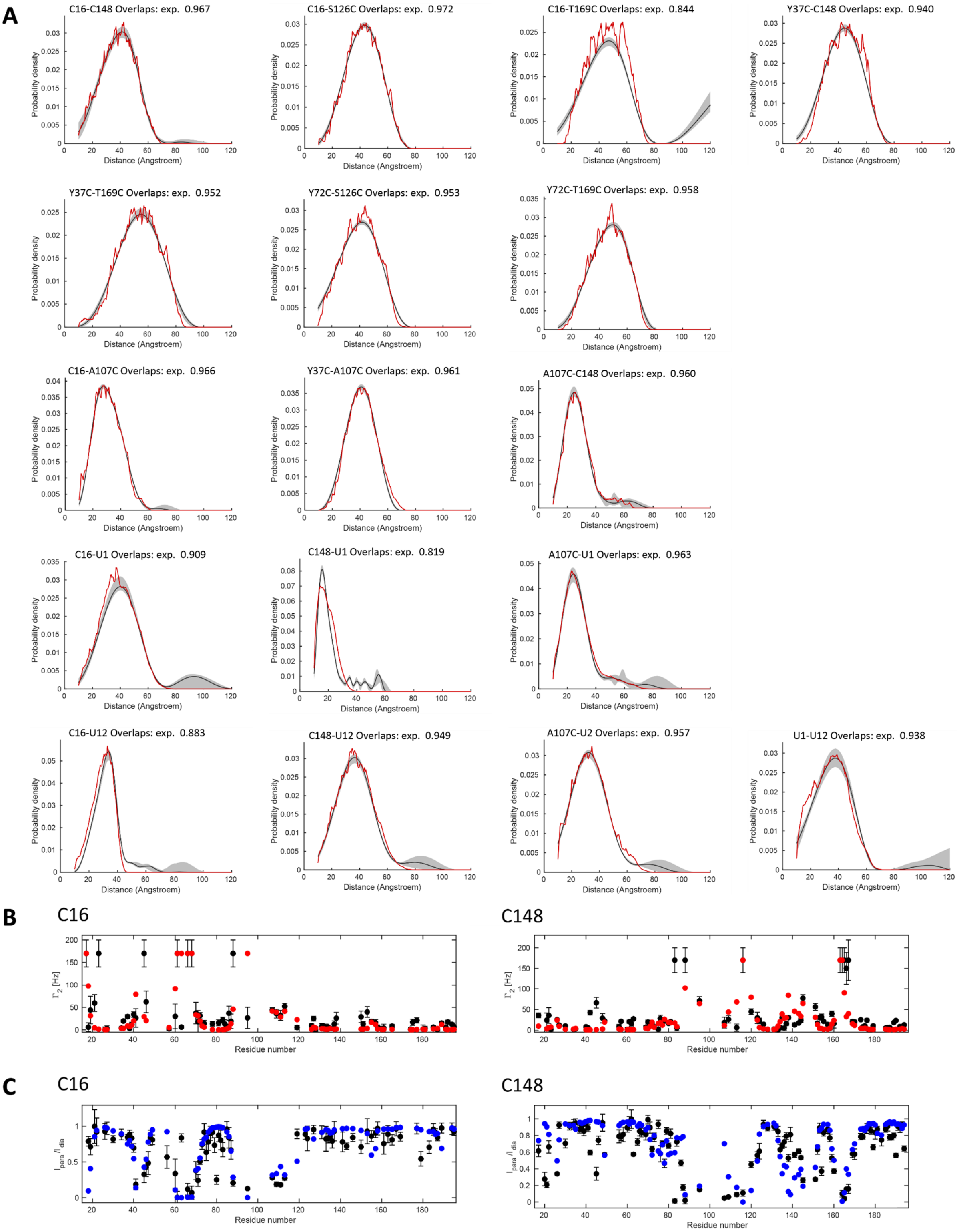


Figure A7 Restraint fit of the ensemble SRSF1 RRM1+2 in complex with 5'-UGGAUUUUUCAU-3' RNA calculated using the RigiFlex modeling. *(A) DEER distance distribution (B) PRE rate restraints (C) PRE ratio restraints. (A-C) Experimentally determined distance distribution (black) and predicted for the entire ensemble (red or blue). Final ensemble = 116 conformers.* $1-\bar{o} = 0.066$, χ^2 *(PRE rate) = 6.94,* χ^2 *(PRE ratio) = 18.43*

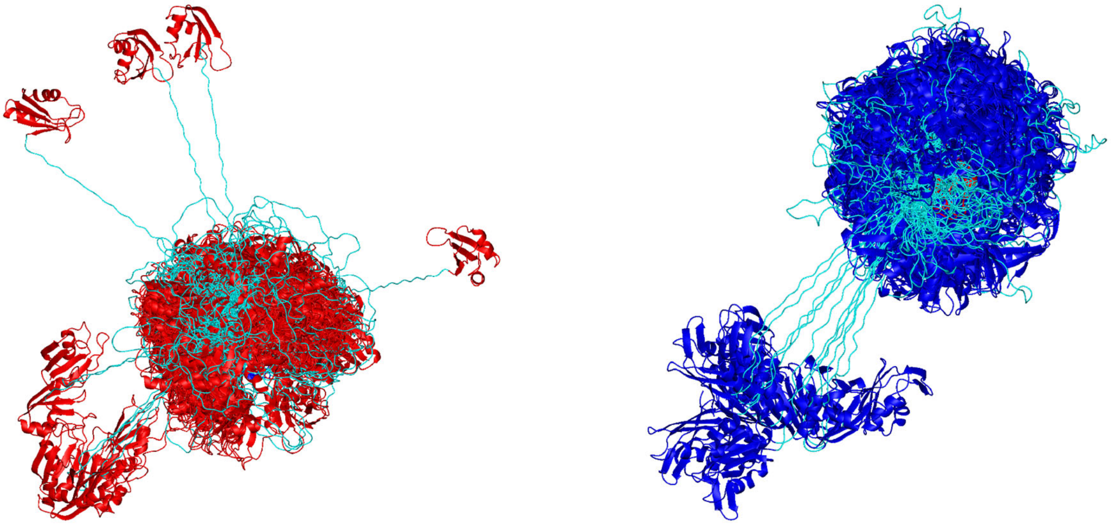


Figure A8 Ensembles of SRSF1 RRM1+2 in the free form calculated using Multistate CYANA calculation. *Raw ensemble obtained after the CYANA Multistate calculation using 213 PRE restraints (9 states, 180 conformers) Left: superimposed on RRM2, right superimposed on RRM1. RRM2 is depicted in blue, with RRM1 in red and the linker in cyan; N-terminal (1-15) is not shown.*

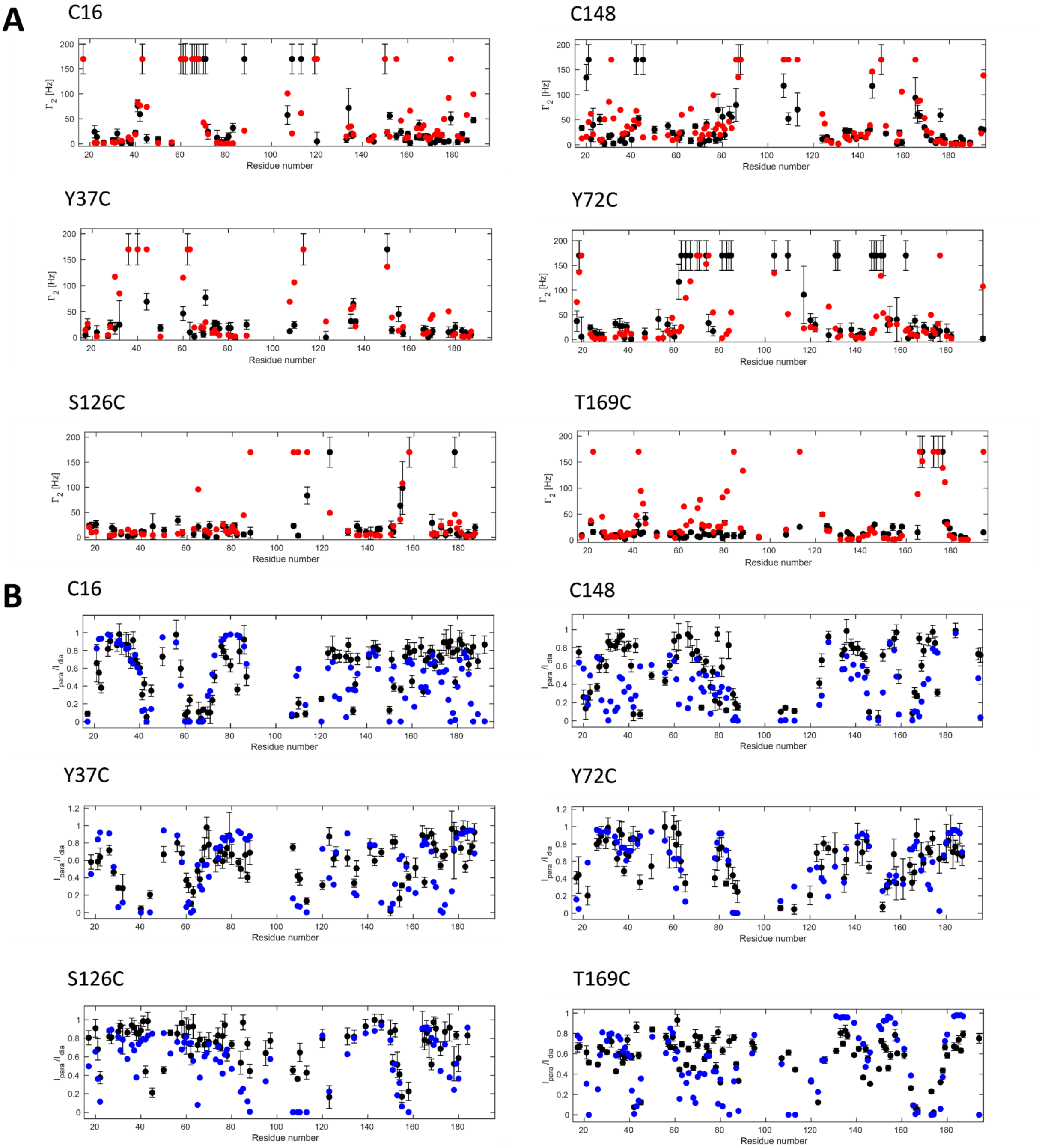


Figure A9 PRE restraint fit of the ensemble SRSF1 RRM1+2 in the free state after Multistate CYANA calculation. *(A) PRE rate restraints (B) PRE ratio restraints. (A-B) Experimentally determined distance distribution (black) and predicted for the entire ensemble (red or blue). Row ensemble = 140 conformers.* χ^2 *(PRE rate) = 19.5273,* χ^2 *(PRE ratio) = 9.4792*

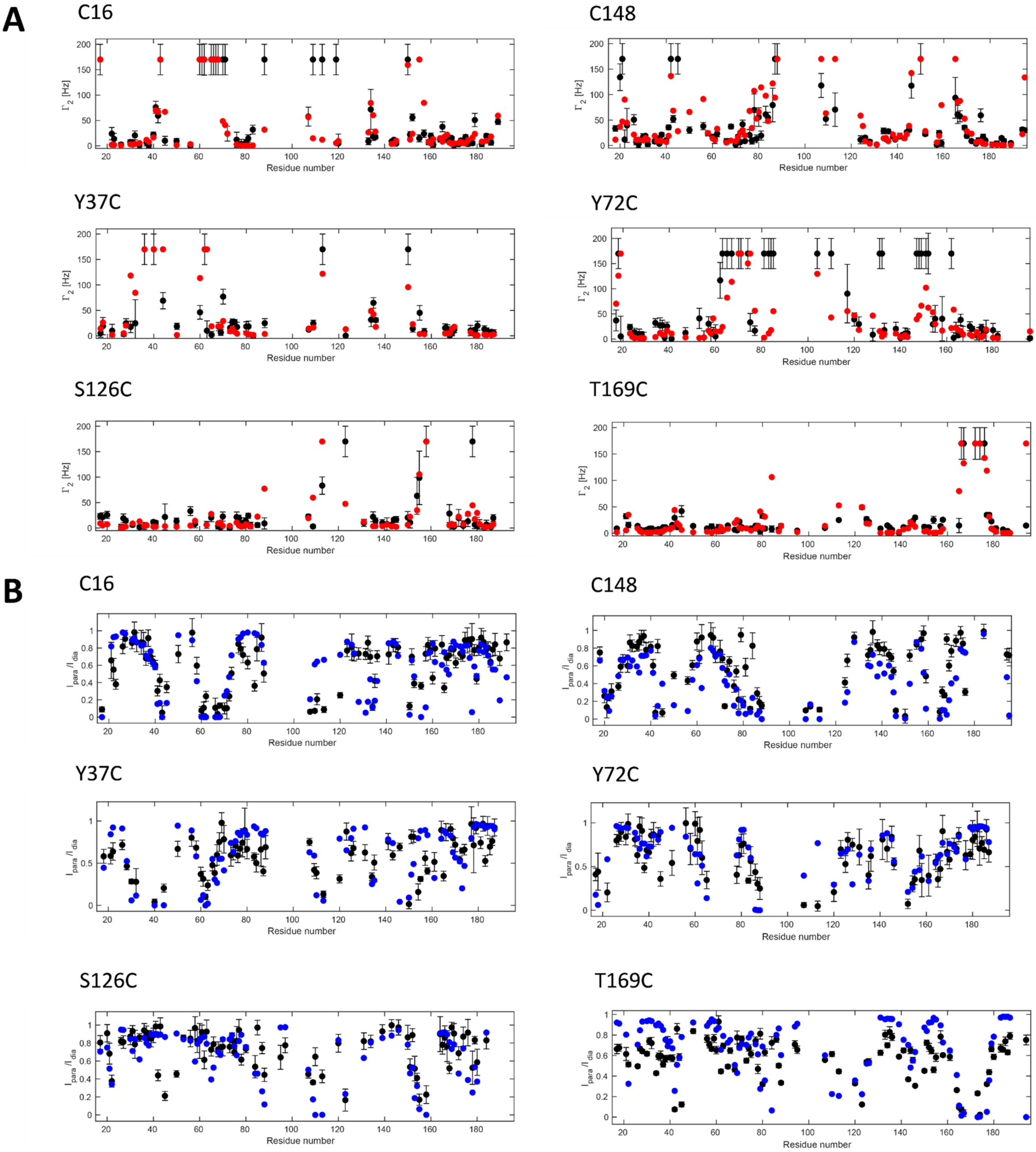
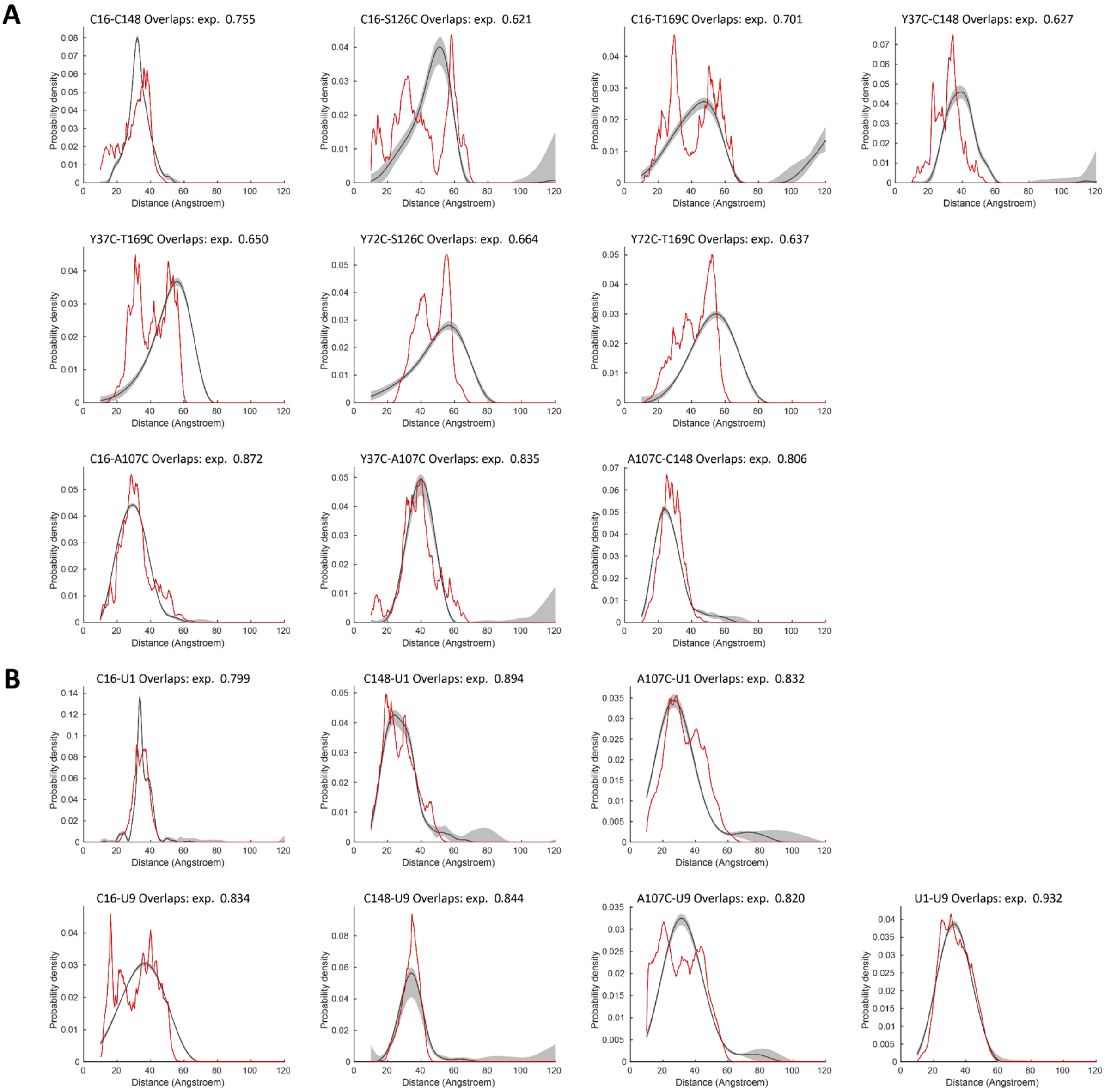


Figure A10 PRE restraint fit of the ensemble SRSF1 RRM1+2 in the free state after the EnsembleFit step. *(A) PRE rate restraints (B) PRE ratio restraints. (A-B) Experimentally determined distance distribution (black) and predicted for the entire ensemble (red or blue). Final ensemble = 49 conformers.* χ^2 *(PRE rate) = 6.6055,* χ^2 *(PRE ratio) = 8.8224*


A
C16-C148 Overlaps: exp. 0.755
C16-S126C Overlaps: exp. 0.621
C16-T169C Overlaps: exp. 0.701
Y37C-C148 Overlaps: exp. 0.627
Y37C-T169C Overlaps: exp. 0.650
Y72C-S126C Overlaps: exp. 0.664
Y72C-T169C Overlaps: exp. 0.637
C16-A107C Overlaps: exp. 0.872
Y37C-A107C Overlaps: exp. 0.835
A107C-C148 Overlaps: exp. 0.806
B
C16-U1 Overlaps: exp. 0.799
C148-U1 Overlaps: exp. 0.894
A107C-U1 Overlaps: exp. 0.832
C16-U9 Overlaps: exp. 0.834
C148-U9 Overlaps: exp. 0.844
A107C-U9 Overlaps: exp. 0.820
U1-U9 Overlaps: exp. 0.932
Probability density
Distance (Angstroem)
0
20
40
60
80
100
120

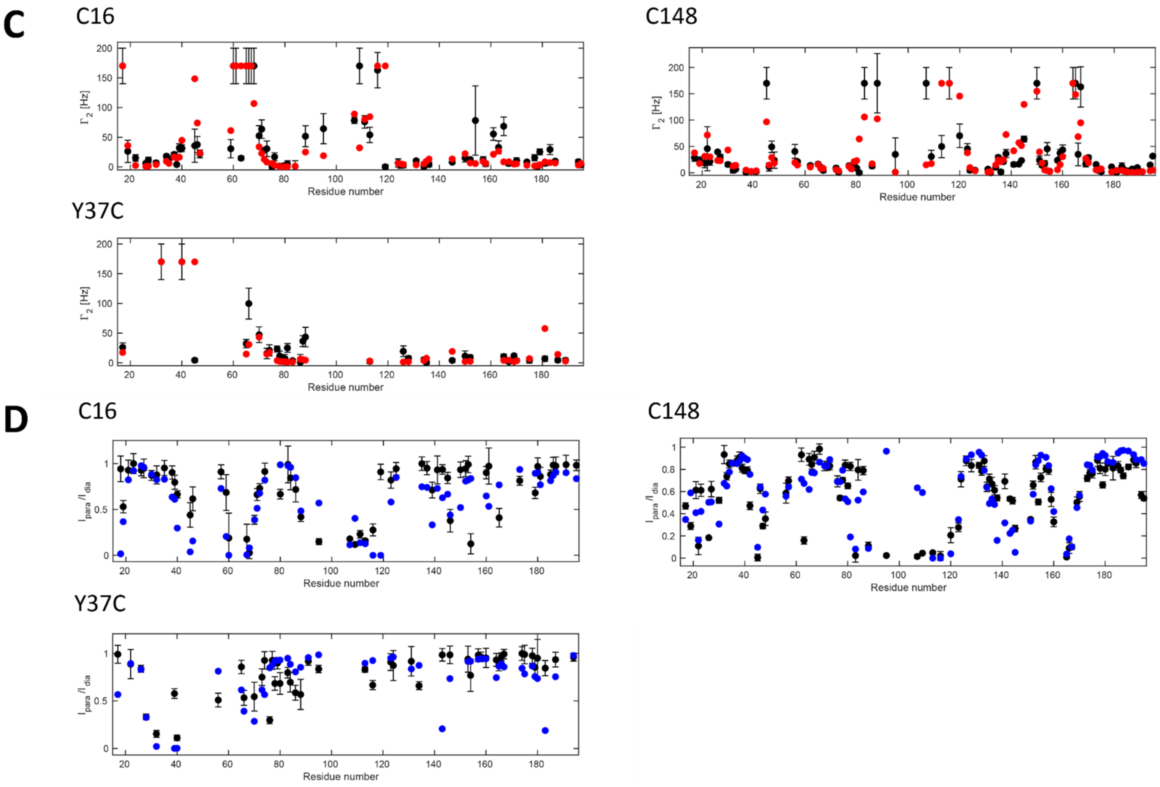
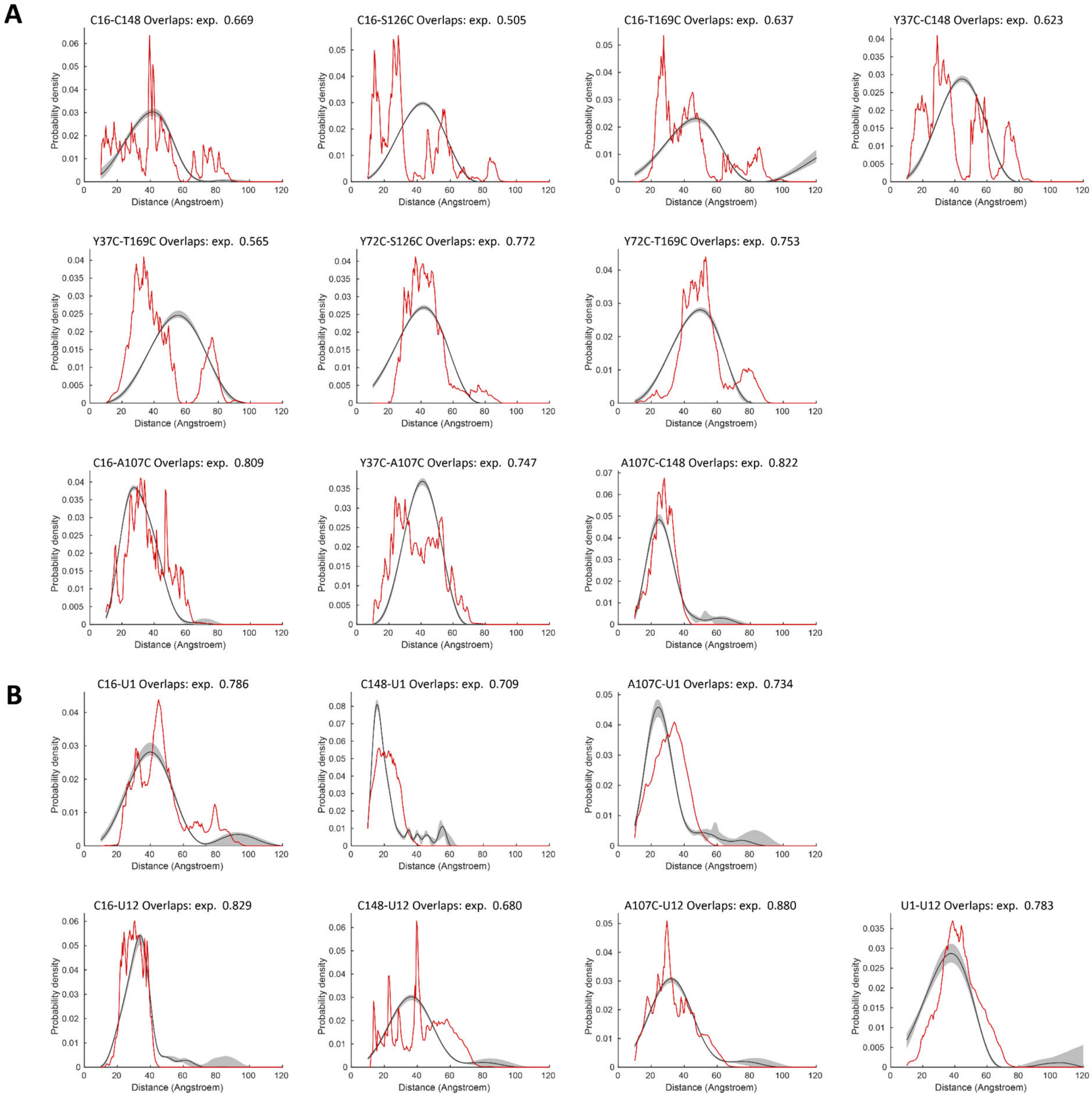


Figure A11 Restraint fit of the ensemble SRSF1 RRM1+2 in complex with 5'-UCAUUGGAU-3' RNA after the EnsembleFit step. *(A) Inter-RRM DEER distance distributions, (B) Intra-RNA and protein-RNA DEER distance distributions (C) PRE rate restraints (D) PRE ratio restraints. (A-D) Experimentally determined distance distribution (black) and predicted for the entire ensemble (red or blue). Final ensemble = 26 conformers. 1−ō = 0.235,* χ^2 *(PRE rate) = 12.2963,* χ^2 *(PRE ratio) = 21.6649*


A
C16-C148 Overlaps: exp. 0.669
C16-S126C Overlaps: exp. 0.505
C16-T169C Overlaps: exp. 0.637
Y37C-C148 Overlaps: exp. 0.623
Y37C-T169C Overlaps: exp. 0.565
Y72C-S126C Overlaps: exp. 0.772
Y72C-T169C Overlaps: exp. 0.753
C16-A107C Overlaps: exp. 0.809
Y37C-A107C Overlaps: exp. 0.747
A107C-C148 Overlaps: exp. 0.822
B
C16-U1 Overlaps: exp. 0.786
C148-U1 Overlaps: exp. 0.709
A107C-U1 Overlaps: exp. 0.734
C16-U12 Overlaps: exp. 0.829
C148-U12 Overlaps: exp. 0.680
A107C-U12 Overlaps: exp. 0.880
U1-U12 Overlaps: exp. 0.783
Probability density
Distance (Angstroem)
0
20
40
60
80
100
120

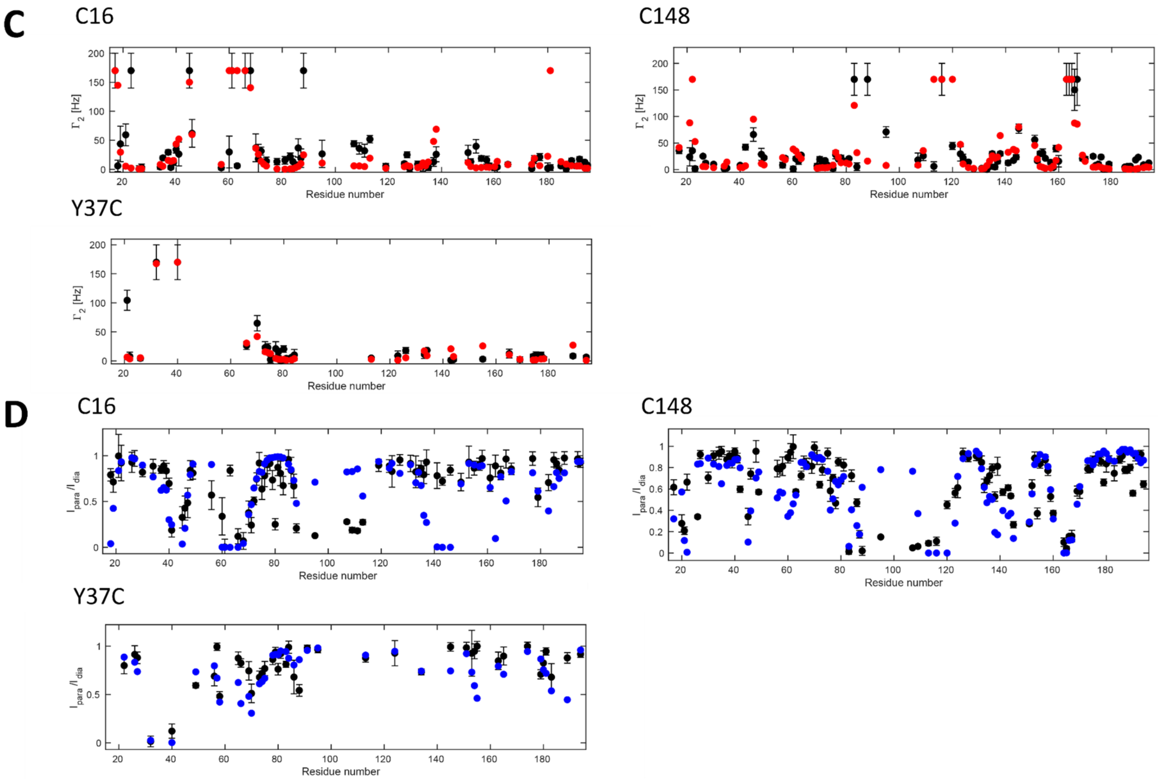


Figure A12 Restraint fit of the ensemble SRSF1 RRM1+2 in complex with 5'-UGGAUUUUUCAU-3' RNA after the EnsembleFit step. *(A) Inter-RRM DEER distance distributions, (B) Intra-RNA and protein-RNA DEER distance distributions (C) PRE rate restraints (D) PRE ratio restraints. (A-D) Experimentally determined distance distribution (black) and predicted for the entire ensemble (red or blue). Final ensemble = 23 conformers.* $1-\bar{o} = 0.283$, χ^2 *(PRE rate) = 8.4931,* χ^2 *(PRE ratio) = 25.8809*

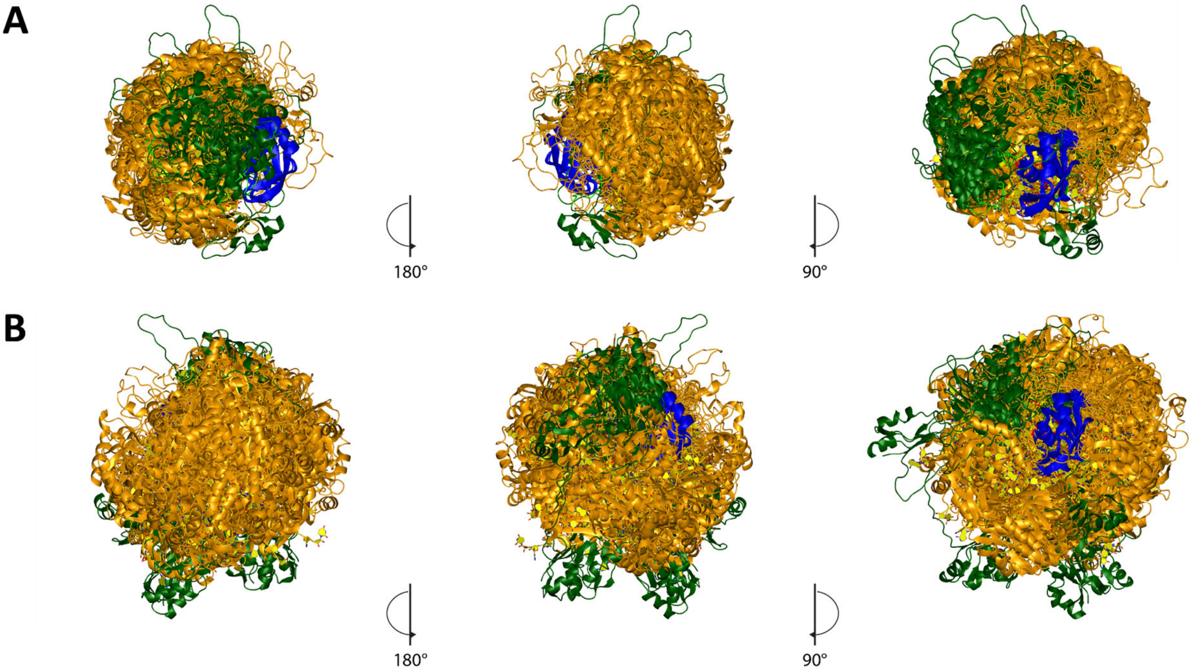


Figure A13 Comparison between ensembles of SRSF1 RRM1+2 in complex with RNA obtained with RigiFlex and the hybrid modeling. *(A) SRSF1 RRM1+2 in complex with 5'-UCAUUGGAU-3' and (B) SRSF1 RRM1+2 in complex with 5'-UGGAUUUUUCAU-3'. Ensembles are superimposed on RRM2 depicted in blue, RRM1 and linker are orange for the ensemble calculates using the RigiFlex modeling and in green for the ensemble obtained using the hybrid method; N-terminal (1-15) is not shown.*

A.3 Supporting Tables

Inter-RRM distances	RRM-linker distances	Protein-RNA distances
C16-C148	C16-A107C	C16-U5´
C16-S126C	Y37C-A107C	C16-U3´
C16-T169C	A107C-C148	C148-U5´
Y37C-C148		C148-U3´
Y37C-T169		A107C-U5´
Y72-S126C		A107C-U3´
Y72C-T196C		U5’-U3´

Table A1 Sites involved in the acquisition of inter-RRMs or linker-RRMs distance restraints

PRE mutants
C16A
C148A
C16A-C148A-Y37C
C16A-C148A-Y72C
C16A-C148A-S126C
C16A-C148A-T169C

Table A2 Mutations performed for the acquisition of PRE restraints

# states	TF	# viol	max
1	5517.85	78	14.92
2	755.93	51	9.16
3	124.68	23	3.26
4	53.36	23	3.12
5	20.78	20	1.82
6	3.90	8	0.8
7	1.43	3	0.41
8	0.80	1	0.3
9	0.77	1	0.29
10	0.82	1	0.33

Table A3 Test Multistate CYANA calculation for the free protein. *Calculations were performed for 1-10 states, calculating 2000 structures, 200000 steps, and the best 20 structures were selected (6 mutants = 213 PRE restraints were included). TF = target function; #viol = number of restraints that are violated, max = the maximal violation.*

# states	TF	# viol	max
1	1.98	4	0.37
2	4.76	8	0.36
3	8.3	13	0.41
4	14.4	19	0.57
5	21.03	27	0.68
6	28.63	33	0.95

Table A4 Test Multistate CYANA calculation for SRSF1 RRM1+2 in complex with 5'-UCAUUGGAU-3' RNA (no PRE data). *Calculations were performed for 1-6 states, calculating 2000 structures, 200000 steps, and the best 20 structures were selected (no PRE restraints were included). TF = target function; #viol = number of restraints that are violated, max = the maximal violation.*

# states	TF	# viol	max
1	1.98	4	0.32
2	4.3	8	0.33
3	8.14	12	0.41
4	15.02	18	0.71
5	24.58	26	1.11
6	34.44	31	1.16

Table A5 Test Multistate CYANA calculation for SRSF1 RRM1+2 in complex with 5'-UGGAUUUUUCAU-3' RNA (no PRE data). *Calculations were performed for 1-6 states, calculating 2000 structures, 200000 steps, and the best 20 structures were selected (no PRE restraints were included). TF = target function; #viol = number of restraints that are violated, max = the maximal violation.*

Appendix Chapter 3

B.1 Supporting figures

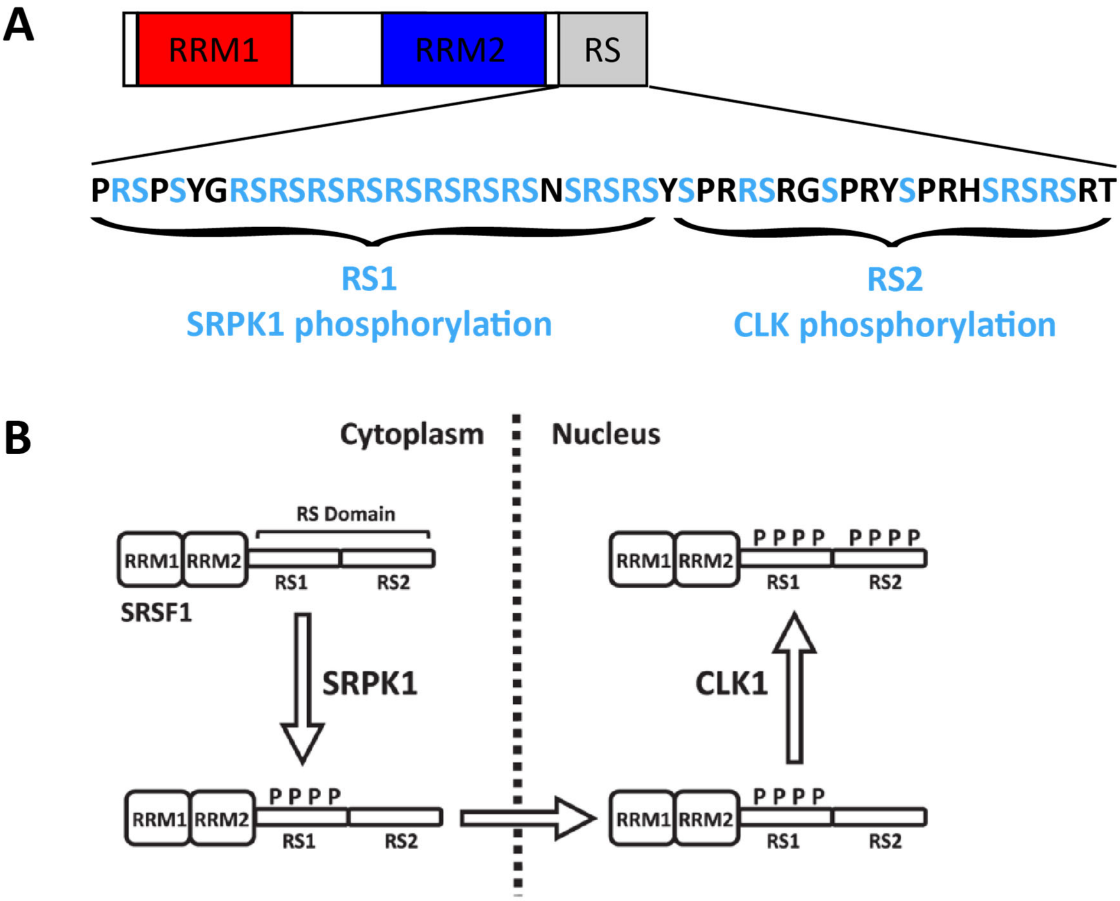


__Figure B1 Phosphorylation of the RS domain.__ (A) RS1 is phosphorylated by the cytoplasmic SRPK1 while RS2 is phosphorylated by the nuclear CLK. Modified from (Ghosh and Adams 2011). (B) Different phosphorylation of the two segments controls SRSF1 subcellular distribution (Aubol et al. 2013).

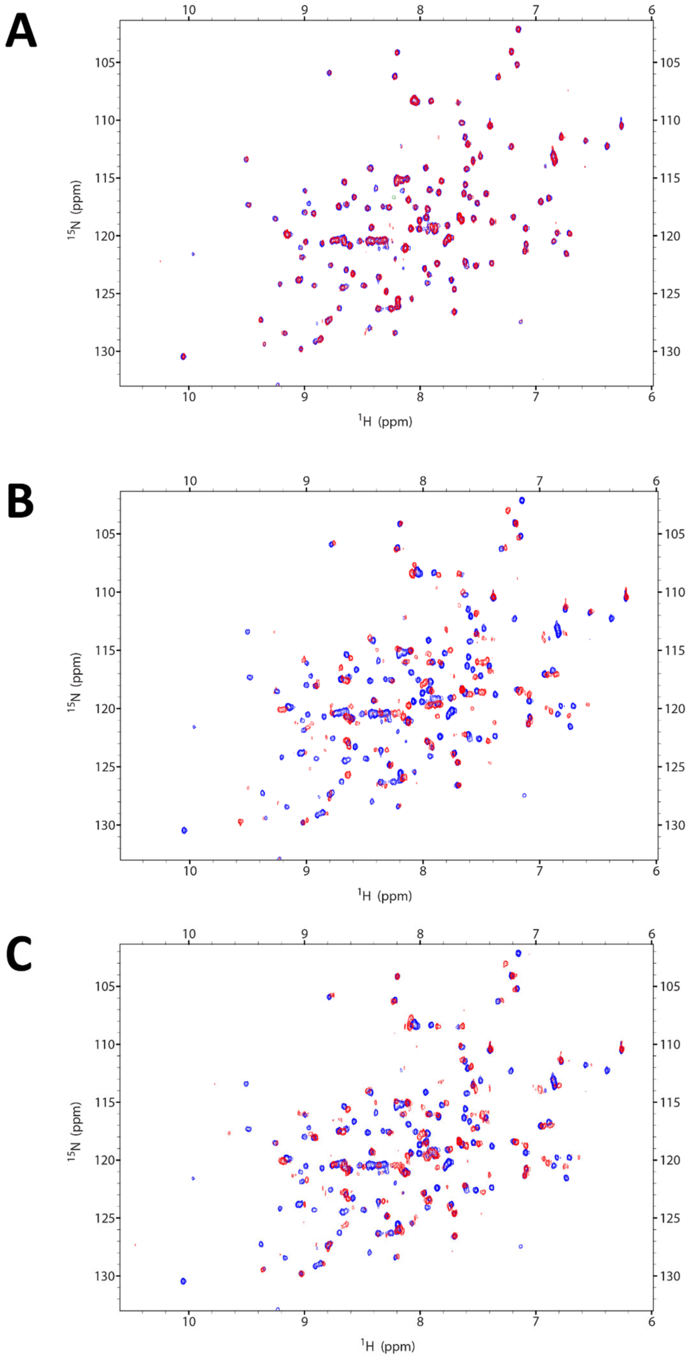


Figure B2 SRSF1 RRM1+2 without GB1 in presence of RNA*. Comparison of ^{1}H-^{15}N HSQC spectra of SRSF1 RRM1+2 YS free form (blue) and in the presence of (A) 5'-UUUUUUUUU-3', (B) 5'-UCAUUGGAU-3' and (C) 5'-UGGAUUUUUCAU-3' RNAs (red) at 1:1 RNA:protein ratios.*

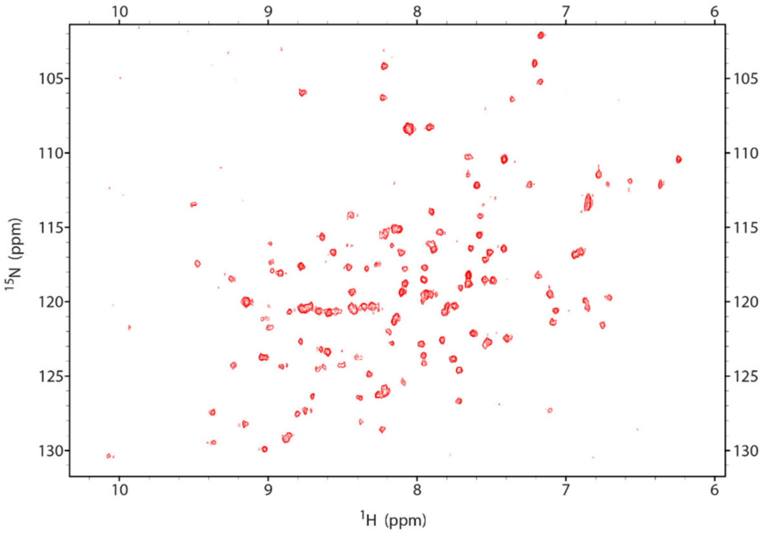


Figure B3 ^{1}H-^{15}N HSQC of SRSF1 RRM1+2 without GB1 stabilized in 0.5% agarose.

Chapter 6: References

The author used Grammarly and ChatGPT to check for grammatical errors and to make slight improvements of the writing and to minimize language errors.

Afroz, T., Z. Cienikova, A. Clery, and F. H. T. Allain. 2015. 'One, Two, Three, Four! How Multiple RRMs Read the Genome Sequence', *Methods Enzymol*, 558: 235-78.

Altenbach, C., T. Marti, H. G. Khorana, and W. L. Hubbell. 1990. 'Transmembrane protein structure: spin labeling of bacteriorhodopsin mutants', *Science*, 248: 1088-92.

Amrein, H., M. Gorman, and R. Nothiger. 1988. 'The sex-determining gene tra-2 of Drosophila encodes a putative RNA binding protein', *Cell*, 55: 1025-35.

Anko, M. L. 2014. 'Regulation of gene expression programmes by serine-arginine rich splicing factors', *Semin Cell Dev Biol*, 32: 11-21.

Ashkinadze, D., H. Kadavath, R. Riek, and P. Guntert. 2022. 'Optimization and validation of multi-state NMR protein structures using structural correlations', *Journal of Biomolecular Nmr*, 76: 39-47.

Aubol, B. E., R. M. Plocinik, J. C. Hagopian, C. T. Ma, M. L. McGlone, R. Bandyopadhyay, X. D. Fu, and J. A. Adams. 2013. 'Partitioning RS domain phosphorylation in an SR protein through the CLK and SRPK protein kinases', *J Mol Biol*, 425: 2894-909.

Aubol, B. E., G. Wu, M. M. Keshwani, M. Movassat, L. Fattet, K. J. Hertel, X. D. Fu, and J. A. Adams. 2016. 'Release of SR Proteins from CLK1 by SRPK1: A Symbiotic Kinase System for Phosphorylation Control of Pre-mRNA Splicing', *Mol Cell*, 63: 218-28.

Babiceanu, M., F. Qin, Z. Xie, Y. Jia, K. Lopez, N. Janus, L. Facemire, S. Kumar, Y. Pang, Y. Qi, I. M. Lazar, and H. Li. 2016. 'Recurrent chimeric fusion RNAs in non-cancer tissues and cells', *Nucleic Acids Res*, 44: 2859-72.

Banani, S. F., H. O. Lee, A. A. Hyman, and M. K. Rosen. 2017. 'Biomolecular condensates: organizers of cellular biochemistry', *Nat Rev Mol Cell Biol*, 18: 285-98.

Baralle, F. E., and J. Giudice. 2017. 'Alternative splicing as a regulator of development and tissue identity', *Nat Rev Mol Cell Biol*, 18: 437-51.

Barraud, P., and F. H. Allain. 2013. 'Solution structure of the two RNA recognition motifs of hnRNP A1 using segmental isotope labeling: how the relative orientation between RRMs influences the nucleic acid binding topology', *J Biomol NMR*, 55: 119-38.

Berget, S. M., C. Moore, and P. A. Sharp. 1977. 'Spliced segments at the 5' terminus of adenovirus 2 late mRNA', *Proc Natl Acad Sci U S A*, 74: 3171-5.

Bessonov, S., M. Anokhina, C. L. Will, H. Urlaub, and R. Luhrmann. 2008. 'Isolation of an active step I spliceosome and composition of its RNP core', *Nature*, 452: 846-50.

Biggiogera, M., and S. Fakan. 1998. 'Fine structural specific visualization of RNA on ultrathin sections', *J Histochem Cytochem*, 46: 389-95.

Boehr, D. D., D. McElheny, H. J. Dyson, and P. E. Wright. 2006. 'The dynamic energy landscape of dihydrofolate reductase catalysis', *Science*, 313: 1638-42.

Boggs, R. T., P. Gregor, S. Idriss, J. M. Belote, and M. McKeown. 1987. 'Regulation of sexual differentiation in D. melanogaster via alternative splicing of RNA from the transformer gene', *Cell*, 50: 739-47.

Braberg, H., H. Jin, E. A. Moehle, Y. A. Chan, S. Wang, M. Shales, J. J. Benschop, J. H. Morris, C. Qiu, F. Hu, L. K. Tang, J. S. Fraser, F. C. Holstege, P. Hieter, C. Guthrie, C. D. Kaplan, and N. J. Krogan. 2013. 'From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II', *Cell*, 154: 775-88.

Brangwynne, C. P., P. Tompa, and R. V. Pappu. 2015. 'Polymer physics of intracellular phase transitions', *Nature Physics*, 11: 899-904.

Bryngelson, J. D., J. N. Onuchic, N. D. Socci, and P. G. Wolynes. 1995. 'Funnels, pathways, and the energy landscape of protein folding: a synthesis', *Proteins*, 21: 167-95.

Bryngelson, J. D., and P. G. Wolynes. 1987. 'Spin glasses and the statistical mechanics of protein folding', *Proc Natl Acad Sci U S A*, 84: 7524-8.

Caceres, J. F., T. Misteli, G. R. Screaton, D. L. Spector, and A. R. Krainer. 1997. 'Role of the modular domains of SR proteins in subnuclear localization and alternative splicing specificity', *J Cell Biol*, 138: 225-38.

Cartegni, L., and A. R. Krainer. 2002. 'Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1', *Nat Genet*, 30: 377-84.

Castello, A., B. Fischer, C. K. Frese, R. Horos, A. M. Alleaume, S. Foehr, T. Curk, J. Krijgsveld, and M. W. Hentze. 2016. 'Comprehensive Identification of RNA-Binding Domains in Human Cells', *Mol Cell*, 63: 696-710.

Cavanagh, J., W. J. Fairbrother, A. G. Palmer, M. Rance, and N. J. Skelton. 2007. 'Protein Nmr Spectroscopy Principles and Practice Second Edition Preface to the First Edition', *Protein Nmr Spectroscopy: Principles and Practice, 2nd Edition*: Vii-X.

Cech, T. R., and J. A. Steitz. 2014. 'The noncoding RNA revolution-trashing old rules to forge new ones', *Cell*, 157: 77-94.

Cerasuolo, A., L. Buonaguro, F. M. Buonaguro, and M. L. Tornesello. 2020. 'The Role of RNA Splicing Factors in Cancer: Regulation of Viral and Human Gene Expression in Human Papillomavirus-Related Cervical Cancer', *Front Cell Dev Biol*, 8: 474.

Che, Y., and L. Fu. 2020. 'Aberrant expression and regulatory network of splicing factor-SRSF3 in tumors', *J Cancer*, 11: 3502-11.

Chen, M., and J. L. Manley. 2009. 'Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches', *Nat Rev Mol Cell Biol*, 10: 741-54.

Cho, S., A. Hoang, S. Chakrabarti, N. Huynh, D. B. Huang, and G. Ghosh. 2011. 'The SRSF1 linker induces semi-conservative ESE binding by cooperating with the RRMs', *Nucleic Acids Res*, 39: 9413-21.

Cho, S., A. Hoang, R. Sinha, X. Y. Zhong, X. D. Fu, A. R. Krainer, and G. Ghosh. 2011. 'Interaction between the RNA binding domains of Ser-Arg splicing factor 1 and U1-70K snRNP protein determines early spliceosome assembly', *Proc Natl Acad Sci U S A*, 108: 8233-8.

Chou, T. B., Z. Zachar, and P. M. Bingham. 1987. 'Developmental expression of a regulatory gene is programmed at the level of splicing', *EMBO J*, 6: 4095-104.

Chow, L. T., R. E. Gelinas, T. R. Broker, and R. J. Roberts. 1977. 'An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA', *Cell*, 12: 1-8.

Clery, A., M. Krepl, C. K. X. Nguyen, A. Moursy, H. Jorjani, M. Katsantoni, M. Okoniewski, N. Mittal, M. Zavolan, J. Sponer, and F. H. Allain. 2021. 'Structure of SRSF1 RRM1 bound to RNA reveals an unexpected bimodal mode of interaction and explains its involvement in SMN1 exon7 splicing', *Nat Commun*, 12: 428.

Clery, A., M. Schubert, and F. H. Allain. 2012. 'NMR spectroscopy: an excellent tool to understand RNA and carbohydrate recognition by proteins', *Chimia (Aarau)*, 66: 741-6.

Clery, A., R. Sinha, O. Anczukow, A. Corrionero, A. Moursy, G. M. Daubner, J. Valcarcel, A. R. Krainer, and F. H. Allain. 2013. 'Isolated pseudo-RNA-recognition motifs of SR proteins can regulate splicing using a noncanonical mode of RNA recognition', *Proc Natl Acad Sci U S A*, 110: E2802-11.

Clore, G. M., and A. M. Gronenborn. 1989. 'Determination of three-dimensional structures of proteins and nucleic acids in solution by nuclear magnetic resonance spectroscopy', *Crit Rev Biochem Mol Biol*, 24: 479-564.

———. 1998a. 'Determining the structures of large proteins and protein complexes by NMR', *Trends Biotechnol*, 16: 22-34.

———. 1998b. 'New methods of structure refinement for macromolecular structure determination by NMR', *Proc Natl Acad Sci U S A*, 95: 5891-8.

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: 4108-39.

Clore, G. M., and V. Venditti. 2013. 'Structure, dynamics and biophysics of the cytoplasmic protein-protein complexes of the bacterial phosphoenolpyruvate: sugar phosphotransferase system', *Trends Biochem Sci*, 38: 515-30.

Colwill, K., L. L. Feng, J. M. Yeakley, G. D. Gish, J. F. Caceres, T. Pawson, and X. D. Fu. 1996. 'SRPK1 and Clk/Sty protein kinases show distinct substrate specificities for serine/arginine-rich splicing factors', *J Biol Chem*, 271: 24569-75.

Colwill, K., T. Pawson, B. Andrews, J. Prasad, J. L. Manley, J. C. Bell, and P. I. Duncan. 1996. 'The Clk/Sty protein kinase phosphorylates SR splicing factors and regulates their intranuclear distribution', *EMBO J*, 15: 265-75.

Corkery, D. P., A. C. Holly, S. Lahsaee, and G. Dellaire. 2015. 'Connecting the speckles: Splicing kinases and their role in tumorigenesis and treatment response', *Nucleus*, 6: 279-88.

Cowper, A. E., J. F. Caceres, A. Mayeda, and G. R. Screaton. 2001. 'Serine-arginine (SR) protein-like factors that antagonize authentic SR proteins and regulate alternative splicing', *J Biol Chem*, 276: 48908-14.

Crawford, T. O., and C. A. Pardo. 1996. 'The neurobiology of childhood spinal muscular atrophy', *Neurobiol Dis*, 3: 97-110.

Dasgupta, T., and A. N. Ladd. 2012. 'The importance of CELF control: molecular and biological roles of the CUG-BP, Elav-like family of RNA-binding proteins', *Wiley Interdiscip Rev RNA*, 3: 104-21.

De Silva, N. I. U., T. Fargason, Z. Zhang, T. Wang, and J. Zhang. 2022. 'Inter-domain Flexibility of Human Ser/Arg-Rich Splicing Factor 1 Allows Variable Spacer Length in Cognate RNA's Bipartite Motifs', *Biochemistry*, 61: 2922-32.

Deka, P., M. E. Bucheli, C. Moore, S. Buratowski, and G. Varani. 2008. 'Structure of the yeast SR protein Npl3 and Interaction with mRNA 3'-end processing signals', *J Mol Biol*, 375: 136-50.

Dill, K. A. 1985. 'Theory for the folding and stability of globular proteins', *Biochemistry*, 24: 1501-9.

Dill, K. A., and H. S. Chan. 1997. 'From Levinthal to pathways to funnels', *Nat Struct Biol*, 4: 10-9.

Dorn, G., G. Gmeiner, T. de Vries, E. Dedic, M Novakovic, F. F. Damberger, C. Maris, E Finol, C. P. Sarnowski, J. Kohlbrecher, T. J. Welsh, S. Bolisetty, R. Mezzenga, R. Aebersold, A. Leitner, M. Yulikov, G. Jeschke, and F. H.-T. Allain. 2022. "Integrative solution structure of a PTBP1-viral IRES complex reveals strong compaction and ordering with residual conformational flexibility." In.: bioRxiv.

Douglas, A. G., and M. J. Wood. 2011. 'RNA splicing: disease and therapy', *Brief Funct Genomics*, 10: 151-64.

Dowling, D., S. Nasr-Esfahani, C. H. Tan, K. O'Brien, J. L. Howard, D. A. Jans, D. F. Purcell, C. M. Stoltzfus, and S. Sonza. 2008. 'HIV-1 infection induces changes in expression of cellular splicing factors that regulate alternative viral splicing and virus production in macrophages', *Retrovirology*, 5: 18.

Draper, D. E. 1995. 'Protein-Rna Recognition', *Annual Review of Biochemistry*, 64: 593-620.

Dreyfuss, G., V. N. Kim, and N. Kataoka. 2002. 'Messenger-RNA-binding proteins and the messages they carry', *Nat Rev Mol Cell Biol*, 3: 195-205.

Dreyfuss, G., M. J. Matunis, S. Pinol-Roma, and C. G. Burd. 1993. 'hnRNP proteins and the biogenesis of mRNA', *Annu Rev Biochem*, 62: 289-321.

Duss, O., E. Michel, M. Yulikov, M. Schubert, G. Jeschke, and F. H. Allain. 2014. 'Structural basis of the non-coding RNA RsmZ acting as a protein sponge', *Nature*, 509: 588-92.

Eddy, S. R. 2001. 'Non-coding RNA genes and the modern RNA world', *Nat Rev Genet*, 2: 919-29.

Eigen, M. 1957. 'Determination of General and Specific Ionic Interactions in Solution', *Discussions of the Faraday Society*: 25-36.

———. 1968. 'New looks and outlooks on physical enzymology', *Q Rev Biophys*, 1: 3-33.

Emmanouilidis, L., L. Esteban-Hofer, F. F. Damberger, T. de Vries, C. K. X. Nguyen, L. F. Ibanez, S. Mergenthal, E. Klotzsch, M. Yulikov, G. Jeschke, and F. H. Allain. 2021. 'NMR and EPR reveal a compaction of the RNA-binding protein FUS upon droplet formation', *Nat Chem Biol*, 17: 608-14.

Esteban-Hofer, L. 2022. 'EPR Characterization of Intrinsic Disorder in the RNA-Binding Proteins FUS and SRSF1', ETH Zürich.

Fabregas Ibanez, L., G. Jeschke, and S. Stoll. 2020. 'DeerLab: a comprehensive software package for analyzing dipolar electron paramagnetic resonance spectroscopy data', *Magn Reson (Gott)*, 1: 209-24.

Fischer, E. 1894. 'The influence of configuration on enzyme activity (Translated from German)', *Dtsch Chem Ges.*: 1894;27:2984–93.

Frankiw, L., D. Baltimore, and G. Li. 2019. 'Alternative mRNA splicing in cancer immunotherapy', *Nat Rev Immunol*, 19: 675-87.

Fu, X. D. 1995. 'The superfamily of arginine/serine-rich splicing factors', *RNA*, 1: 663-80.

Fu, X. D., and T. Maniatis. 1992. 'The 35-kDa mammalian splicing factor SC35 mediates specific interactions between U1 and U2 small nuclear ribonucleoprotein particles at the 3' splice site', *Proc Natl Acad Sci U S A*, 89: 1725-9.

Gallinaro, H., E. Lazar, M. Jacob, A. Krol, and C. Branlant. 1981. 'Small RNAs in HnRNP fibrils and their possible function in splicing', *Mol Biol Rep*, 7: 31-9.

Garbuio, L., E. Bordignon, E. K. Brooks, W. L. Hubbell, G. Jeschke, and M. Yulikov. 2013. 'Orthogonal spin labeling and Gd(III)-nitroxide distance measurements on bacteriophage T4-lysozyme', *J Phys Chem B*, 117: 3145-53.

Ge, H., and J. L. Manley. 1990. 'A protein factor, ASF, controls cell-specific alternative splicing of SV40 early pre-mRNA in vitro', *Cell*, 62: 25-34.

Gerstberger, S., M. Hafner, and T. Tuschl. 2014. 'A census of human RNA-binding proteins', *Nat Rev Genet*, 15: 829-45.

Geuens, T., D. Bouhy, and V. Timmerman. 2016. 'The hnRNP family: insights into their role in health and disease', *Hum Genet*, 135: 851-67.

Ghosh, G., and J. A. Adams. 2011. 'Phosphorylation mechanism and structure of serine-arginine protein kinases', *FEBS J*, 278: 587-97.

Glatz, D. C., D. Rujescu, Y. Tang, F. J. Berendt, A. M. Hartmann, F. Faltraco, C. Rosenberg, C. Hulette, K. Jellinger, H. Hampel, P. Riederer, H. J. Moller, A. Andreadis, K. Henkel, and S. Stamm. 2006. 'The alternative splicing of tau exon 10 and its regulatory proteins CLK2 and TRA2-BETA1 changes in sporadic Alzheimer's disease', *J Neurochem*, 96: 635-44.

Gossert, A. D., and W. Jahnke. 2016. 'NMR in drug discovery: A practical guide to identification and validation of ligands interacting with biological macromolecules', *Prog Nucl Magn Reson Spectrosc*, 97: 82-125.

Guerrier-Takada, C., K. Gardiner, T. Marsh, N. Pace, and S. Altman. 1983. 'The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme', *Cell*, 35: 849-57.

Gui, J. F., W. S. Lane, and X. D. Fu. 1994. 'A serine kinase regulates intracellular localization of splicing factors in the cell cycle', *Nature*, 369: 678-82.

He, F. , Y. Muto, M. Inoue, T. Kigawa, M. Shirouzu, T. Terada, S. Yokoyama, and RIKEN Structural Genomics/Proteomics Initiative (RSGI). 2005. "Solution structure of RRM domain in splicing factor SF2." In. Protein Data Bank

Hentze, M. W., A. Castello, T. Schwarzl, and T. Preiss. 2018. 'A brave new world of RNA-binding proteins', *Nat Rev Mol Cell Biol*, 19: 327-41.

Henzler-Wildman, K. A., V. Thai, M. Lei, M. Ott, M. Wolf-Watz, T. Fenn, E. Pozharski, M. A. Wilson, G. A. Petsko, M. Karplus, C. G. Hubner, and D. Kern. 2007. 'Intrinsic motions along an enzymatic reaction trajectory', *Nature*, 450: 838-44.

Herrmann, T., P. Guntert, and K. Wuthrich. 2002. 'Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS', *J Biomol NMR*, 24: 171-89.

Hirose, T., K. Ninomiya, S. Nakagawa, and T. Yamazaki. 2022. 'A guide to membraneless organelles and their various roles in gene regulation', *Nat Rev Mol Cell Biol*.

Hofmann, Y., and B. Wirth. 2002. 'hnRNP-G promotes exon 7 inclusion of survival motor neuron (SMN) via direct interaction with Htra2-beta1', *Hum Mol Genet*, 11: 2037-49.

Hooks, K. B., D. Delneri, and S. Griffiths-Jones. 2014. 'Intron evolution in Saccharomycetaceae', *Genome Biol Evol*, 6: 2543-56.

House, A. E., and K. W. Lynch. 2006. 'An exonic splicing silencer represses spliceosome assembly after ATP-dependent exon recognition', *Nat Struct Mol Biol*, 13: 937-44.

Huang, Y., T. A. Yario, and J. A. Steitz. 2004. 'A molecular link between SR protein dephosphorylation and mRNA export', *Proc Natl Acad Sci U S A*, 101: 9666-70.

Humphrey, W., A. Dalke, and K. Schulten. 1996. 'VMD: visual molecular dynamics', *J Mol Graph*, 14: 33-8, 27-8.

Irimia, M., A. Denuc, D. Burguera, I. Somorjai, J. M. Martin-Duran, G. Genikhovich, S. Jimenez-Delgado, U. Technau, S. W. Roy, G. Marfany, and J. Garcia-Fernandez. 2011. 'Stepwise assembly of the Nova-regulated alternative splicing network in the vertebrate brain', *Proc Natl Acad Sci U S A*, 108: 5319-24.

Iwahara, J., D. E. Anderson, E. C. Murphy, and G. M. Clore. 2003. 'EDTA-derivatized deoxythymidine as a tool for rapid determination of protein binding polarity to DNA by intermolecular paramagnetic relaxation enhancement', *J Am Chem Soc*, 125: 6634-5.

Iwahara, J., C. D. Schwieters, and G. M. Clore. 2004. 'Ensemble approach for NMR structure refinement against (1)H paramagnetic relaxation enhancement data arising from a flexible paramagnetic group attached to a macromolecule', *J Am Chem Soc*, 126: 5879-96.

Iwahara, J., C. Tang, and G. Marius Clore. 2007. 'Practical aspects of (1)H transverse paramagnetic relaxation enhancement measurements on macromolecules', *J Magn Reson*, 184: 185-95.

Jankowsky, E., and M. E. Harris. 2015. 'Specificity and nonspecificity in RNA-protein interactions', *Nat Rev Mol Cell Biol*, 16: 533-44.

Jeong, S. 2017. 'SR Proteins: Binders, Regulators, and Connectors of RNA', *Mol Cells*, 40: 1-9.

Jeschke, G. 2013. "Interpretation of Dipolar EPR Data in Terms of Protein Structure." In *Structural Information from Spin-Labels and Intrinsic Paramagnetic Centres in the Biosciences*, 88-120. C. R.

Timmel, J. R. Harmer, Eds. (Springer Berlin Heidelberg, 2013).

Jeschke, G , and L. Esteban-Hofer. 2022. 'Integrative ensemble modeling of proteins and their complexes with distance distribution restraints.' in, *Advances in Biomolecular EPR (ed Britt, R. D.)* (Academic Press).

Jeschke, G. 2012. 'DEER distance measurements on proteins', *Annu Rev Phys Chem*, 63: 419-46.

———. 2016. 'Ensemble models of proteins and protein domains based on distance distribution restraints', *Proteins*, 84: 544-60.

———. 2018. 'MMM: A toolbox for integrative structure modeling', *Protein Sci*, 27: 76-85.

———. 2021. 'MMM: Integrative ensemble modeling and ensemble analysis', *Protein Sci*, 30: 125-35.

Joshi, A., V. Esteve, A. N. Buckroyd, M. Blatter, F. H. Allain, and S. Curry. 2014. 'Solution and crystal structures of a C-terminal fragment of the neuronal isoform of the polypyrimidine tract binding protein (nPTB)', *PeerJ*, 2: e305.

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. 2021. 'Highly accurate protein structure prediction with AlphaFold', *Nature*, 596: 583-89.

Jurica, M. S., and M. J. Moore. 2003. 'Pre-mRNA splicing: awash in a sea of proteins', *Mol Cell*, 12: 5-14.

Kampa, D., J. Cheng, P. Kapranov, M. Yamanaka, S. Brubaker, S. Cawley, J. Drenkow, A. Piccolboni, S. Bekiranov, G. Helt, H. Tammana, and T. R. Gingeras. 2004.

'Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22', *Genome Res*, 14: 331-42.

Kashima, T., and J. L. Manley. 2003. 'A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy', *Nat Genet*, 34: 460-3.

Kataoka, N., J. L. Bachorik, and G. Dreyfuss. 1999. 'Transportin-SR, a nuclear import receptor for SR proteins', *J Cell Biol*, 145: 1145-52.

Kay, L. E., D. A. Torchia, and A. Bax. 1989. 'Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease', *Biochemistry*, 28: 8972-9.

Keil, P., A. Wulf, N. Kachariya, S. Reuscher, K. Huhn, I. Silbern, J. Altmuller, M. Keller, R. Stehle, K. Zarnack, M. Sattler, H. Urlaub, and K. Strasser. 2023. 'Npl3 functions in mRNP assembly by recruitment of mRNP components to the transcription site and their transfer onto the mRNA', *Nucleic Acids Research*, 51: 831-51.

Keiten-Schmitz, J., L. Roder, E. Hornstein, M. Muller-McNicoll, and S. Muller. 2021. 'SUMO: Glue or Solvent for Phase-Separated Ribonucleoprotein Complexes and Molecular Condensates?', *Front Mol Biosci*, 8: 673038.

Keiten-Schmitz, J., K. Wagner, T. Piller, M. Kaulich, S. Alberti, and S. Muller. 2020. 'The Nuclear SUMO-Targeted Ubiquitin Quality Control Network Regulates the Dynamics of Cytoplasmic Stress Granules', *Mol Cell*, 79: 54-67 e7.

Koradi, R., M. Billeter, and K. Wuthrich. 1996. 'MOLMOL: a program for display and analysis of macromolecular structures', *J Mol Graph*, 14: 51-5, 29-32.

Kosen, P. A. 1989. 'Spin labeling of proteins', *Methods Enzymol*, 177: 86-121.

Krainer, A. R., G. C. Conway, and D. Kozak. 1990. 'Purification and characterization of pre-mRNA splicing factor SF2 from HeLa cells', *Genes Dev*, 4: 1158-71.

Kramer, A. 1996. 'The structure and function of proteins involved in mammalian pre-mRNA splicing', *Annu Rev Biochem*, 65: 367-409.

Krieger, E., K. Joo, J. Lee, J. Lee, S. Raman, J. Thompson, M. Tyka, D. Baker, and K. Karplus. 2009. 'Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8', *Proteins*, 77 Suppl 9: 114-22.

Krivov, G. G., M. V. Shapovalov, and R. L. Dunbrack, Jr. 2009. 'Improved prediction of protein side-chain conformations with SCWRL4', *Proteins*, 77: 778-95.

Kruger, K., P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech. 1982. 'Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena', *Cell*, 31: 147-57.

Kumar, D., M. Das, C. Sauceda, L. G. Ellies, K. Kuo, P. Parwal, M. Kaur, L. Jih, G. K. Bandyopadhyay, D. Burton, R. Loomba, O. Osborn, and N. J. Webster. 2019. 'Degradation of splicing factor SRSF3 contributes to progressive liver disease', *J Clin Invest*, 129: 4477-91.

Lai, M. C., R. I. Lin, S. Y. Huang, C. W. Tsai, and W. Y. Tarn. 2000. 'A human importin-beta family protein, transportin-SR2, interacts with the phosphorylated RS domain of SR proteins', *J Biol Chem*, 275: 7950-7.

Larrasa-Alonso, J., M. Villalba-Orero, C. Marti-Gomez, P. Ortiz-Sanchez, M. M. Lopez-Olaneta, M. A. Rey-Martin, F. Sanchez-Cabo, F. McNicoll, M. Muller-McNicoll, P. Garcia-Pavia, and E. Lara-Pezzi. 2021. 'The SRSF4-GAS5-Glucocorticoid Receptor Axis Regulates Ventricular Hypertrophy', *Circ Res*, 129: 669-83.

Lee, D., C. Hilty, G. Wider, and K. Wuthrich. 2006. 'Effective rotational correlation times of proteins from NMR relaxation interference', *J Magn Reson*, 178: 72-6.

Li, K., and Z. Wang. 2021. 'Splicing factor SRSF2-centric gene regulation', *Int J Biol Sci*, 17: 1708-15.

Li, P., S. Banjade, H. C. Cheng, S. Kim, B. Chen, L. Guo, M. Llaguno, J. V. Hollingsworth, D. S. King, S. F. Banani, P. S. Russo, Q. X. Jiang, B. T. Nixon, and M. K. Rosen. 2012. 'Phase transitions in the assembly of multivalent signalling proteins', *Nature*, 483: 336-40.

Liao, S. E., and O. Regev. 2021. 'Splicing at the phase-separated nuclear speckle interface: a model', *Nucleic Acids Res*, 49: 636-45.

Long, J. C., and J. F. Caceres. 2009. 'The SR protein family of splicing factors: master regulators of gene expression', *Biochem J*, 417: 15-27.

Lunde, B. M., C. Moore, and G. Varani. 2007. 'RNA-binding proteins: modular design for efficient function', *Nat Rev Mol Cell Biol*, 8: 479-90.

Lundstrom, P., A. Ahlner, and A. T. Blissing. 2012. 'Isotope labeling methods for large systems', *Adv Exp Med Biol*, 992: 3-15.

Ma, B., S. Kumar, C. J. Tsai, and R. Nussinov. 1999. 'Folding funnels and binding mechanisms', *Protein Eng*, 12: 713-20.

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: 408-11.

Manley, J. L., and R. Tacke. 1996. 'SR proteins and splicing control', *Genes Dev*, 10: 1569-79.

Martin, E. W., A. S. Holehouse, I. Peran, M. Farag, J. J. Incicco, A. Bremer, C. R. Grace, A. Soranno, R. V. Pappu, and T. Mittag. 2020. 'Valence and patterning of aromatic residues determine the phase behavior of prion-like domains', *Science*, 367: 694-99.

Martinez-Contreras, R., J. F. Fisette, F. U. Nasim, R. Madden, M. Cordeau, and B. Chabot. 2006. 'Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing', *PLoS Biol*, 4: e21.

Marzahn, M. R., S. Marada, J. Lee, A. Nourse, S. Kenrick, H. Zhao, G. Ben-Nissan, R. M. Kolaitis, J. L. Peters, S. Pounds, W. J. Errington, G. G. Prive, J. P. Taylor, M. Sharon, P. Schuck, S. K. Ogden, and T. Mittag. 2016. 'Higher-order oligomerization promotes localization of SPOP to liquid nuclear speckles', *EMBO J*, 35: 1254-75.

Masliah, G., C. Maris, S. L. Konig, M. Yulikov, F. Aeschimann, A. L. Malinowska, J. Mabille, J. Weiler, A. Holla, J. Hunziker, N. Meisner-Kober, B. Schuler, G. Jeschke, and F. H. Allain. 2018. 'Structural basis of siRNA recognition by TRBP double-stranded RNA binding domains', *EMBO J*, 37.

Matlin, A. J., and M. J. Moore. 2007. 'Spliceosome assembly and composition', *Adv Exp Med Biol*, 623: 14-35.

Mintz, P. J., S. D. Patterson, A. F. Neuwald, C. S. Spahr, and D. L. Spector. 1999. 'Purification and biochemical characterization of interchromatin granule clusters', *EMBO J*, 18: 4308-20.

Misteli, T., and D. L. Spector. 1996. 'Serine/threonine phosphatase 1 modulates the subnuclear distribution of pre-mRNA splicing factors', *Mol Biol Cell*, 7: 1559-72.

More, D. A., and A. Kumar. 2020. 'SRSF3: Newly discovered functions and roles in human health and diseases', *Eur J Cell Biol*, 99: 151099.

Morin, S. 2011. 'A practical guide to protein dynamics from 15N spin relaxation in solution', *Prog Nucl Magn Reson Spectrosc*, 59: 245-62.

Moursy A., Cléry A., Gerhardy S., Betz K. M., Rao S., Mazur J., Campagne S., Beusch I., Duszczyk M. M. , Robinson M. D., Panse V. G., and Allain F. H.-T. 2023. "RNA recognition by Npl3p reveals U2 snRNA-binding compatible with a chaperone role during splicing." In.: bioRxiv.

Moursy, A., F. H. Allain, and A. Clery. 2014. 'Characterization of the RNA recognition mode of hnRNP G extends its role in SMN2 splicing regulation', *Nucleic Acids Res*, 42: 6659-72.

Ninomiya, K., S. Adachi, T. Natsume, J. Iwakiri, G. Terai, K. Asai, and T. Hirose. 2020. 'LncRNA-dependent nuclear stress bodies promote intron retention through SR protein phosphorylation', *EMBO J*, 39: e102729.

Nott, T. J., E. Petsalaki, P. Farber, D. Jervis, E. Fussner, A. Plochowietz, T. D. Craggs, D. P. Bazett-Jones, T. Pawson, J. D. Forman-Kay, and A. J. Baldwin. 2015. 'Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles', *Mol Cell*, 57: 936-47.

Okada, M., Y. Tateishi, E. Nojiri, T. Mikawa, S. Rajesh, H. Ogasa, H. Yagi, T. Kohno, T. Kigawa, I. Shimada, P. Güntert, Y. Ito, and T. Ikeya. 2021. "Multi-state structure determination and dynamics analysis reveals a new ubiquitin-recognition mechanism in yeast ubiquitin C-terminal hydrolase." In.

Ortiz-Sanchez, P., M. Villalba-Orero, M. M. Lopez-Olaneta, J. Larrasa-Alonso, F. Sanchez-Cabo, C. Marti-Gomez, E. Camafeita, J. M. Gomez-Salinero, L. Ramos-Hernandez, P. J. Nielsen, J. Vazquez, M. Muller-McNicoll, P. Garcia-Pavia, and E. Lara-Pezzi. 2019. 'Loss of SRSF3 in Cardiomyocytes Leads to Decapping of Contraction-Related mRNAs and Severe Systolic Dysfunction', *Circ Res*, 125: 170-83.

Ottoz, D. S. M., and L. E. Berchowitz. 2020. 'The role of disorder in RNA binding affinity and specificity', *Open Biol*, 10: 200328.

Pannier, M., S. Veit, A. Godt, G. Jeschke, and H. W. Spiess. 2000. 'Dead-time free measurement of dipole-dipole interactions between electron spins', *J Magn Reson*, 142: 331-40.

Parenteau, J., M. Durand, S. Veronneau, A. A. Lacombe, G. Morin, V. Guerin, B. Cecez, J. Gervais-Bird, C. S. Koh, D. Brunelle, R. J. Wellinger, B. Chabot, and S. Abou Elela. 2008. 'Deletion of many yeast introns reveals a minority of genes that require splicing for function', *Mol Biol Cell*, 19: 1932-41.

Pascual, M., M. Vicente, L. Monferrer, and R. Artero. 2006. 'The Muscleblind family of proteins: an emerging class of regulators of developmentally programmed alternative splicing', *Differentiation*, 74: 65-80.

Patel, A. A., and J. A. Steitz. 2003. 'Splicing double: insights from the second spliceosome', *Nat Rev Mol Cell Biol*, 4: 960-70.

Pederson, T. 2011. 'The nucleolus', *Cold Spring Harb Perspect Biol*, 3.

Peran, I., and T. Mittag. 2020. 'Molecular structure in biomolecular condensates', *Curr Opin Struct Biol*, 60: 17-26.

Phair, R. D., and T. Misteli. 2000. 'High mobility of proteins in the mammalian cell nucleus', *Nature*, 404: 604-9.

Polyhach, Y., E. Bordignon, R. Tschaggelar, S. Gandra, A. Godt, and G. Jeschke. 2012. 'High sensitivity and versatility of the DEER experiment on nitroxide radical pairs at Q-band frequencies', *Phys Chem Chem Phys*, 14: 10762-73.

Pozzi, N., A. D. Vogt, D. W. Gohara, and E. Di Cera. 2012. 'Conformational selection in trypsin-like proteases', *Curr Opin Struct Biol*, 22: 421-31.

Rappsilber, J., U. Ryder, A. I. Lamond, and M. Mann. 2002. 'Large-scale proteomic analysis of the human spliceosome', *Genome Res*, 12: 1231-45.

Reyes, C. M., and P. A. Kollman. 2000. 'Structure and thermodynamics of RNA-protein binding: using molecular dynamics and free energy analyses to calculate the free energies of binding and conformational change', *J Mol Biol*, 297: 1145-58.

Rogozin, I. B., L. Carmel, M. Csuros, and E. V. Koonin. 2012. 'Origin and evolution of spliceosomal introns', *Biol Direct*, 7: 11.

Roth, M. B., C. Murphy, and J. G. Gall. 1990. 'A monoclonal antibody that recognizes a phosphorylated epitope stains lampbrush chromosome loops and small granules in the amphibian germinal vesicle', *J Cell Biol*, 111: 2217-23.

Saitoh, N., C. S. Spahr, S. D. Patterson, P. Bubulya, A. F. Neuwald, and D. L. Spector. 2004. 'Proteomic analysis of interchromatin granule clusters', *Mol Biol Cell*, 15: 3876-90.

Sakharkar, M. K., B. S. Perumal, K. R. Sakharkar, and P. Kangueane. 2005. 'An analysis on gene architecture in human and mouse genomes', *In Silico Biol*, 5: 347-65.

Sanford, J. R., and J. P. Bruzik. 1999. 'Developmental regulation of SR protein phosphorylation and activity', *Genes Dev*, 13: 1513-8.

Sanford, J. R., J. Ellis, and J. F. Caceres. 2005. 'Multiple roles of arginine/serine-rich splicing factors in RNA processing', *Biochem Soc Trans*, 33: 443-6.

Schiemann, O., N. Piton, J. Plackmeyer, B. E. Bode, T. F. Prisner, and J. W. Engels. 2007. 'Spin labeling of oligonucleotides with the nitroxide TPA and use of PELDOR, a pulse EPR method, to measure intramolecular distances', *Nat Protoc*, 2: 904-23.

Serrano, P., B. E. Aubol, M. M. Keshwani, S. Forli, C. T. Ma, S. K. Dutta, M. Geralt, K. Wuthrich, and J. A. Adams. 2016. 'Directional Phosphorylation and Nuclear Transport of the Splicing Factor SRSF1 Is Regulated by an RNA Recognition Motif', *J Mol Biol*, 428: 2430-45.

Shen, H., and M. R. Green. 2004. 'A pathway of sequential arginine-serine-rich domain-splicing signal interactions during mammalian spliceosome assembly', *Mol Cell*, 16: 363-73.

Shen, H., J. L. Kan, and M. R. Green. 2004. 'Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly', *Mol Cell*, 13: 367-76.

Shin, Y., and C. P. Brangwynne. 2017. 'Liquid phase condensation in cell physiology and disease', *Science*, 357.

Skelton, N.J., A.G. Palmer, M. Akke, J. Kordel, M. Rance, and W.J. Chazin. 1993. 'Practical Aspects of Two-Dimensional Proton-Detected 15N Spin Relaxation Measurements', *Journal of Magnetic Resonance, Series B*.

Skotheim, R. I., and M. Nees. 2007. 'Alternative splicing in cancer: noise, functional, or systematic?', *Int J Biochem Cell Biol*, 39: 1432-49.

Skrisovska, L., and F. H. Allain. 2008. 'Improved segmental isotope labeling methods for the NMR study of multidomain or large proteins: application to the RRMs of Npl3p and hnRNP L', *J Mol Biol*, 375: 151-64.

Sliskovic, I., H. Eich, and M. Muller-McNicoll. 2022. 'Exploring the multifunctionality of SR proteins', *Biochem Soc Trans*, 50: 187-98.

Soret, J., M. Gabut, and J. Tazi. 2006. 'SR proteins as potential targets for therapy', *Prog Mol Subcell Biol*, 44: 65-87.

Sreedharan, J., I. P. Blair, V. B. Tripathi, X. Hu, C. Vance, B. Rogelj, S. Ackerley, J. C. Durnall, K. L. Williams, E. Buratti, F. Baralle, J. de Belleroche, J. D. Mitchell, P. N. Leigh, A. Al-Chalabi, C. C. Miller, G. Nicholson, and C. E. Shaw. 2008. 'TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis', *Science*, 319: 1668-72.

Stamm, S., S. Ben-Ari, I. Rafalska, Y. Tang, Z. Zhang, D. Toiber, T. A. Thanaraj, and H. Soreq. 2005. 'Function of alternative splicing', *Gene*, 344: 1-20.

Stoltzfus, C. M., and J. M. Madsen. 2006. 'Role of viral splicing elements and cellular RNA binding proteins in regulation of HIV-1 alternative RNA splicing', *Curr HIV Res*, 4: 43-55.

Strotz, D., J. Orts, C. N. Chi, R. Riek, and B. Vogeli. 2017. 'eNORA2 Exact NOE Analysis Program', *J Chem Theory Comput*, 13: 4336-46.

Tacke, R., and J. L. Manley. 1999. 'Determinants of SR protein specificity', *Curr Opin Cell Biol*, 11: 358-62.

Tan, W., W. Wang, and Q. Ma. 2018. 'Physiological and Pathological Function of Serine/Arginine-Rich Splicing Factor 4 and Related Diseases', *Biomed Res Int*, 2018: 3819719.

Tang, C., C. D. Schwieters, and G. M. Clore. 2007. 'Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR', *Nature*, 449: 1078-82.

Tang, L. 2019. 'Liquid phase separation', *Nat Methods*, 16: 18.

Tazi, J., N. Bakkour, and S. Stamm. 2009. 'Alternative splicing and disease', *Biochim Biophys Acta*, 1792: 14-26.

Tintaru, A. M., G. M. Hautbergue, A. M. Hounslow, M. L. Hung, L. Y. Lian, C. J. Craven, and S. A. Wilson. 2007. 'Structural and functional analysis of RNA and TAP binding to SF2/ASF', *EMBO Rep*, 8: 756-62.

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: 61-76.

Twyffels, L., C. Gueydan, and V. Kruys. 2011. 'Shuttling SR proteins: more than splicing factors', *FEBS J*, 278: 3246-55.

Ule, J., and B. J. Blencowe. 2019. 'Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution', *Mol Cell*, 76: 329-45.

Venables, J. P. 2006. 'Unbalanced alternative splicing and its significance in cancer', *Bioessays*, 28: 378-86.

Vitali, F., A. Henning, F. C. Oberstrass, Y. Hargous, S. D. Auweter, M. Erat, and F. H. Allain. 2006. 'Structure of the two most C-terminal RNA recognition motifs of PTB using segmental isotope labeling', *EMBO J*, 25: 150-62.

Vogeli, B., P. Guntert, and R. Riek. 2013. 'Multiple-state ensemble structure determination from eNOE spectroscopy', *Molecular Physics*, 111: 437-54.

Vogt, A. D., N. Pozzi, Z. Chen, and E. Di Cera. 2014. 'Essential role of conformational selection in ligand binding', *Biophys Chem*, 186: 13-21.

Wagner, R. E., and M. Frye. 2021. 'Noncanonical functions of the serine-arginine-rich splicing factor (SR) family of proteins in development and disease', *Bioessays*, 43: e2000242.

Waudby, C. A., A. Ramos, L. D. Cabrita, and J. Christodoulou. 2016. 'Two-Dimensional NMR Lineshape Analysis', *Sci Rep*, 6: 24826.

Weidtkamp-Peters, S., T. Lenser, D. Negorev, N. Gerstner, T. G. Hofmann, G. Schwanitz, C. Hoischen, G. Maul, P. Dittrich, and P. Hemmerich. 2008. 'Dynamics of component exchange at PML nuclear bodies', *J Cell Sci*, 121: 2731-43.

Weikl, T. R., and F. Paul. 2014. 'Conformational selection in protein binding and function', *Protein Sci*, 23: 1508-18.

Will, C. L., and R. Luhrmann. 2001. 'Spliceosomal UsnRNP biogenesis, structure and function', *Curr Opin Cell Biol*, 13: 290-301.

———. 2011. 'Spliceosome structure and function', *Cold Spring Harb Perspect Biol*, 3.

Williamson, M. P. 2013. 'Using chemical shift perturbation to characterise ligand binding', *Progress in Nuclear Magnetic Resonance Spectroscopy*, 73: 1-16.

Wilusz, J. E., H. Sunwoo, and D. L. Spector. 2009. 'Long noncoding RNAs: functional surprises from the RNA world', *Genes Dev*, 23: 1494-504.

Wu, J. Y., and T. Maniatis. 1993. 'Specific interactions between proteins implicated in splice site selection and regulated alternative splicing', *Cell*, 75: 1061-70.

Wurz, J. M., S. Kazemi, E. Schmidt, A. Bagaria, and P. Guntert. 2017. 'NMR-based automated protein structure determination', *Arch Biochem Biophys*, 628: 24-32.

Xiang, S., V. Gapsys, H. Y. Kim, S. Bessonov, H. H. Hsiao, S. Mohlmann, V. Klaukien, R. Ficner, S. Becker, H. Urlaub, R. Luhrmann, B. de Groot, and M. Zweckstetter. 2013. 'Phosphorylation drives a dynamic switch in serine/arginine-rich proteins', *Structure*, 21: 2162-74.

Yulikov, M. 2015. *Electron Paramagnetic Resonance*.

Zahler, A. M., W. S. Lane, J. A. Stolk, and M. B. Roth. 1992. 'SR proteins: a conserved family of pre-mRNA splicing factors', *Genes Dev*, 6: 837-47.

Zhang, Y. J., Y. F. Xu, C. A. Dickey, E. Buratti, F. Baralle, R. Bailey, S. Pickering-Brown, D. Dickson, and L. Petrucelli. 2007. 'Progranulin mediates caspase-dependent cleavage of TAR DNA binding protein-43', *J Neurosci*, 27: 10530-4.

Zheng, X., Q. Peng, L. Wang, X. Zhang, L. Huang, J. Wang, and Z. Qin. 2020. 'Serine/arginine-rich splicing factors: the bridge linking alternative splicing and cancer', *Int J Biol Sci*, 16: 2442-53.

Zhong, X. Y., P. Wang, J. Han, M. G. Rosenfeld, and X. D. Fu. 2009. 'SR proteins in vertical integration of gene expression from transcription to RNA processing to translation', *Mol Cell*, 35: 1-10.

Zhou, Z., and X. D. Fu. 2013. 'Regulation of splicing by SR proteins and SR protein-specific kinases', *Chromosoma*, 122: 191-207.

Zhou, Z., L. J. Licklider, S. P. Gygi, and R. Reed. 2002. 'Comprehensive proteomic analysis of the human spliceosome', *Nature*, 419: 182-5.

Zhu, J., A. Mayeda, and A. R. Krainer. 2001. 'Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins', *Mol Cell*, 8: 1351-61.

Acknowledgement

Six years ago, I started my PhD, my first experience away from home and my family. This provided me with the opportunity to meet many people who helped me both professionally and personally.

Firstly, I would like to thank **Prof. Frédéric Allain** for accepting me as a PhD student, even though I was a "catastrophe" or a "disaster," and for letting me share the office with him! Overall, to give me the chance to grow and improve.
Secondly, **Antoine** for the help and support over the years and for answering many questions about SRSF1 and how to prepare presentations, a skill that I improved over time with your help. **Fred D** for the big help in performing structure calculations and all the zoom calls during the Covid period.

In addition, I am grateful to my committee, **Prof. Remco Sprangers** and **Prof. Karsten Weis**, for their helpful feedback and support during our committee meetings.
On the NMR side, **Alvar** and **Simon** for providing significant assistance and support when I had to use the NMR facility, and the nice work in preparing the NMR seminars. I also want to thank **Prof. Peter Güntert** for introducing me to the Multistate CYANA calculation, which was an essential step for my project!

Next, I want to thank the "EPR people", **Prof Gunnar Jeschke**, **Maxim**, and **Daniel**, for continuing the "traditional collaboration" between the EPR and NMR groups. However, my greatest thank goes to **Laurina**; it was fantastic to work with you as collaborator and as friend. I enjoyed working with you, chatting, and reciprocating support in any situation!
On this note, I would like to thank all the people I met through the **Sinergia** network (**Schuler**, **Polymenidou** and **Leitner** groups). It was a great collaboration full of productive discussions and feedback.

I would like to express my gratitude to all my colleagues on the **HPP-L floor**.
First and foremost, I want to thank **Tebbino** and **LORON**: I know it was not easy having me around all the time and, complaining in most of the situations; but I am sure I made your PhD less boring! Of course, **Leonidas**, the "big brother" who was always there to listen and give wise suggestions, as well as for the enjoyable moments outside of the lab. I also thank **Antje** for all the lovely dinners and tea-chatting moments we had together! **Ahmed**, for your significant support at the beginning of my PhD, I always kept your advice and kind words in mind. The PhD mum **Irene**, always ready to assist me, even from the other side of the world. **Dominik**, the "rompipalle," for making noise with me, and **Rahel**, for your immense patience.

Lastly, I want to thank all the other members from the **Allain, Jonas, and Gossert labs**. Everyone had to listen to me complain at least once! Thanks for the all the beers-moment after working hours too.

Giacomo, **Chiaretta**, **Martina** e tutti gli altri italiani… è bello avere altri che ti fanno compagnia e con cui potersi lamentare in ITALIANO!

I matematici: **Giuliano**, **Kircheis**, **Mascotto**, **Il Mantegazza** e la **Panzer** per le Skypate e le grandi risate
I pregnanesi: **Silvietta**, **Cri**, **Saretta**, **Alessio** (Mario) e **Lara**… grazie per i momenti (anche se brevi) quando torno a casa, per farmi sentire come se non fossi quasi mai partita… e per tutti i pettegolezzi e i vari aggiornamenti sulla mitica Pregna!
Giulia: non mi stancherò mai di commentare con te film e serie TV!
E ovviamente la **Debby**, la storica amica su cui posso contare sempre e ovunque!
Grazie alla mia **Famiglia**, in particolare **la Mami** per le lunghe chiamate e **il Gu**, avere scorte di cibo italiano è sempre stato un grande sostegno.
E infine, ma non di certo per importanza, **il Fontana**: in 10 anni, hai davvero visto la mia crescita sia come persona che come scienziata; hai vissuto l'intero percorso di PhD con me con annessi alti e bassi… non servono altre parole, grazie di tutto!

Curriculum Vitae

Personal information

Name: Cristina Kim Xuan Nguyen
Date of birth: October 4, 1990
Place of birth: Milan, Italy
Nationality: Italian
Email: cr.nguyen@outlook.com

Education

2017-now	PhD in Structural Biology ETH Zürich (Switzerland)
2013-2016	Master's Degree in Biology University of Milano-Bicocca (Milan, Italy)
2010-2013	Bachelor's Degree in Biology University of Milano-Bicocca (Milan, Italy)

Research experience

06/2017-current	PhD student Institute of Biochemistry, D-BIOL, ETH Zürich (Switzerland) Supervisor: Prof. Dr. Frédéric Allain
03/2016-04/2017	Fellowship Division of Genetics and Cell Biology, DIBIT, San Raffaele Scientific Institute, Milan (Italy) Supervisor: Dr. Daniela Talarico
10/2014-02/2016	Internship for Master's Degree, experimental thesis Division of Genetics and Cell Biology, DIBIT, San Raffaele Scientific Institute, Milan (Italy) Supervisor: Dr. Daniela Talarico
09/2013-10/2013	Stage Department of Biotechnology and Biosciences, Università degli Studi di Milano-Bicocca (Milan, Italy) Supervisor: Prof. Dr. Silvia Maria Luisa Barabino

Language skills

Italian	Native Language
English	Full professional proficiency
German	Basic

Publications

Leonidas Emmanouilidis, Laura Esteban-Hofer, Fred F Damberger, Tebbe de Vries, **Cristina K X Nguyen**, Luis Fábregas Ibáñez, Simon Mergenthal, Enrico Klotzsch, Maxim Yulikov, Gunnar Jeschke, Frédéric H-T Allain (**2021**). NMR and EPR reveal a compaction of the RNA-binding protein FUS upon droplet formation. *Nat Chem Biol*, 17(5):608-614

Antoine Cléry, Miroslav Krepl, **Cristina K X Nguyen**, Ahmed Moursy, Hadi Jorjani, Maria Katsantoni, Michal Okoniewski, Nitish Mittal, Mihaela Zavolan, Jiri Sponer, Frédéric H-T Allain (**2021**). Structure of SRSF1 RRM1 bound to RNA reveals an unexpected bimodal mode of interaction and explains its involvement in SMN1 exon7 splicing. *Nat Commun*, 12(1):428

Antoine Cléry, Laurent Gillioz, **Cristina K X Nguyen**, Frédéric H-T Allain (**2019**). A Step-by-Step Guide to Study Protein-RNA Interactions. *CHIMIA*, 73(6):406-414