TUM School of Natural Sciences
Technische Universität München

Structural analysis of multi-domain RNA binding proteins in 3' splice site recognition and mRNA-protein assembly

Nitin Kachariya

Vollständiger Abdruck der von der TUM School of Natural Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:

Prof. Dr. Bernd Reif

Prüfer*innen der Dissertation:

1. Prof. Dr. Michael Sattler
2. Prof. Dr. Cathleen Zeymer

Die Dissertation wurde am 26.06.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Natural Sciences am 05.09.2023 angenommen.

“What we know is a drop,
What we don’t know is an Ocean”
- **Isaac Newton**

“The thing that doesn’t fit is the thing that’s the most
Interesting: the part that doesn’t go according to what
you expected.”
- **Richard Feynman**

“You dig deeper and it gets more and more complicated,
and you get confused, and it’s tricky and it’s hard,
but… it is beautiful.”
- **Brian Cox**

Table of Contents

Abstract

RNA binding proteins (RBPs) and ribonucleoprotein particles (RNPs) play a significant role in regulating post-transcriptional gene expression. In eukaryotes, the production of mature mRNAs involves the removal of non-coding introns through pre-mRNA splicing. In humans, specific cis-regulatory RNA elements, the 5' splice site (ss), the branch point site (BPS), the polypyrimidine tract (PPT), and 3' ss are found at the intron/exon boundaries of the pre-mRNA transcripts and are recognized by the U1 small nuclear ribonucleoprotein particle (snRNP), the heterodimeric U2 auxiliary factor 1 and 2 (U2AF1/U2AF2), and splicing factor 1 (SF1), respectively. Protein-protein interactions between these factors also contribute to the assembly, such as U2AF2 binds to U2AF1 and SF1 via UHM (U2AF homology motif) and ULM (UHM ligand motif) interactions. SF1 has a multidomain architecture at N-terminal ULM, followed by HH (helix-hairpin), KH (hnRNP K homology), Qua2 (Quaking 2), and a C-terminal proline-rich motif, whereas KH-Qua2 recognizes “YNYURAY” (A: adenosine branch site, R: any nucleotide, Y: pyrimidine) BPS motif.

The work presented in the thesis shows that altering the consensus “U” at the +2 upstream position of BPS abolishes the Qua2 interactions and reduces SF1 specificity, which can lead to an abrupt splicing products in subsequent steps. SF1 significantly impacts U2AF2 to recognize PPT indicated by *in vitro* iCLIP study. In addition, SF1’s HH domain has a conserved “RSPSP” motif known for serine phosphorylation and iCLIP analysis shows no significant binding preference between non- and phosphorylated SF1-U2AF2 complex with a library of pre-mRNA sequences. However, minor changes in NMR chemical shifts observed upon phosphorylation advocate the SF1-U2AF2's weak and transient interactions, which may not differentiate the 3’ splice site recognition. Moreover, intron sequences have varying strengths and spacing between BPS and PPT, or multiple degenerate BPS-like sequences. In this work it is shown that the SF1-U2AF2 complex can adopt compact to extended conformations depending upon the strength of the BPS and PPT sites, enabled by flexible domain connections. Moreover, the SF1-U2AF2 complex shows a binding preference for nearby splice sites in the case of multiple BPS-like sites on the same intron. A proposed ensemble model elucidates the flexible conformations of the SF1-U2AF2 complex, where in the absence of RNA, SF1’s KH-Qua2 and U2AF2’s RRM1,2 domains stay apart, while RNA with strong BPS and PPT brings both of them in the proximity to attain a compact conformation.

These findings highlight the dynamic role of SF1-U2AF2 in 3’ splice site recognition in the early stage of spliceosome assembly.

Second part of the thesis addresses the role of Npl3 in mRNA stability and nuclear export. In *Saccharomyces cerevisiae*, Npl3, Yra1, Nab2, with other facilitator proteins recruit nascent mRNA and export to the nuclear pore complex through the adapter proteins. The second part of the thesis focuses on the structure, dynamics, and function of the Npl3 protein in *S. cerevisiae*. Npl3 has tandem RRMs in the middle for RNA recognition, where RRM1 has a canonical RRM fold with conserved RNP1 and RNP2 sites, however RRM2 has a conserved α1 helix with “SWQDLKD” motif with non-conserved RNPs, making it a pseudo-RRM. The current work shows that RRM1 and RRM2 of Npl3 precisely recognize “CC” and “GG” RNA motifs, respectively, and both RRMs are oriented facing the RNA binding interface in solution, forming the positively charged surface area to facilitate the RNA binding. Mass spectrometry-based derived RRM1 (F162), linker (P196, A197), and RRM2 (F245) mutants show variable temperature-sensitive phenotypes, reduced mRNA binding, and mRNP assembly in yeast. The structural analysis demonstrated that RRM1 or linker mutations do not affect the RRM structure, whereas the RRM2 mutant results in the miss-folding of tertiary structure, so it acts as a loss of function *in vivo*. Moreover, reduced RNA binding was observed for the linker, RRM1, and RRM2 mutants with respect to the wildtype *in vitro* binding study. In conclusion, the study highlights the novel structural and functional regulation of Npl3 in RNA recognition and nuclear mRNA assembly.

The work presents in this thesis highlights how multidomain RNA binding proteins modulate the early stage of spliceosome assembly, nuclear mRNA assembly and export.

Chapter 1 – Introduction

RNA binding proteins (RBPs) and applied methods to study RBPs involved in splicing and mRNP assembly

RNA binding proteins (RBPs)

1.1 Messenger RNA (mRNA) processing

Eukaryotic cells have a precise system that controls gene expression, processing, and metabolism for smooth cell functioning, handle stress conditions, and hence is crucial for cell survival. These processes happen in different cell compartments, such as the nucleus, where pre-messenger RNA (pre mRNA) is synthesized, the cytoplasm, where messenger RNA is translated, and mRNA and protein breakdown follows (Coppin et al., 2018).

The journey of mRNA begins with transcription, continues through translation and ends in degradation. Throughout this journey, mRNA interacts with RNA binding proteins (RBPs) and small non-coding RNAs (snRNAs) to form a messenger ribonucleoprotein (mRNP). In the nucleus, mRNA interacts with various RBPs and mRNPs to regulate transcription initiation. The nascent transcripts undergo further processing by RBPs and mRNPs, such as 5' capping, intron splicing, cleavage, and 3' poly-adenylation. Proofreading steps ensure mRNA quality before it is exported to the cytoplasm. In the cytoplasm, mRNA has multiple fates, including translation, subcellular localization, and mRNA turnover, which involves changing the cohort of associated RBPs (**Figure 1.1**). Together, these processes make up the mRNA life cycle.

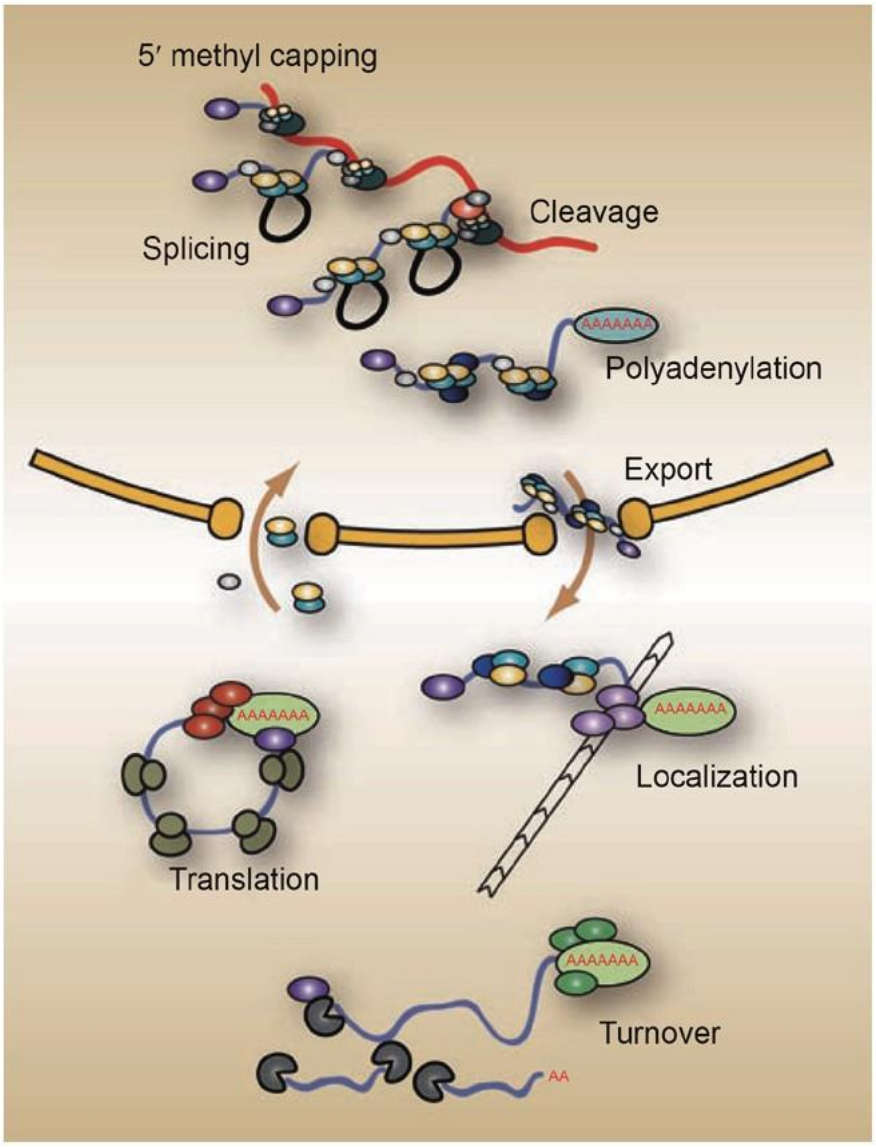


Figure 1.1. Schematic illustration of the general mRNA life cycle. The pre-mRNA (shown in red) undergoes post-transcriptional processing, including 5' capping, splicing, 5' polyadenylation, and mRNPs assembly in the nucleus. Mature mRNA is then transported to the cytoplasm for subcellular localization, translation, and degradation (Figure adapted from McKee and Silver 2007).

Some RNA binding proteins (RBPs) stay attached to RNA for most of their lifespan, while others bind only for a brief period to serve a specific purpose. However, it is not yet fully clear how RBPs control gene regulation by being selective and specific throughout the lifespan of mRNA. To gain a better insight, further studies are needed at both the cellular and structural levels of RBPs (Coppin et al., 2018) (McKee & Silver, 2007).

1.2 RNA binding proteins (RBPs)

RBPs are ubiquitously expressed proteins with central and conserved roles in gene regulation. They are found in abundance throughout the human genome, with 1542 genes accounting for over 7.5% of protein-coding genes. RBPs interact with a variety of RNAs, including mRNAs, ribosomal RNAs (rRNAs), microRNAs (miRNAs), transfer RNAs (tRNAs), small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs), PIWI-interacting RNAs (piRNAs), long non-coding RNAs (lncRNAs) and other regulatory RNAs (**Figure 1.2 A**). These interactions can either regulate RNA metabolism, including processing, stability, translation, import and export, or stabilize the RBPs for subsequent function, localization, and more (**Figure 1.2 B**). Thus, understanding the dynamic, competitive, and complex associations between RBPs and RNA targets is vital to comprehending RNA metabolism (Gerstberger et al., 2014).

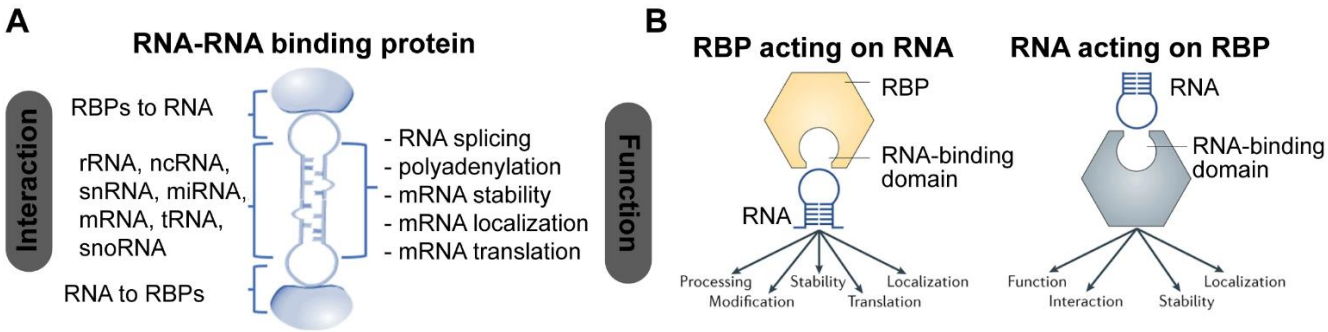


Figure 1.2. The role of RNA binding proteins (RBPs). (A) Various types of interactions between RBPs and RNA are shown. (B) Schematic shows the function of RBPs involved in RNA metabolism (left) and vice-versa RNA involvement in stability and localization of RBP (right). (figure obtained from Qin, Ni et al. 2020 and Hentze, Castello et al. 2018).

Post-translational modifications (PTMs), such as phosphorylation, methylation, acetylation, O-GlcNAcylation (O-Linked β-N-acetylglucosamine modification) and ubiquitination tightly regulate the competitive interactions of RBPs and RNPs (Garcia-Maurino et al., 2017). Additionally, various pathological conditions, including developmental disorders, neurodegenerative diseases, muscular atrophies and cancers, have been linked to aberrant expression, mutation, or deregulation of RBPs (Pereira et al., 2017), (Blech-Hermoni & Ladd, 2013) (Cheng & Jansen, 2017) (Ottoz & Berchowitz, 2020) (Qin et al., 2020) (Hentze et al., 2018).

1.3 Types of RNA binding domains in RBPs

RBPs consist of various types of RNA-binding domains (RBDs), which are often present in single or multiple copies. RBDs are classified based on structural conservation such as: (i) RRM (RNA recognition motif), (ii) KH (K-homology) domain, (iii) dsRBD (double-stranded RNA binding domain), (iv) ZnF (zinc-finger) domain, (v) SAM (sterile alpha motif) domain, (vi) TRAP (trp RNA-binding attenuation protein), (vii) PIWI (p-element induced wimpy testis; piRNA binding), (viii) PUF (Pumilio homology) domain, (ix) PAZ (Piwi-Argonaute-Zwille) domain, (x) S1 domain, (xi) DEAD motif, (xii) YTH (YT521-B homology) domain, (xiii) CSD (cold-shock) domain. Most RBDs recognize 4 to 6 nucleotide (nt) segments and are often found in combinations or repeats of RBDs, which enhances the specificity of RBPs for the nucleotide sequences. Additionally, nucleotide recognition can be precise to single or double nucleotides in length, as observed with PUF and ZnF domains (Lunde et al., 2007).

1.4 RNA recognition by multidomain RBPs

RBPs are involved in a vast regulatory mechanism that defines protein function and RNA binding specificity. Most of the RBPs have a multidomain architecture that expands the functional repertoire of the proteins. By combining different domain types and arrangements, multi-domain RBPs create a larger surface area for binding, resulting in improved RNA binding affinity and specificity compared to a single-domain.

Multi-domain RBPs rely on the connecting linkers between their domains to recognize RNA. These linkers are typically unstructured, which enables the RBDs to search and scan RNA targets of different lengths or to modulate RNA binding by intra-molecular interactions. On the other hand, domains connected by short linkers bind to a continuous stretch of RNA motifs.

The position of the RNA motifs, the relative position of the RNA-binding domains, and the flexibility of the linkers together contribute to the RBPs' function, as depicted in **Figure 1.3**. In addition, the multi-domain architecture enables RBPs to arrange flexible RNA in a way that suits their specific function, while the structured domains arrange themselves topologically to interact with structured RNAs. (Ottoz & Berchowitz, 2020) (Garcia-Maurino et al., 2017) (Coppin et al., 2018) (Kelaini et al., 2021) (Castello et al., 2012) (Lunde et al., 2007).

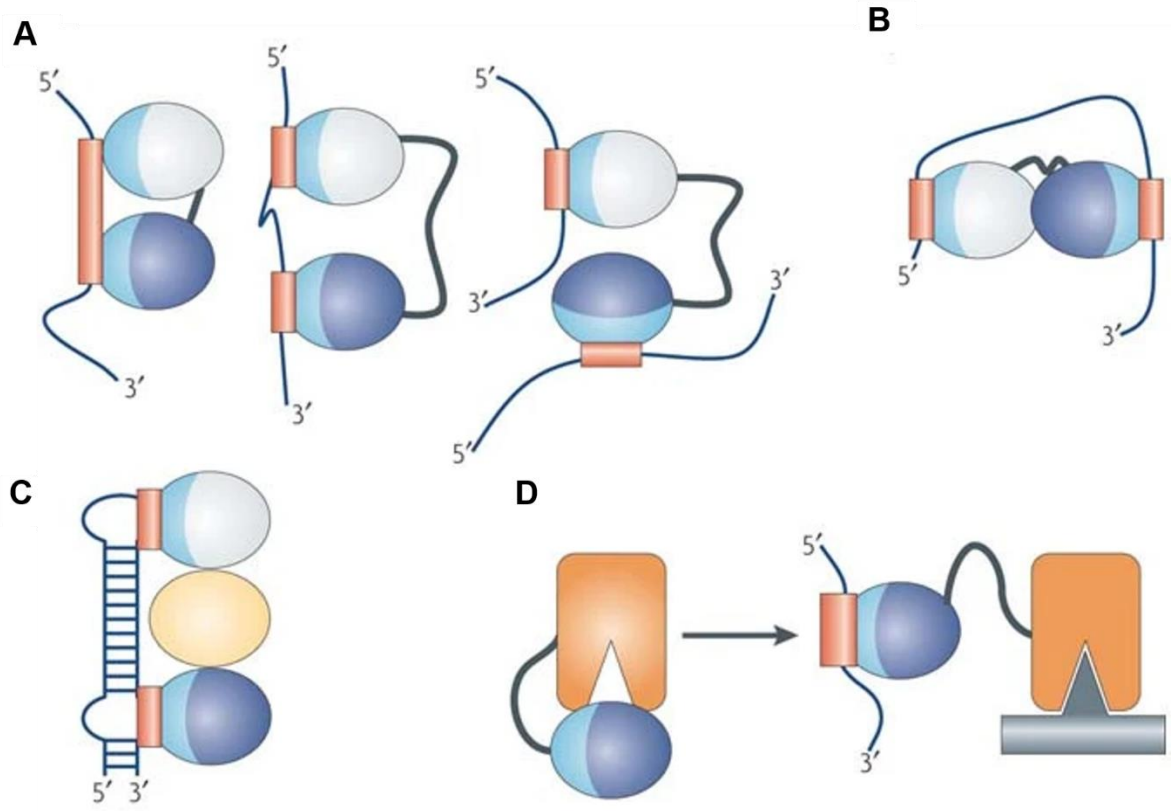


Figure 1.3. Various RNA-binding domains (RBD) and their functions for RNA regulation. (A), (B) Schematic illustration displays multi-domain arrangement and the linker connecting domains regulates the RNA affinity, selectivity, and specificity. (C) Spacer protein assists in positioning RBDs for RNA binding is shown. (D) RBD with enzymatic domain for substrate specificity is illustrated (Schematic is taken from Lunde, Moore et al. 2007).

A typical example of a multidomain RBP is U2 small nuclear RNA auxiliary factor 2 (U2AF2) that has three RRM domains, where RRM1 and RRM2 are connected by a 25 amino acid unstructured linker that mediates recognition of longer stretch of poly-pyrimidine tracts at 3' splice sites. Additionally, U2AF2 has a linker of 35 amino acids that connects to the UHM (U2AF homology motif) domain that adopts an RRM fold but mediates a protein-protein

interaction to a ULM (UHM ligand motif) peptide in the splicing factor SF1, which itself exhibits a KH domain to recognized a branch point sequence upstream of the polypyrimidine tract (PPT). As a result, the U2AF2-SF1 complex can easily find nearby 3' RNA splice sites on the intron due to the inter-domain flexibility it provides.

In contrast to U2AF2 RRM1-2 domains, the Npl3 protein from *Saccharomyces cerevisiae*, which is similar to the SRSF1 (serine-arginine rich splicing factor 1) in humans, has a short linker made up of eight amino acids that more rigidly connects its two RRMs. This linker limits the protein's flexibility and allows it to recognize the nearby nucleotide motifs for each RRM. Some RBPs harbor an enzymatic domain where the RBDs are used to identify the targets for catalytic activity. For instance, KIS kinase has a C-terminal UHM domain that identifies target proteins via the ULM domain and a N-terminal kinase domain that has phosphorylation activity. Therefore, to comprehend the function of RBPs, it is essential to understand their domain function, along with their associated linker and RNA recognition.

Chapter 3 provides a thorough analysis of human U2AF2 and SF1 multi-domain RBP proteins, including their structural and dynamic behaviors in recognizing different 3' splice sites RNA and splicing regulations. Chapter 4 focuses on the structural properties of yeast Npl3 to understand the RNA specificity of RBDs and their involvement in nuclear mRNP assembly and export. The principles, technical details, and practical applications of biophysical techniques utilized to study these RNA-binding proteins are explored in the following chapter.

Background of applied methods

1.5 Basics of NMR theory

Nuclear Magnetic Resonance (NMR) spectroscopy is a powerful and widely used method for studying different types of molecules in both solid and liquid states. This technique involves using radio frequency to stimulate the specific atomic nuclei or "spin" of the molecule, observing its behavior, and extracting structural and dynamic information. The spin of an atomic nucleus is a positively charged spherical object with a magnetic dipole moment (μ) that rotates around an axis, resulting in a precession spin frequency (as shown in **Figure 1.4 A-B**).

Nuclei have a nuclear spin quantum number (m), where those with non-zero values ($m \neq 0$) can be detected by NMR and those with $m = 0$ are NMR-inactive. When there is no external magnetic field (B_0), the direction of magnetic dipole moment (μ) is randomly oriented, and the

energy levels can cancel each other out. However, in the presence of an external B_0 field, the magnetic momentum of NMR active spin will align with the direction of the external B_0 field with a precession frequency known as the Larmor frequency. The Larmor frequency is proportional to the external B_0 and the spin's gyromagnetic ratio (γ). In the presence of an external B_0 field, the spin aligns to the z-axis and creates specific energy levels based on the nuclear spin quantum number (m), described by the equation: 2(I) + 1. For instance, ^{1}H nucleus has m = ½, which generates two energy levels in the presence of B_0 field corresponding to m_I = +½ and m_I = −½, which are called α-state (lower energy) and β-state (higher energy), respectively. The energy difference between spin ±1/2 can be expressed as ΔE= (h/2π) γ B_0, where h is the Planck constant (**Figure 1.4**).

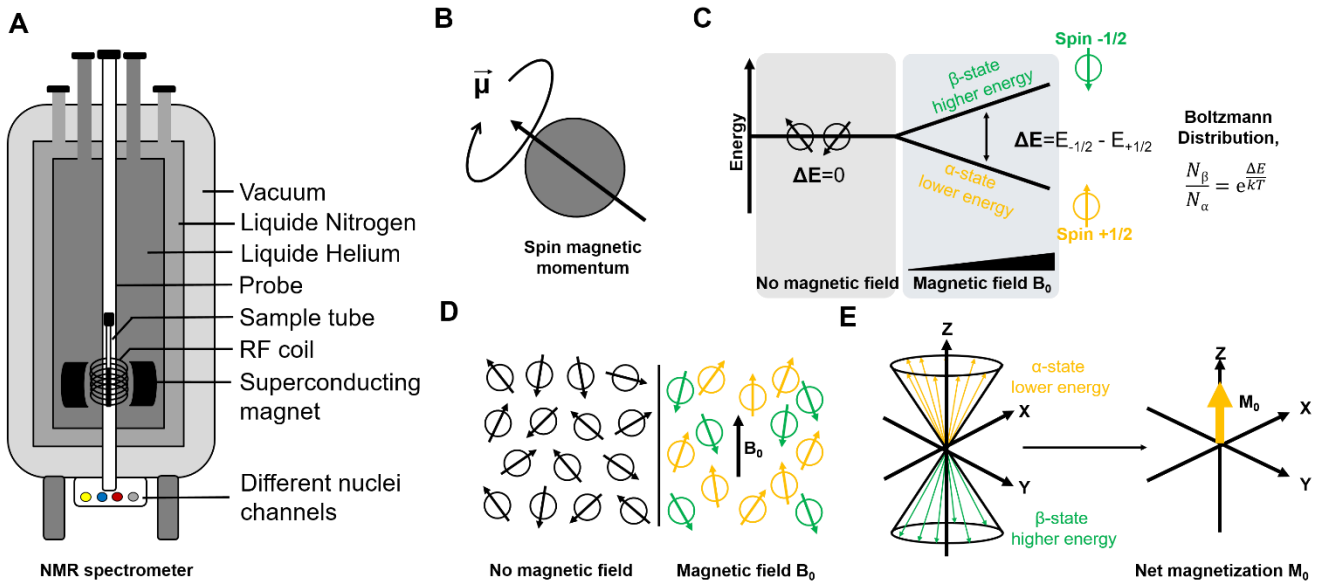


Figure 1.4. Schematic of NMR spectrometer and NMR active spin detection. (A) NMR spectrometer and its various sub-components. (B) Schematic of nuclear spin precession and the direction of magnetic dipolar moment. (C, D) the randomization of the spin magnetic moment without magnetic field, and the alignment of the spin with (in yellow, low energy state) or against (in green, high energy state) the external magnetic field. (E) Schematic of the direction of bulk magnetization along the external magnetic field (z-axis).

1.6 Sensitivity of NMR

The energy difference between the α- and β-states (for spin 1/2) is proportional to the external magnetic field and the gyromagnetic ratio of the spin. This energy difference is small and the population ratio can be calculated using Boltzmann's distribution formula, N_β/N_α = exp (ΔE/kT); where ΔE is the energy difference between the two states, T is the temperature in

Kelvin, k is Boltzmann's constant, N_β and N_α are the spin population in α- and β-state, respectively. For a ^1H spin, the population ratio of α- and β-state is 0.999936 for a 9.4 Tesla magnetic field at 25 °C. However, this small population ratio makes NMR a less sensitive technique, requiring a higher sample concentration for accurate measurements. Nevertheless, the small energy difference between α- and β-states has its advantages. It results in a longer lifetime of the excited state, which is on the milliseconds to seconds timescale and generates narrow line width, allowing for designing multidimensional experiments. Despite being less sensitive than other spectroscopic methods, NMR provides atom-specific high-resolution structural insight of chemical or biomolecular materials. Additionally, NMR sensitivity depends on several factors, including the strength of the external magnetic field, the size of the molecules, the gyromagnetic ratio of excited and observed NMR active nuclei, the concentration of molecules, the number of scans used during measurements and temperature. NMR sensitivity is measured by extracting the signal-to-noise ratio (S/N) of one- or n-dimension (1D or nD) spectra and defined as below equation,

$$\text{NMR Sensitivity, } \frac{S}{N} \propto N \cdot A \cdot T^{-1} \cdot (B_0)^{\frac{3}{2}} \cdot (\gamma_{ext}) \cdot (\gamma_{obs})^{\frac{3}{2}} \cdot T_2^* \cdot (NS)^{\frac{1}{2}}$$

Where, N represents the number of molecules, A is the abundance of NMR active spin, B_0 is the strength of the magnetic field, T refers to the temperature in Kelvin, T_2^* is the transverse relaxation time of NMR active spin, and NS is the number of scans used during the NMR measurement.

1.7 One-dimensional proton NMR measurement

In the presence of an external B_0 field, the NMR active spin precesses at a Larmor frequency (ω) which is an equilibrium state. To transition from its equilibrium state to the excited state between two energy levels, an oscillating magnetic field close to the Larmor frequency is applied along the x or y-axis. This is achieved by applying a millisecond range of radio frequency (RF) pulse (90° or 180° pulse) to the sample, which produces the optimal response for an NMR-active spin. The transition of excited states generates an oscillating magnetic field, and the bulk magnetization returns to equilibrium when the RF pulse switches off. This transition induces a current in the receiver coil, which is recorded over time as a Free Induction Decay (FID). Then applying a Fourier transformation to the FID generates an NMR spectrum with respective frequencies.

As shown in **Figure 1.5**, the simple 1D NMR experiment involves two distinct periods: preparation and detection. During the preparation period, the spins remain in thermal equilibrium, and the bulk magnetization is aligned along the z-axis, which is parallel to the external magnetic field. Following the 90° excitation pulse applied with an oscillating B_1 field, the magnetization rotates in the xy-plane. The detection period begins after the 90° pulse is switched off and measures the decay of the time domain FID as the excited spins return to thermal equilibrium under the external B_0 field.

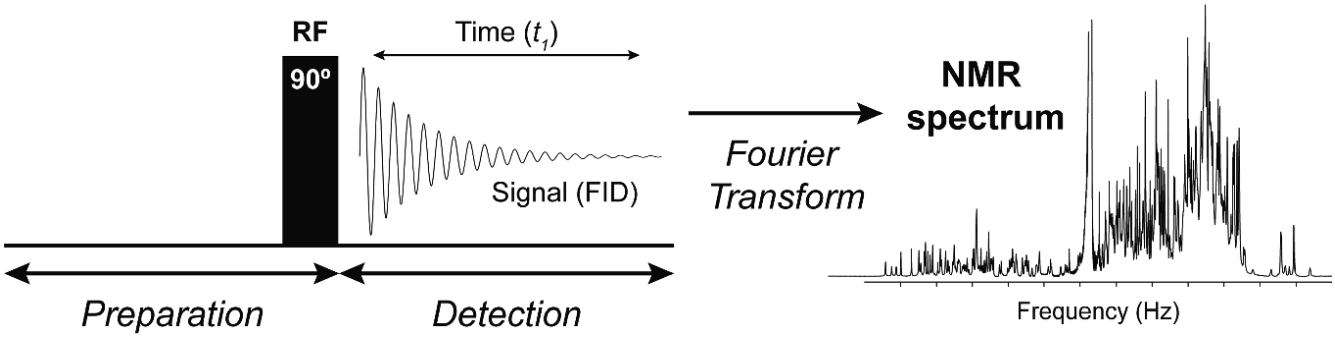


Figure 1.5. Schematic of a basic 1D pulse scheme and NMR spectrum. The left side displays the preparation and detection of a simple 1D NMR pulse scheme. The right side shows a 1D NMR spectrum obtained after Fourier transformation of the time domain FID.

1.8 2D and 3D NMR for protein backbone assignment

1D NMR spectra of biomolecule often have poor resolution due to overlapping signals, making it difficult to identify them. As a result, multi-dimensional NMR experiments are commonly used in biomolecular NMR. In a 1D experiment, the FID is recorded with a single time variable, "t_1 evolution time" and a 1D spectrum is generated through Fourier transformation. On the other hand, 2D NMR records the time domain signal as a function of two time variables, t_1 and t_2, resulting in a 2D spectrum after Fourier transformation in direct and indirect dimensions. Therefore, 2D spectra can be viewed as a sequence of multiple 1D spectra obtained with varying t_1 points, with the signal being obtained during t_2 evolution time at the end of the pulse sequence. Additional pulses and delays are added between t_1 and t_2 evolution, depending on whether the 2D spectra are homo- or hetero-nuclear.

Heteronuclear experiments based on ^{1}H-^{15}N and ^{1}H-^{13}C are highly useful in biomolecular NMR and more often uniformly isotopically labeled protein samples are required for measurements. The ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectrum is

considered the fingerprint spectra of a protein, where all backbone residues having amide groups (except for proline) are detected. In protein NMR, multiple 2D pulse elements are combined to generate 3D NMR experiments such as ^{1}H, ^{15}N NOESY-HSQC and ^{1}H, ^{15}N TOCSY-HSQC. A conventional 3D NMR pulse program has three evolution periods, where t_1 and t_2 evolution times correspond to the indirect dimensions, and t_3 evolution corresponds to the direct dimension.

To characterize a protein by NMR, the initial step involves completing backbone and side-chain resonance assignments by combining of 2D and 3D spectra. For sequence assignment of backbone resonances, conventional 3D experiments can be measured using a uniformly ^{15}N and ^{13}C labeled protein, such as HNCACB, HN(CO)CACB, HNCO, and HN(CA)CO (as shown in **Figure 1.6 A**). These experiments are carried out to assign Cα, Cβ, CO, N_H, and H_N backbone atoms of self and preceding residues, which can be utilized to calculate secondary structure and dihedral angles. Similarly, to assign side-chain resonances, different sets of experiments can be recorded, like (H)CC(CO)NH, H(CCCO)NH, HCCH-TOCSY, HCCH-COSY, TOSCY-HSQC, HiSQC, etc. These side-chain experiments are useful in assigning proton and carbon atoms of side-chain resonances, as shown in **Figure 1.6** (Ikura et al., 1990) (Sattler, 1990).

To derive protein structure by NMR, different sets of NMR experiments are required. For instance, the ^{1}H, ^{15}N NOESY-HSQC experiment provides proton-proton intra- and inter-molecular NOE resonances that are close in space, with distances of less than 5 Å. On the other hand, ^{1}H, ^{15}N TOCSY-HSQC experiment is complementary to ^{1}H, ^{15}N NOESY-HSQC, which transfers the magnetization through multiple bonds via J-coupling. So, NOESY and TOCSY experiments can be used for differentiating the intra- and inter-molecular resonances for protein structure calculation. Additionally, experiments such as the ^{1}H, ^{15}N-edited NOSY-HSQC, ^{1}H, ^{13}C-edited NOESY-HSQC, and ^{13}C-aromatic NOESY-HSQC can be performed to obtain inter-molecular distance restraints for protein structure calculation. These distance restraints and Talos-derived dihedral angle restraints can be employed for ensemble structure calculation by utilizing computational tools like Cyana, Aria, and Xplore-NIH.

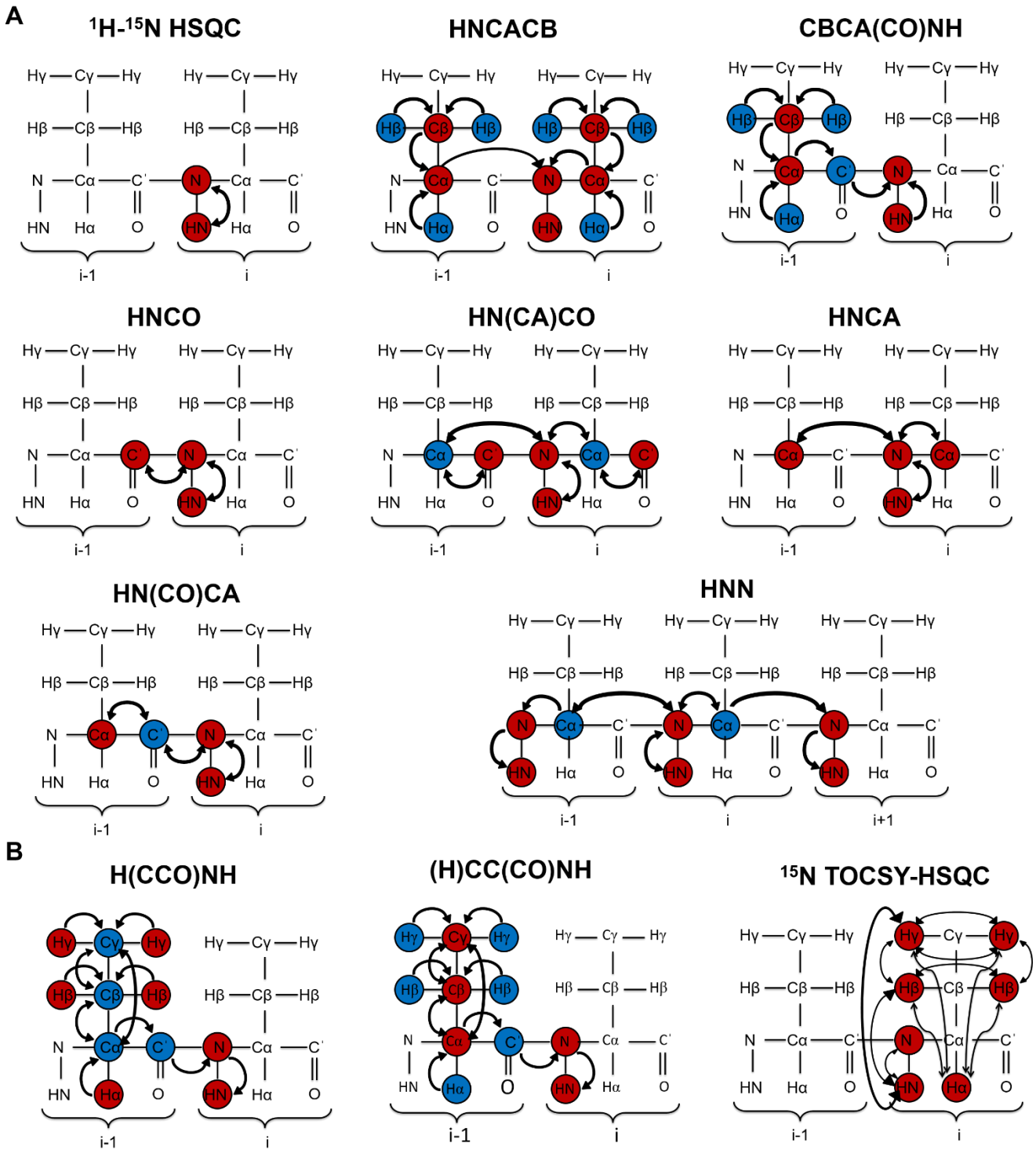


Figure 1.6. 2D and 3D NMR experiments for protein. Backbone (A) and side-chain (B) experiments are shown. The selected atoms transferring magnetization through a bond (J-coupling) from both the current residue (i) and the preceding residue (i-1) are indicated by an arrow. The detected and carrier nuclei are highlighted in red and yellow, respectively.

1.9 Protein-ligand interaction by NMR

Solution NMR is an effective method for investigating the interactions between proteins and ligands, as well as their structural and dynamic properties. The chemical shifts of functional groups in NMR are highly sensitive to changes in chemical environments, which can be used to study protein-ligand binding. NMR can detect a wide range of binding affinities, from weak (millimolar) to strong (nanomolar) binding affinities. By analyzing resonance changes in 1D or 2D spectra of a biomolecule when a ligand is added, various quantitative parameters can be obtained, such as (a) active or passive binding sites, (b) chemical exchange regime (association and dissociation rates), (c) dissociation constant, K_D and (d) single-step or multistep (allosteric) binding modes.

1.10 Chemical exchange between free and ligand-bound form

In a protein-ligand binding study, NMR titration experiments are conducted by recording multiple 1D (^{1}H or ^{13}C) or 2D ^{1}H-^{15}N (or ^{13}C) HSQC spectra of the isotopically labeled protein alone and after stepwise addition of non-labeled target ligand. The change in resonance frequency ($\Delta\omega$ in Hz) that occurs upon ligand binding depends on the chemical exchange rates (k_{ex}) between the free and ligand-bound complex. This exchange regime can be represented as, P + L $\leftrightarrow$ [PL] and $k_{ex} = k_2+k_1$; where P and L are the protein and ligand concentration, respectively, [PL] represents the protein-ligand complex, while k_1 and k_2 are the forward and reverse rate constants, respectively. Typically, Δω ranges from 10 to 10,000 sec^{-1}, while the timescale for chemical exchange (k_{ex}) is from 10 µsec to 100 msec, which is well-suited for the NMR timescale. Also, the timescale for chemical exchange can be represented by the equation as; $\tau = (\sqrt{2}\ \pi\ \Delta\omega)^{-1}$.

As shown in **Figure 1.7**, if the exchange rate $k_{ex} >> \Delta\omega$, the peak positions change in a progressive fashion with each titration step, known as the fast exchange regime. In contrast, when the exchange rate, $k_{ex} << \Delta\omega$, the peak position is populated only by the free and bound state in each NMR titration step, known as a slow exchange regime. When $k_{ex} \sim \Delta\omega$, an intermediate exchange regime is observed where the peak line-width becomes broad in a progressive manner and changes to a sharp line-width in a fully ligand-bound state during the titration. The timescale of the chemical exchange rate also depends on the strength of the magnetic field and measurement temperature. Decreasing magnetic field strength and (or)

increasing the temperature would alter the slow exchange to fast exchange timescale regime for protein-ligand binding. In summary, the resonance frequencies are well resolved for the interchanging species for slow timescale, while frequencies averaged out to one single line in the case of fast timescale (Waudby et al., 2016) (Hiroaki & Kohda, 2018) (Bryant, 1983) (Waudby et al., 2016) (Feng et al., 2019) (Becker et al., 2018).

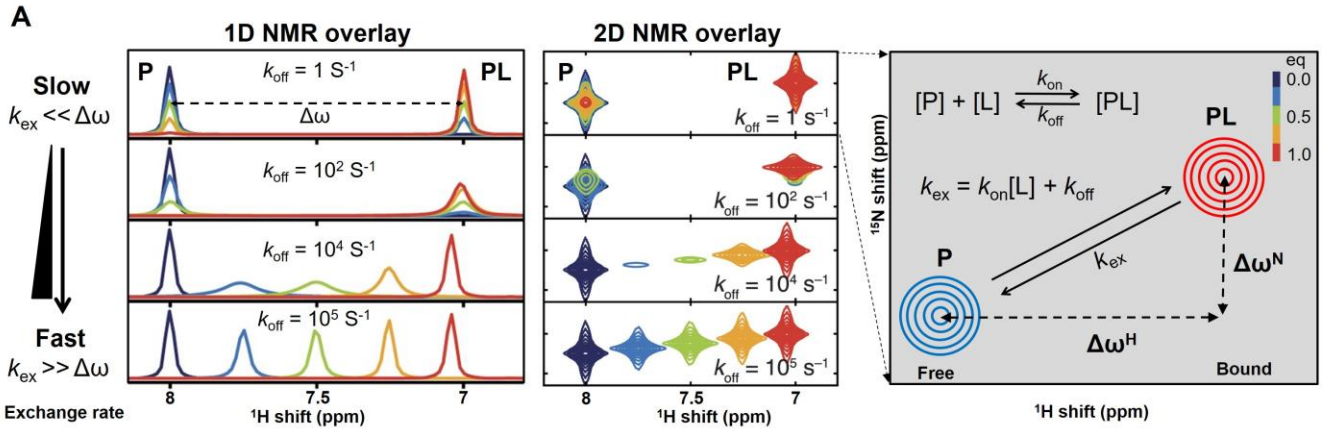


Figure 1.7. Solution NMR method for differentiating the chemical exchange regime for protein-ligand interactions. (A) Schematic of 1D (in left) and 2D (middle) NMR titration experiments are shown for the different exchange rates and frequency differences of protein in free (P) and ligand-bound (PL) states. (Figure is adapted with minor modification from Waudbly, Ramos et al. 2016).

1.11 Analysis of protein-ligand binding surface area

Overlaying the multiple 2D ^{1}H-^{15}N (or ^{1}H-^{13}C HSQC) spectra from NMR titration experiments of isotopically labeled protein in the absence and presence of increasing amounts of target ligand can give a clear idea of the exchange regime and protein-ligand interactions. For detailed analysis, the chemical shift perturbation (CSP) can be calculated between the free and bound forms of spectra using the formula as; $CSP = \left[(\Delta\delta\, ^1H_N)^2 + \left(\frac{\Delta\delta\, ^{15}N_H}{5}\right)^2 \right]^{\frac{1}{2}}$; where Δδ ^1H_N and Δδ $^{15}N_H$ annotate for amide proton and nitrogen shift differences between the free and bound form of ^{1}H-^{15}N HSQC spectra, respectively. The maximum CSPs above +2 of standard deviation are considered active (direct) binding sites, while other CSPs are considered passive binding sites. Those CSPs can be mapped on available structures to find the ligand interacting surface area. If binding is in the fast exchange regime, a dissociation constant (K_D) can be extracted from each of the resonances using a quadratic equation for protein-ligand complex as mentioned below:

$\delta_{obs} = \frac{\delta_{max}}{2P_0}\left[(P_0 + L_0 + K_D) - \{(P_0 + L_0 + K_D)^2 - 4P_0L_0\}^{\frac{1}{2}}\right]$; where δ_{max} and δ_{obs} are the maximum and observed chemical shifts (in ppm) upon addition of titrant, P_0, L_0, and K_D are total protein, total ligand, and dissociation constant. Using this equation, the dissociation constant can be extracted from NMR titration by assuming the single binding site with a 1:1 stoichiometry of protein to ligand ratio. Additionally, this fit function can be used for weaker affinity (micromolar to milimolar) of protein-ligand binding and is valid for a peak shift in a linear direction (**Figure 1.8**) (Aguirre et al., 2015) (Hiroaki & Kohda, 2018). However, more complex protein-ligand interactions are recognized for the non-linear peak shifts suggesting the allosteric or conformational change upon ligand binding (**Figure 1.8**).

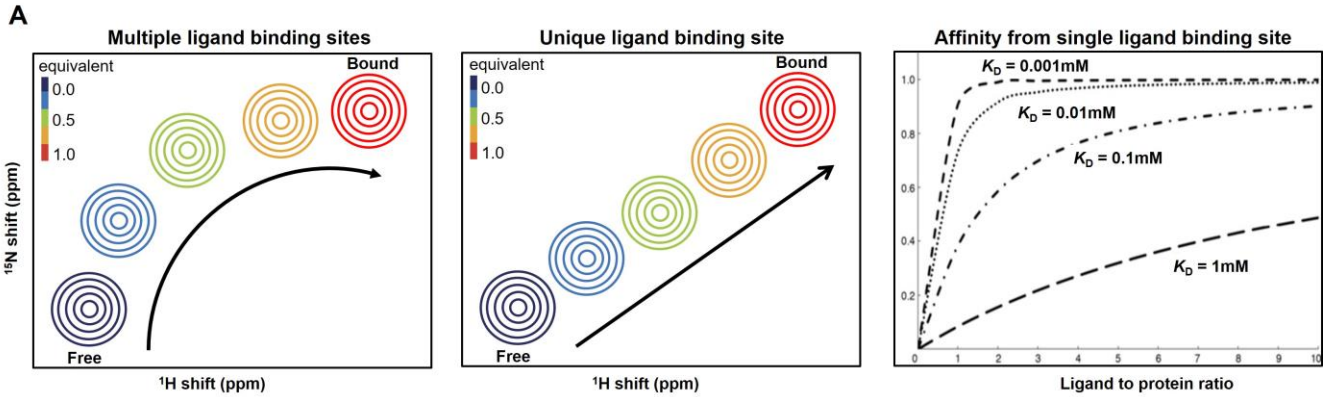


Figure 1.8. Schematic representation of NMR peak shifts based on single or multiple ligand binding sites and binding affinity. (A) The left figure shows the peak change in the direction upon the addition of the ligand showing the more complex binding event. (B) In the middle, the peak shift is shown in a linear direction while titrating the ligand that represents the single (active or passive) binding sites. (C) Right: simulated curves for dissociation constants were extracted for the different affinity of ligands (simulated K_D curves taken from Aguirre, Cala et al., 2015).

1.12 Protein backbone dynamics

Longitudinal spin relaxation

Solution NMR is one of the best techniques to study the dynamics of biomolecules at the residue level in solution. A wide range of dynamics and timescales can be studied by NMR, from fast motion with pico-second to very slow dynamics in the second timescale. NMR dynamics is described as an evolution of signal intensity over time after applying a radio frequency pulse. During this period, the bulk magnetization will move from the z-axis to the x-y transverse plane and return gradually to the initial equilibrium position along the longitudinal (z)

axis, called T_1 relaxation. T_1 relaxation is also known as spin-lattice relaxation and can be measured with an inversion recovery experiment, as shown in **Figure 1.9 A, B**. T_1 relaxation time gets shorter for any factor that slowdown the molecular motion, such as self-association, aggregation, solvent viscosity, ligand-bound substrate, etc.

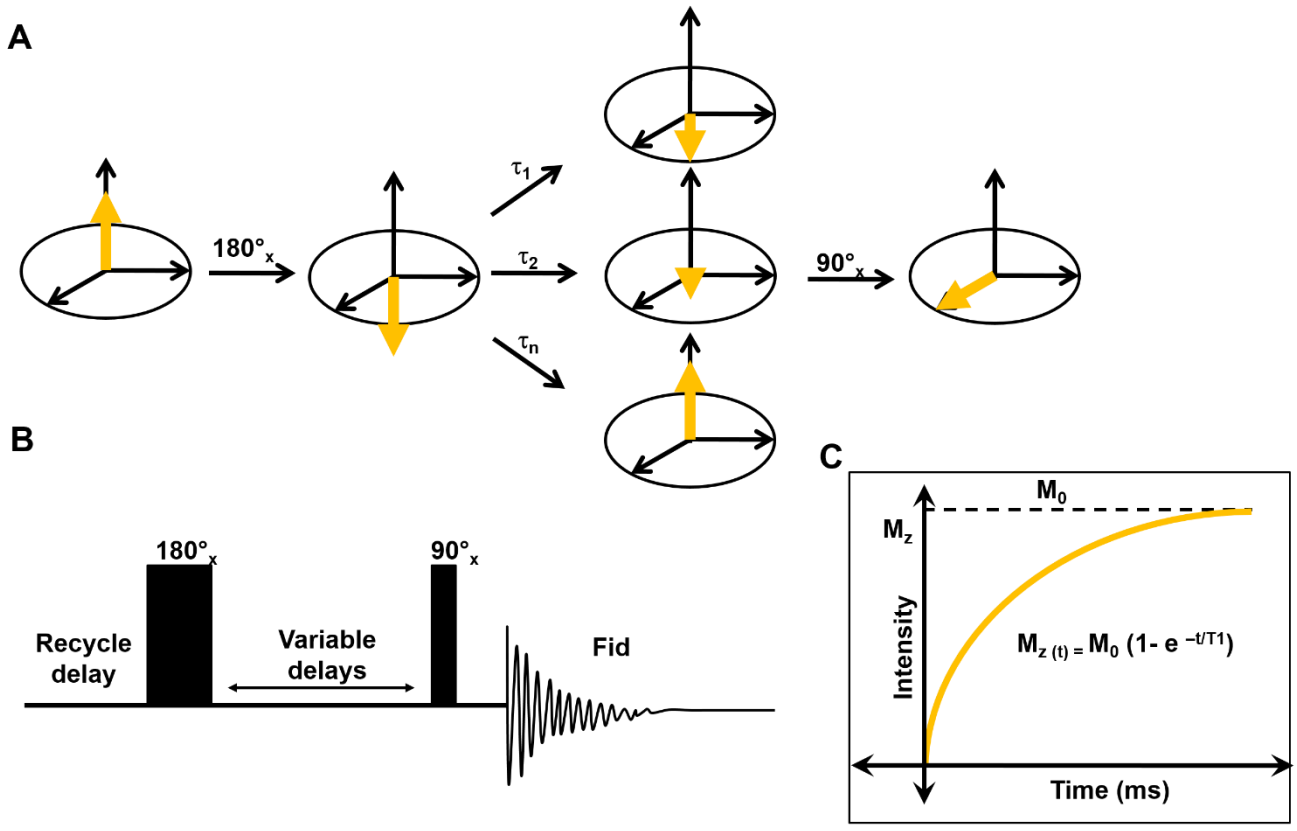


Figure 1.9. Schematic representation of spin-lattice relaxation (T_1). (A) After a 180º inversion pulse, bulk magnetization (yellow) exchanges energy with the surrounding to return to the equilibrium state (z-axis) (B) Graphical NMR pulse scheme for recording longitudinal relaxation time (T_1) and derive rates (R_1) using series of delays. (C) Representative plot for spin intensity (detection in the x-y plan) vs. delays shown as an exponential decay in single intensity in the x-y plane and intensity increases exponentially in the z-axis.

Spin transverse relaxation (spin-spin relaxation)

Besides T_1 relaxation, the second component also contributes known as spin-spin relaxation or T_2 relaxation. After applying an RF pulse, the bulk magnetization moves from the z-axis to the x,y plane, and precessing spins lose energy due to the loss of coherence in the x,y plane. Hence, it's also called transverse relaxation. T_2 is independent of the T_1 relaxation. Also, $T_1 \sim T_2$ is observed when small molecules tumble faster than Larmor frequency. T_2 relaxation gets shorter with increasing the molecule size due to dipole-dipole interactions

(**Figure 1.10**). In protein NMR, the frequently measured relaxation parameters are ^{1}H-^{15}N HSQC based T_1 longitudinal and T_2 transverse relaxation time for backbone amide together with {^{1}H}-^{15}N NOEs (Palmer, 2004) (Jaremko et al., 2018) (Reddy & Rainey, 2010) .

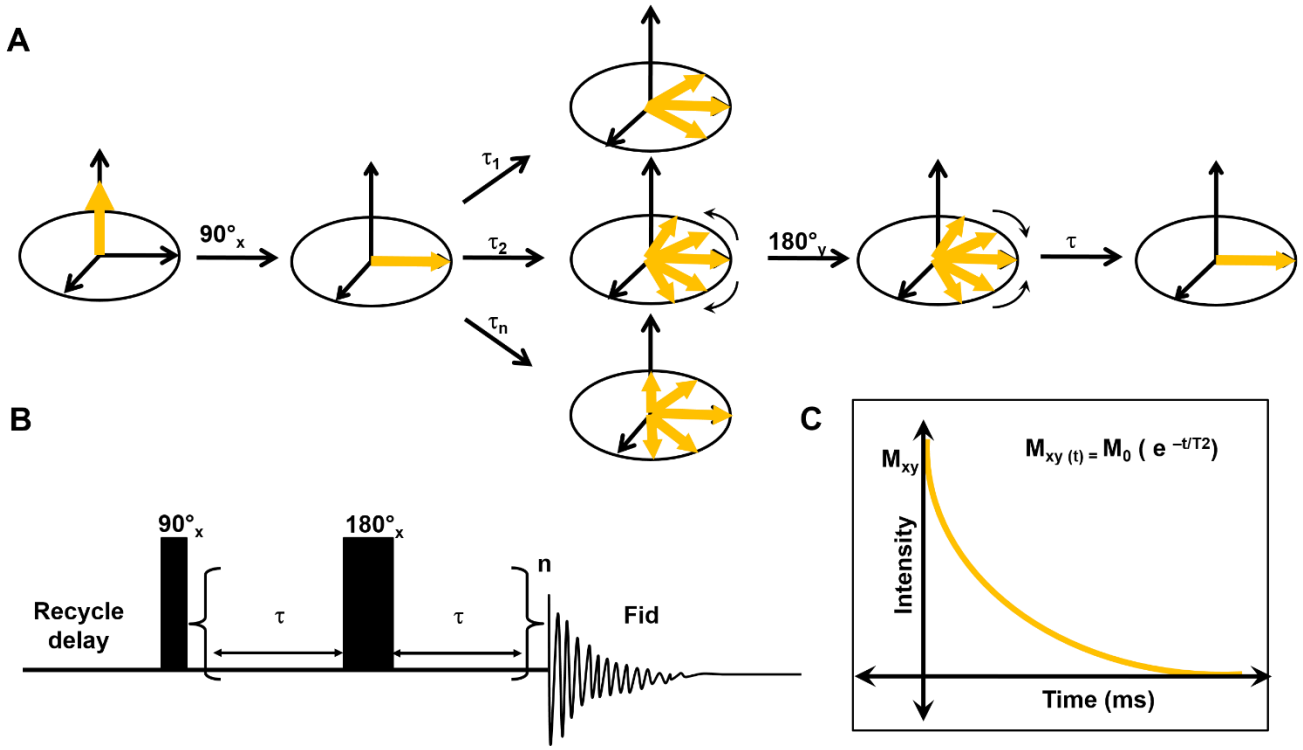


Figure 1.10. Proton transverse relaxation measurement using NMR. (A) Schematic representation of coherence loss over time after the first 90^{o} pulse. (B) NMR pulse scheme to measure CPMG based transverse relaxation rates. (C) Plot for signal intensity (in the x,y plane) vs. delays shown as an exponential decay.

1.13 Paramagnetic relaxation enhancement NMR method

Paramagnetic relaxation enhancement (PRE) is one of the well-explored solution NMR techniques for biomolecules. PRE-based NMR can provide various types of information: (a) long-range distance restraints to derive multi-domain arrangements (b) protein dynamics (c) protein-protein, protein-membrane, and protein-ligand interactions (d) detection of lowly populated species. The PRE effect is mainly due to the magnetic dipolar interactions between the NMR active spins and the unpaired electron of a paramagnetic moiety. The electron's gyromagnetic ratio is much larger (~ 658 times) than that of the proton. This results in the relaxation rate enhancement of nuclear magnetization, and it is proportional to an average of "r^{-6}" distance between the unpaired electron and NMR-active nucleus, allowing PREs to provide long-range distance information up to the ~20-25 Å protons away from the unpaired electron

(**Figure 1.11**). In contrast, the proton-proton distances obtained from the traditional nuclear Overhauser effect (NOE) method can extend only up to ~5 Å (Olivieri et al., 2018).

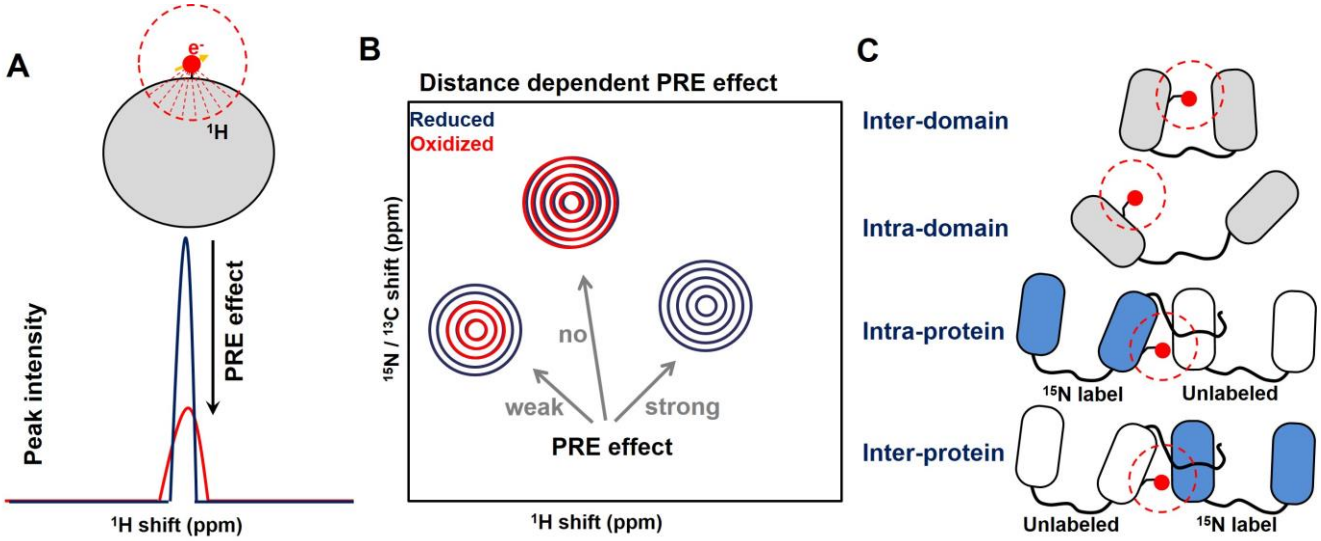


Figure 1.11. Principles and analysis of paramagnetic relaxation enhancement (PRE) experiment. (A) Schematic 1D overlay spectra are shown for the reduced and oxidized state of the sample. (B) Schematic illustration of strong, moderate and weak PRE effect is shown with ^{1}H-^{15}N/^{13}C HSQC spectra. (C) Different types of PRE experiments are shown: intra- and inter-domain PRE for protein along and/or with protein-protein complex using differential labeling scheme).

1.14 Spin-labeling for PRE measurements

To conduct PRE experiments, it is necessary to introduce a single paramagnetic spin label (SL) onto the biomolecule (i.e., protein) surface. For that, cysteine is an ideal target for SL in proteins due to its reactive thiol (sulfhydryl "SH") group, which can be easily introduced via point mutation. As shown in (**Figure 1.12**), the most commonly studied SLs for this purpose are nitroxide spin-labels like MTSL ((1-Oxyl-2,2,5,5-tetramethyl-3-pyrroline-3-methyl) methanethiosulfonate), IPSL (3-(2-Iodoacetamido)-2,2,5,5-tetramethyl-1-pyrrolidinyloxy), MPSL (3-Maleimido-2,2,5,5-tetramethyl-1-pyrrolidinyloxy), IDSL (bis-(2,2,5,5-Tetramethyl-3-imidazoline-1-oxyl-4-yl)disulfide; Noxygen), etc. To ensure efficient labeling, a 3-fold excess concentration of nitroxide SL should be added to the protein sample at pH 8 overnight at 4 °C. Reducing agents should be avoided in the reaction buffer as they can interfere with labeling reactions. The success of labeling and labeling efficiency can be confirmed by mass spectrometry (Hennig et al., 2015) (Klare & Steinhoff, 2009) (Ackermann et al., 2021).

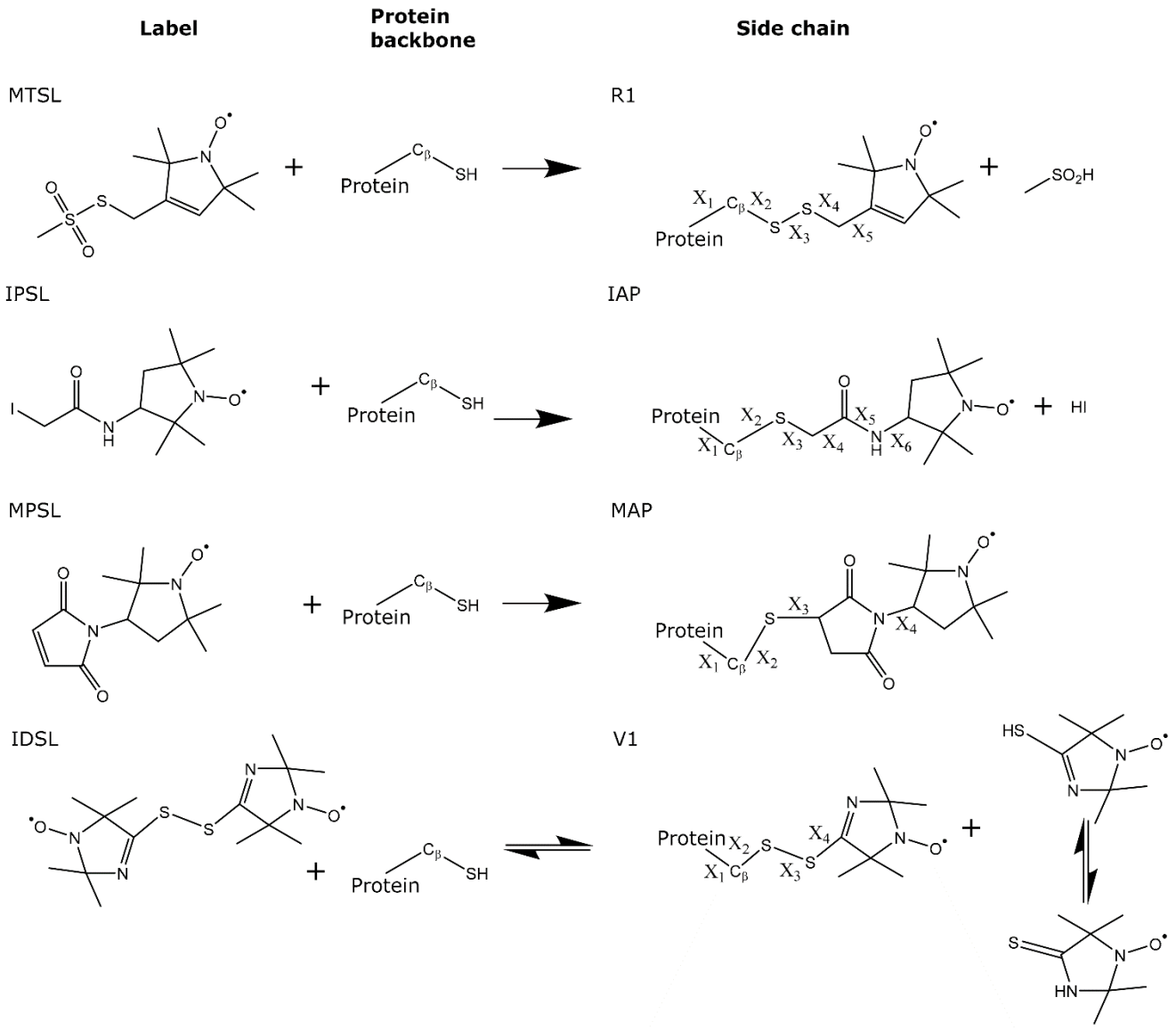


Figure 1.12. Type of nitroxide PRE tags used in NMR. A figure displays the most common nitroxide spin labels MTSL, IPSL, MPSL and IDSL available with cysteine conjugates for PRE experiments. X represents rotatable bonds.(Figure is taken from Ackermann, Chapman et al. 2021).

1.15 PRE measurements and structure refinement

The PRE can be measured using 2D ^{1}H-^{15}N/^{13}C HSQC and/or with ^{15}N-R_2 rates for the paramagnetic and diamagnetic states of the samples. For quantitative analysis of PREs, the intensity ratio and/or ^{15}N transverse relaxation rates (R_2) are derived between the paramagnetic and the diamagnetic states of the sample. The qualitative analysis of PRE is useful for studying protein-protein, protein-membrane interactions, as well as transient interactions in intrinsically

disordered proteins (IDPs). PRE-derived distance restraints can aid in protein structure calculation using a rigid-body refinement approach. The intensity ratio of para- and diamagnetic states of the sample can be used to calculate long-range intra- and inter-molecular distance restraints between unpaired electrons from SL and proton spin with the equation provided below:

$$\text{Intensity ratio, } \frac{I_{oxi}}{I_{red}} = \frac{R_2 . e^{-R_2{}^{sp} \cdot t}}{R_2 + R_2{}^{sp}} \quad \text{(eq. 1)}$$

R_2^{sp} can be derived using Solomon-Bloembergen equation mensioned below:

$$\text{Spin Label, } R_2^{sp} = \left(\frac{1}{15}\right)\left(\frac{\mu_0}{4\pi}\right)^2 \gamma_H^2 \, g^2 \, \mu_B^2 \, S(S+1) r^{-6} \left(4\tau_c + \frac{3\tau_c}{1+(\omega_H \tau_c)^2}\right) \quad \text{(eq. 2)}$$

$$\text{Distance, } r = \left[\frac{K}{R_2^{sp}}\left(4\tau_c + \frac{3\tau_c}{1+(\omega_H \tau_c)^2}\right)\right]^{\frac{1}{6}} \quad \text{(eq. 3)}$$

$$\text{Correlation time, } \tau_c = \frac{1}{\tau_r} + \frac{1}{\tau_s} + \frac{1}{\tau_m} \quad \text{(eq. 4)}$$

where, R_2 and R_2^{sp} are transverse relaxation rates for diamagnetic and paramagnetic sample, K is a constant (1.23 x 10^{-32} cm^6 sec^{-2}) for nitroxide spin labels (MTSL or IPSL), ω_H is proton Larmor frequency, γ_H is the proton gyromagnetic ratio, g is the electron g-factor, S is a spin quantum number, μ_0 is the permeability of a vacuum, μ_B is the magnetic moment of the free electron, τ_c is an apparent PRE correlation time. Also, τ_r is the rotational correlation time for paramagnetic protein, τ_s is the electron spin relaxation time, and τ_m is the lifetime of the complex. In equations, proton Larmor frequency(ω_H) and PRE correlation time (τ_c) can be derived from external magnetic field strength and molecular weight of the protein, respectively. Based on the above equations, the simulated distance calibration curve can be calculated (**Figure 1.13**). The nitroxide spin labels (SL), such as IPSL or MTSL are effective within a radius of 13 to 25 Å. This means an intensity ratio (I_{ox}/I_{red}) between 0 to 1 indicates that the spin label and proton spin are separated by <10 Å and >30 Å, respectively (Iwahara et al., 2004) (Softley et al., 2020) (Sjodt & Clubb, 2017).

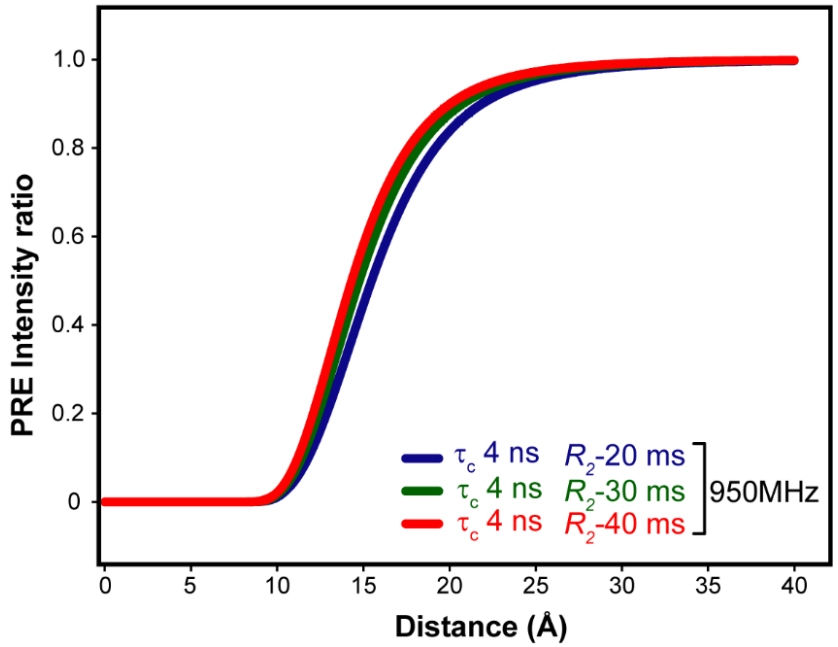


Figure 1.13. Calibration curves plotted for the intensity ratio with estimation of PRE distances. The curves were generated as per Equations 1 and 2 using an average R_2 rate and correlation time (τ_c) for SF1 protein, respectively.

While using Nitroxide SL, it's important to consider that it has rotatable bonds which provide significant flexibility and thus, PRE-derived distances should consider having +/-5 Å errors in estimation for each extracted distance (**Figure 1.14**). This error value may increase depending on the local flexibility of the protein. To minimize error, selecting the SL site on the rigid part of the protein is best. Based on derived distances, the protein structural models can be derived by rigid-body refinement protocol using computation tools such as CNS, Aria, Xplore-NIH, MMMx etc. Depending on the protein size or protein-complex, multiple SL positions may be required to generate good quality structural models with enough distance restraints. To assess the refined structures, a Q-factor (quality factor) can be used to measure the agreement between the refined structural models and the experimental PRE datasets. (Klare & Steinhoff, 2009) (Clore, 2011) (Barnes et al., 2021).

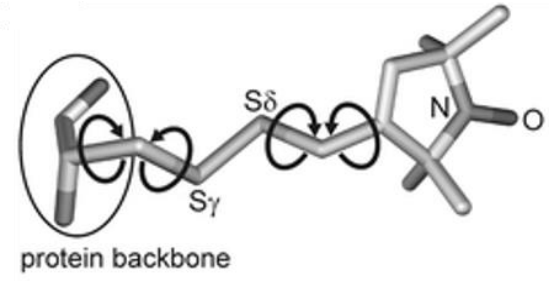


Figure 1.14. Rotatable bonds of IPSL nitroxide spine label attached to cysteine residue are shown with arrows (in right).

1.16 Small-angle X-ray scattering

Small-angle X-ray scattering (SAXS) is a well-established technique used to characterize the structure of biomolecules, such as proteins, RNA, and nano-discs, at lower resolutions. It is a solution-based method that does not require crystallization or vitrification procedures like crystallography. In SAXS, a monochromatic X-ray beam is passed through the sample solution, and the radiation scattered at low angles is measured by a detector, which provides the scattering vector "*q*" and scattering intensity "*I*" (**Figure 1.15**). The interaction between x-ray photons with a wavelength of "*λ*" and the electron density in the sample generates constructive interference patterns along certain angles, based on Bragg's law of scattering: $n\lambda = 2d \sin(\theta)$ and the scattering vector, $q = (4\pi \sin \theta) / \lambda$, where "n" is an integer, "d" is the displacement between reflection sites, and "θ" is the scattering angle. The inhomogeneity in the electron density observed at lower (<10°) angles provides details on the size and shape of the macromolecules. However, since scattering patterns are a combination of macromolecules and surrounding buffer components, background subtraction is required to remove unwanted scattering from the buffer. The final processed SAXS curve yields intensity from the macromolecules as a function of the scattering angle (**Figure 1.15**) (Brosey & Tainer, 2019) (Fullmer, 2015) chapter 7).

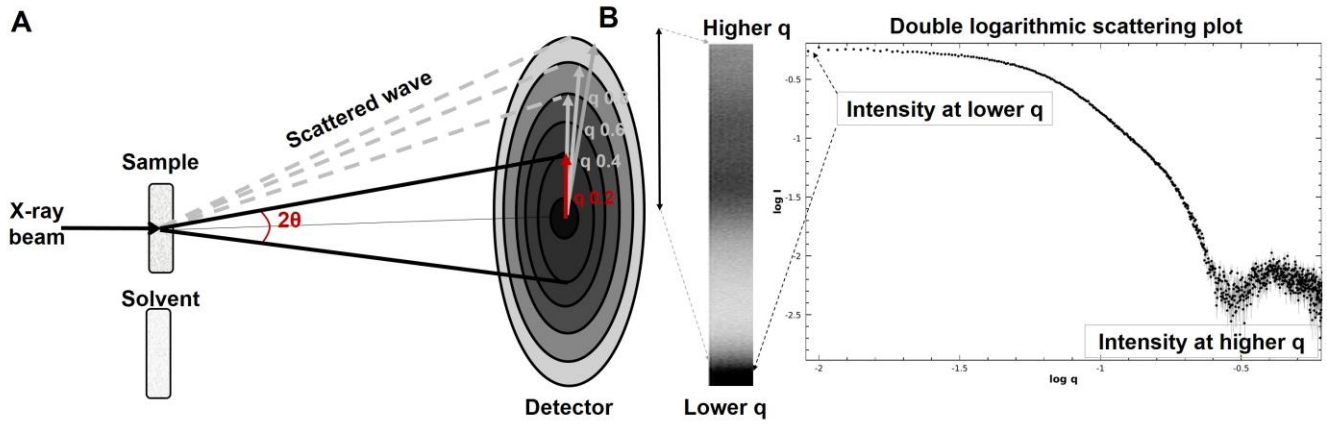


Figure 1.15. Basic Small Angle X-ray Scattering (SAXS) instrument setup and data processing. (A) A monochromatic X-ray beam passes through the aqueous sample and the X-rays scattered at low angles (θ) are collected by a detector (B) Double logarithmic scattering plot is shown after background subtraction for Npl3 protein.

1.17 Practical applications of SAXS

The experimental SAXS curve provides valuable information to extract biophysical

parameters such as radius of gyration (Rg), overall shape, size (Dmax), and molecular mass. These parameters are obtained from the Guinier plot (***log I vs. q^2***), pair distance distribution function (***Pr vs. r***), and Kratky plot (***q^2I vs. q***). R_g is proportional to the size of molecules, particle shape, size distribution, and interactions between particles. D_{max} is determined by transforming reciprocal space data to real space, which gives a geometrical representation of the scattering species as a histogram of ***P(r) vs r*** (radius) plot. ***P(r)*** goes to zero at the maximum diameter of the particle, providing structural information for protein and RNA, such as shape, compactness, molecular flexibility, folding, and unfolding states of the protein (**Figure 1.16**). Additionally, SAXS data can also complement other techniques such as solution NMR, X-ray crystallography, EPR, FRET, and molecular dynamics (MD) simulation to understand the complex structure behavior of macromolecules in solution (Fullmer, 2015); chapter 7) (Dmitri & Michel, 2003).

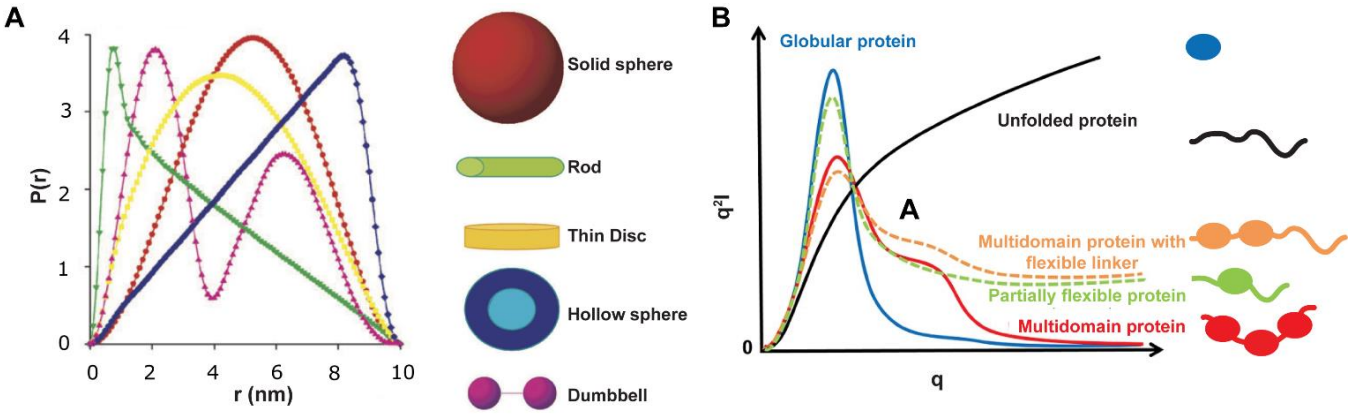


Figure 1.16. SAXS analysis of biomolecules. (A) Pair distance distribution function "*P(r)* vs *r*" plots are shown for different geometric shapes of biomolecules (Figure is taken from reference Dimitri and Michel, 2003). (B) A schematic representation of Kratky plots are shown for different protein system (figure is adapted from https://www-ssrl.slac.stanford.edu/~saxs/analysis/assessment.htm).

A few open-source software packages are available such as ATSAS, SCATTER, etc., to analyze macromolecules' R_g, D_{max}, *MW*, and shape from experimental SAXS data. The ATSAS package includes DAMMIF and DAMMIN tools for generating ab initio models from SAXS curves. Additionally, CRYSOL can generate theoretical SAXS curves from available high-resolution crystal or NMR structures and match them with experimental data. For rigid body modeling and refinement from known high-resolution structures, CORAL and SREFLEX are

useful tools. The ATSAS package also offers the ensemble optimization method (EOM) for providing a detailed analysis of the ensemble representation of flexible proteins like IDP. Furthermore, experimental SAXS data can be used to compare and screen structural ensemble models generated from other methods, such as NMR, EPR, and MD simulation (Kikhney & Svergun, 2015) (Da Vela & Svergun, 2020) (Brosey & Tainer, 2019).

1.18 Protocols for applied methods

To gain a comprehensive understanding of how RBPs and RNA recognitions are regulated, a range of biophysical methods including NMR binding, PRE, T_1/T_2 fast relaxation, SAXS, and ITC were utilized. In this study, these methods were applied to investigate the pre-mRNA splicing and nuclear mRNP assembly. The protocols for these methodologies are mentioned in Chapter 2.

Aim and scope of the thesis

RNA-binding proteins (RBPs) are essential in regulating post-transcriptional gene expression. They can act as splicing enhancers or repressors and work with other proteins to maintain cellular homeostasis during critical processes such as transcription, pre-mRNA processing, splicing, nuclear mRNA export, cytoplasmic localization, translation, and degradation. Chapter 1 of the thesis provides a comprehensive overview of the intricate architecture of multidomain RBPs, including various types of RNA binding domains and the role of flexible linkers in RNA regulation. This thesis focuses on studying the dynamic regulation of multidomain RBPs involved in **pre-mRNA splicing** and **nuclear mRNPs assembly** using various biophysical techniques such as NMR, SAXS, and ITC.

(a) RBP regulation in pre-mRNA splicing

RNA binding proteins (RBPs) determine the fate of pre-mRNA splicing for both constitutively and alternatively. During the early stages of spliceosome assembly, two proteins SF1 and U2AF heterodimer, are the first target proteins to recognize the intron splice sites. While there have been various cellular and biochemical studies on intron splice site recognition, there is still a need to understand the dynamic recognition and regulation of splice sites by splicing factors at atomic details. Despite having high-resolution structures of individual domains of SF1 and U2AF2, there is a lack of structural understanding of splice site recognition by SF1-U2AF2 complex in the context of variable strength, distances, and multiple intron splice sites. Additionally, further investigation is needed to understand the mode of interactions, structural details, and underlying molecular mechanism of SF1 in both canonical and disease-associated splice sites, and the role of SF1 phosphorylation. To address these gaps, the first aim of the project is to investigate the underlying molecular mechanism of how human SF1-U2AF2 complex recognizes variable 3' splice sites in the early stage of splice sites. The work presented in Chapter 3 will explore the structural features of SF1 and its recognition of canonical and disease-associated branch point sites of introns. The study will expand upon existing knowledge of subdomains by analyzing the structural features of the whole SF1-U2AF2 complex. Additionally, it will investigate the dynamic regulations of the SF1-U2AF2 complex, the crucial role of their inter-domain flexibility in recognizing the diverse strength of BPS and PPT splice sites, and the subsequent impact of this process on splicing regulation in humans.

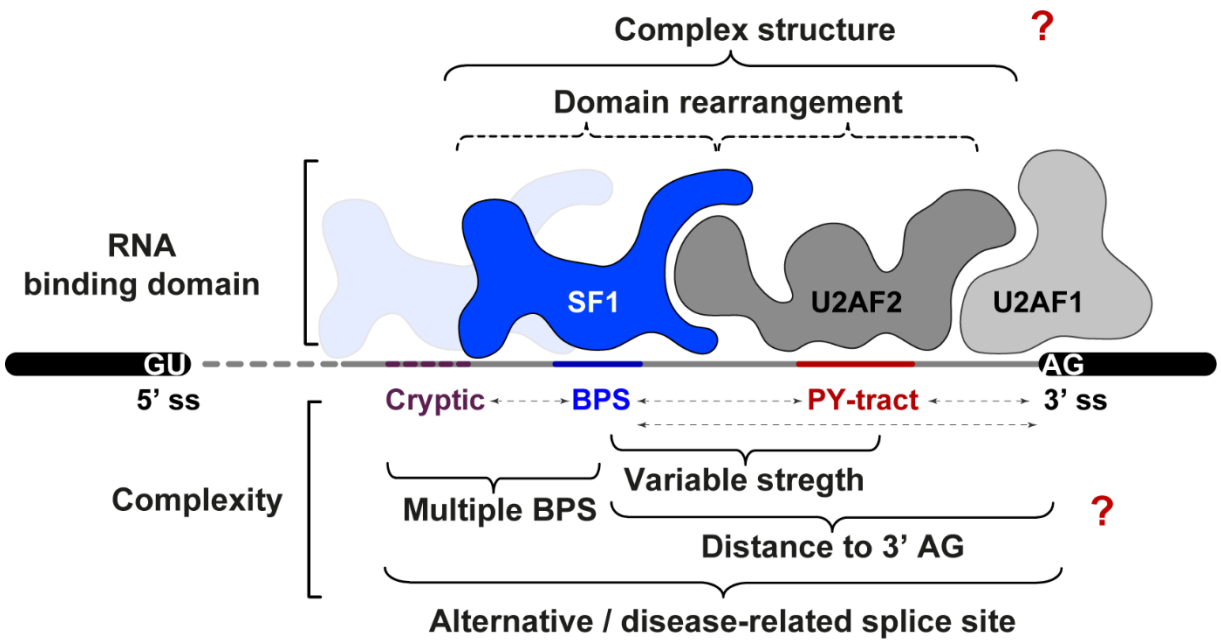


Figure 1.17. The branch point site, the polypyrimidine tract, and the 3' "AG" splice sites RNA recognition by splicing factors SF1, U2AF2, and U2AF1.

(b) RBP regulation in nuclear mRNP assembly

Besides splicing, RBPs are essential for mRNA stability and mRNP assembly through direct protein-RNA interactions. In *Saccharomyces cerevisiae*, Npl3, Yra1, Nab2 and Mex67-Mtr2 with other proteins are responsible for recruiting nascent mRNA and exporting to the nuclear pore complex. Amongst these, Npl3 is a multidomain SR-like protein that contributes to transcription, splicing, mRNP assembly, mRNA processing, and nuclear mRNA export. It has two RNA recognition motifs (RRMs) that facilitate RNA recognition. The individual structure of RRMs in RNA-free form has been reported previously. However, the molecular functions of each RRM, RNA specificity, mode of interactions, and structure of tandem RRMs for RNA recognition are not well characterized. Therefore, the second aim of the thesis is to investigate the structure, dynamics, and functional role of Npl3 in nuclear mRNP components, with a focus on understanding the recruitment of Npl3 in nuclear mRNP assembly and nuclear export in yeast. Chapter 4 will present a study of Npl3's structure and function, highlighting its molecular functions both *in vitro* and *in vivo*. This chapter will also provide insight into the atomistic details of Npl3 RRMs, their specificity of RNA recognition, the structural analysis of Npl3 mutants, and their respective roles in nuclear mRNP assembly.

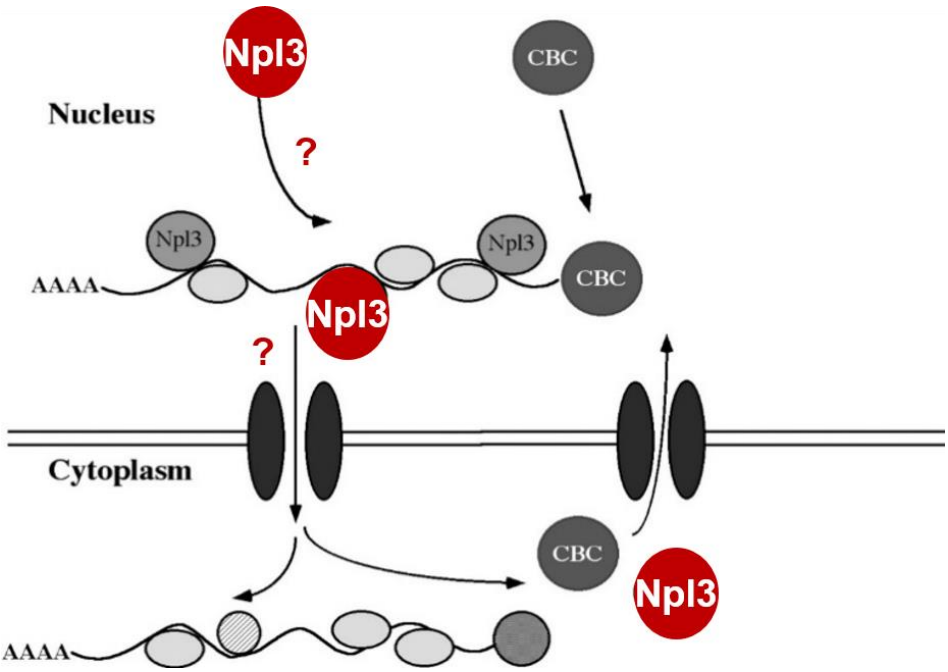


Figure 1.18. Role of Npl3 in nuclear mRNA recognition and mRNP assembly.

Chapter 2 – Materials and Methods

Materials

2.1 Chemicals and reagents

Acrylamide/Bis, Serva
Ammonium persulfate, Merck
Ammonium chloride-^{15}N,SIGMA-ALDRICH
Biotin, SERVA
β-Mercaptoethanol, ROTH
Glucose-^{13}C, Cambridge Isotope Laboratories
Calcium chloride, SIGMA-ALDRICH
Cobalt chloride, Merck
Cupper chloride, Merck
Coomassie Stain, SERVA
D2O 99%, Sigma-Aldrich and Silantes
D(+)-Glucose, SIGMA-ALDRICH
di-Sodium hydrogen phosphate, SIGMA--ALDRICH
Dodecylsulfate-Na-salt pellets, SERVA
DTT, SERVA and SIGMA-ALDRICH
Ethanol, VWR
Hydrochloric acid, SIGMA-ALDRICH
Iron chloride, VWR
Glycerole, ROTH
Glycine, ROTH
Imidazole, ROTH
IPSL, SIGMA-ALDRICH
IPTG, SERVA
Kanamycin, SERVA
Lysozyme, SERVA
Magnesium chloride, VWR Chemicals
MES, SIGMA-ALDRICH
Ammonium chloride ^{15}N, Cambridge Isotope Laboratories
Ni-NTA, Merck
Prestained SDS PAGE Protein Marker 6.5 - 200 kDa, SERVA
Protease Inhibitor Mix HP, SERVA
Sodium Chloride, ROTH
Sodium dihydrogen phosphate dehydrate, SIGMA-ALDRICH
Sodium hydroxide, Merck
Sodium perchlorate, SIGMA-ALDRICH
TEMED, Sigma-Aldrich
TEV protease, self-prepared in lab
Thiamine, SERVA
Tris/HCl, ROTH
Urea, SIGMA-ALDRICH
Yeast extract, SIGMA-ALDRICH
Zinc chloride, SIGMA-ALDRIC

2.2 Regular disposable materials

Amicon Ultra Centrifugal Filters 3k, 10k cut-off with 0.5 ml, 5 ml, 15 ml, MERCK
Cellulose filter 0.20 µm, 0.45 µm pore size, Sartorius and Cytiva

Sterican Disposable inoculation loop, VWR
Eppendorf tubes 1.5 ml, 2 ml, 15 ml, 50 ml
Gravity Flow Columns, BIO-RED
Hand gloves, Starlab
Cuvette (polystyrene), Bio-Rad
Inject syringes 1 ml, 3 ml, 5 ml
NAP-5 columns, Cytiva
NMR tubes 5mm and 3mm, Duran
NMR Shigemi tubes, SHIGEMI Co., LTD
Paper towel
Pasteur pipettes,
PD-10 columns, GE
Pipette tips 20 µl, 200 µl, 1000 µl, Starlab
Pipette boy, Starlab

2.3 Laboratory devices

Autoclave, Systec VX-150
AKTA Pure and AKTA Go protein purification system, GE
BioSAXS (in-house) Rigaku instrument
Bradford reagent (5x), Bio-Rad
Bruker NMR spectrometers 500, 3x600, 800, 900, 950, 1200 MHz
Centrifuge SORVALL LYNX 600, Thermo -Scientific
Centrifuge Mega star 1.6R with swinging -bucket, VWR
Centrifuge (Cooling table-top) 5424R, -eppendorf
Eppendorf tubes, tube stands; pipettes
Flasks 2L, 5L; beakers, VWR
Freezer -80 C, U725 Innova
Fridge 4 ^{o}C & -20 ^{o}C storage
French press cell lysis, DIGI
GELDOC XR, Biored
Ice machine, Ziegra Eismaschinen
Incubator Incu-Line, VWR Internationals
Isothermal titration calorimetry, Malvern -Panalytical
NanoDrop ND-1000 Spectrophotometer, -PEQlab-Biotechnology GmBh
Magnetic stirrer IKAMAG REO S6, IKA
Measuring cylinders, VWR
Microwave NN-E202CB, Panasonic
NMR tube 5mm, 3mm size
pH meter InLab routine
Resource Q and S, 6ml columns, Cytiva
Shaker, New Brunswick Scientific
SmartSpec Plus spectrophotometer, Bio-Rad
Sonicator cell lysis,
Superdex 75 HiLoad 16/600 column, GE
Superdex 200 Increase 10/300 GL, Cytiva
Table top centrifuge 5810R, eppendorf
Thermomixer confort, eppendorf
Thermocycler MyCycler, BioRad
Transilluminator for DNA/protein,blueBox S
Transilluminator UV 254nm, Bentchtop
Vacuum pump, MZ-2C-NT Vacuubrand
Vortex-Mixer, VWR
Water bath, VWR

2.4 Cell strains

Escherichia coli BL21 (DE3) *Escherichia coli* DH10β

Escherichia coli DH5α *Escherichia coli* Mach 1

- All cell strains were provided by Dr. Arie Geerlof, from Protein Expression and Purification Facility (PEPF), Institute of Structural Biology (STB), Helmholtz Zentrum München.

2.5 Bacterial expression vectors

pETM-10 pETM-11 pET-24d pET-Trx_1a pET-GB1_1a

- All expression vectors were provided by Dr. Arie Geerlof, from Protein Expression and Purification Facility (PEPF), Institute of Structural Biology (STB), Helmholtz Zentrum München.

2.6 Program and web servers

Software	**Cites**
Adobe illustrator CS6	www.adobe.com/products/illustrator
ATSAS v3.0.5	https://www.embl-hamburg.de/biosaxs/software.html
CCPN v2.5	https://ccpn.ac.uk/
CNS v1.2	https://www.mrc-lmb.cam.ac.uk/public/xtal/doc/cns/cns_1.2/
Expasy protparam	https://web.expasy.org/protparam/
ESPript 3.0	https://espript.ibcp.fr/ESPript/ESPript/
Expasy translate	https://web.expasy.org/translate/
Feedly	https://feedly.com/
Gene Runner	http://www.generunner.net/
Gnuplot v5.4	http://www.gnuplot.info/
Haddock v2.2	https://www.bonvinlab.org/software/haddock2.2/
MicroCal PEAQ-ITC	https://www.malvernpanalytical.com/en/support/product-support/microcal-range/microcal-itc-range/microcal-peaq-itc
JalView v2.11.2.0	https://www.jalview.org/
KEGG	https://www.genome.jp/kegg/kegg2.html
Muscles	https://www.ebi.ac.uk/Tools/msa/muscle/
NMRPipe	https://www.ibbr.umd.edu/nmrpipe/index.html

onlyoffice	https://www.onlyoffice.com/desktop.aspx
PDB	https://www.rcsb.org/
PDBePISA	https://www.ebi.ac.uk/pdbe/pisa/
Procheck-NMR	https://www.ebi.ac.uk/thornton-srv/software/PROCHECK/
Pymol v1.8.6.0 & v2.5.2	https://pymol.org/2/
SBGrid	https://sbgrid.org/
SnapGene v6.2	https://www.snapgene.com/
Topspin v3.5pl6	https://www.bruker.com/en/products-and-solutions/mr/nmr-software/topspin.html
Uniprot	https://www.uniprot.org/
Zotero v6.0.23	https://www.zotero.org/

Buffers, stocks and protocols

2.7 Bradford assay

Estimate unknown protein concentration by Bradford assay (1ml reaction)			
BSA standard concentration series		**Volume of buffer**	**Volume from 5x Bradford reagent**
BSA concentration (mg/ml)	**Volume of BSA (stock: 10mg/ml)**		
Blank	0 µl	800 µl	200 µl
0.01	1 µl	799 µl	200 µl
0.02	2 µl	798 µl	200 µl
0.03	3 µl	797 µl	200 µl
0.04	4 µl	796 µl	200 µl
0.05	5 µl	795 µl	200 µl
Unknown concentration	**Volume of sample**		
Protein sample S1	1 µl	799 µl	200 µl
Protein sample S2	2 µl	798 µl	200 µl
Protein sample S3	3 µl	797 µl	200 µl
Protein sample S4	4 µl	796 µl	200 µl
Mix well, incubate for 10 minutes and measure the O.D. at 595 nm			
Plot the graph with BSA concentration versus O.D. and use linear fit to derive the slope. Estimate the protein concentration from slop and the OD of the unknown protein concentration.			

2.8 Competent cell preparation

-Streak *E. coli* DH5a or *E. coli* BL21 strain on an LB plate and allow them to grow at 37^{o}C overnight. Inoculate single colony in 10 ml of LB media and grow overnight at 37^{o}C

-Transfer 5 ml of overnight grown culture into 500 ml of LB media and allow cells to grow at 37^{o}C until OD600 reaches 0.4 (~2-3 hours).

-Place culture flask on ice for 30 mins. Cells must remain cold for the rest of the procedure. Harvest the cells using a centrifuge at 3,000 g for 15 min.

-Resuspend the cells in 50 ml of cold sterile 0.1 M CaCl2 solution and transfer into 50 ml falcon tube. Incubate on ice for 30 mins

-Centrifuge cells at 4^{o}C for 15 mins at 3,000 g (~2500 rpm) and remove the supernatant.

-Resuspend the cells (by slow pipetting and cutting the tip) in 5 ml cold 0.1M CaCl2 containing 15% glycerol.

-Transfer 100µl volume into sterile ice-cold 1.5 ml Eppendorff tubes. Freeze the cells in liquid nitrogen and store them at -80^{o}C for up to six months.

NOTE: through the process, cells should be treated with care. No vortexing or excess pipetting should be performed, especially when the cells are treated in $CaCl_2$ solution.

2.9 M9 minimal medium

For 1 liter M9 mineral medium, add to 867 ml sterile water and below volume of stocks		**Components**	**Final concentration per liter**
Volume	**Stocks**		
50 ml	M9 salt solution (20X)	Na_2HPO4 KH_2PO4 NaCl NH_4CL (^{15}N)	33.7 mM 22.0 mM 8.55 mM 9.35 mM
20 ml	20% Glucose	Glucose ($^{13}C/^{12}C$)	0.4 %
1 ml	1 M $MgCl_2$	$MgCl_2$	1 mM
0.3 ml	1 M $CaCl_2$	$CaCl_2$	0.3 mM
1 ml	Biotin (1 mg/ml)	Biotin	1 µg
1 ml	Thiamin (1 mg/ml)	Thiamin	1 µg
10 ml	Trace elements solution (100 x)	Trace elements	1 x

2.10 M9 stock solutions

<table>
<tr><th colspan="2">M9 Stock preparation</th></tr>
<tr><td>M9 salt solution (20x)
-Na_2HPO_4-$2H_2O$: 150.4 gm/L
-KH_2PO_4 : 60 gm/L
-NaCl : 10 gm/L
-NH_4Cl : 10 gm/L
Dissolve the salts in 800 ml water and adjust the pH to 7.2 with NaOH. Add water to a final volume of 1 L and autoclave for 15 min at 121°C.
———————————————————
20% Glucose
For 500 ml stock solution, add 100 gm glucose to 440 ml water. Sterilize the solution over a 0.22-µm filter.</td><td>100X trace elements solution
-EDTA : 5 gm/L (13.4 mM)
-$FeCl_3$-6H2O : 0.83 gm/L (3.1 mM)
-$ZnCl_2$: 84 mg/L (0.62 mM)
-$CuCl_2$-$2H_2O$: 13 mg/L (76 µM)
-$CoCl_2$-$2H_2O$: 10 mg/L (42 µM)
-H_3BO^3 : 10 mg/L (162 µM)
-$MnCl^2$-$4H_2O$: 1.6 mg/L (8.1 µM)

Dissolve 5 gm EDTA in 800 ml water and adjust the pH to 7.5 with NaOH. Add the other components and add water to a final volume of 1 L. Sterilize the solution over a 0.22-µm filter.</td></tr>
<tr><td>Biotin (1 mg/ml)
For 50 ml stock solution, dissolve 50 mg biotin in 45 ml water. Add small aliquots of 1N NaOH until the biotin has dissolved. Add water to a final volume of 50 ml. Sterilize the solution over a 0.22 µM filter. Prepare 1 ml aliquots and store at -20°C.</td><td>Thiamin-HCl (1 mg/ml)
For 50 ml stock solution dissolve 50 mg thiamin-HCl in 45 ml water. Add water to a final volume of 50 ml. Sterilize the solution over a 0.22 µm filter. Prepare 1 ml aliquots and store at -20°C.</td></tr>
<tr><td>1 M $MgSO_4$
$MgSO_4$-$7H_2O$: 24.65 gm/100 ml
For 100 ml stock solution dissolve 24.65 gm $MgSO_4$-$7H_2O$ in 87 ml water. Autoclave for 15 min at 121°C.</td><td>1 M $CaCl_2$
$CaCl_2$-$2H_2O$: 14.70 g/100 ml
For 100 ml stock solution dissolve 14.70 gm $CaCl_2$-$2H_2O$ in 94.5 ml water. Autoclave for 15 min at 121°C.</td></tr>
</table>

2.11 Polymerase chain reaction (PCR) NEB protocol

PCR assay system			**PCR cycle**			
Taq reaction buffer (10x)	5 µl		Step 1	Denaturation	95°C	2 min
dNTPs mix (10 mM)	1 µl		Step 2 **25 cycles**	Denaturation Annealing Elongation	95°C 55°C-65°C 72°C	1 min 1 min 2 min
Forward primer (10 mM)	1 µl					
Revere primer (10 mM)	1 µl					
Template DNA (100 ng/µl)	2 µl		Step 3	Elongation	72°C	20 min
Taq /Q5/Phusion polymerase	2.5 µl		Step 4	Hold	4°C	10 hr
Water	Up to 50 µl		Stored PCR product at 4°C			

2.12 Restriction digestion and ligation protocol

Restriction enzyme digestion			**Ligation reaction**	
PCR product	27 µl (1 µg)		Digested vector	100 µg
NEB buffer (10x)	3 µl (1 x)		Digested PCR product	3x molar conc.
Restriction enzyme-1	1 µl (10 unit)		Ligase buffer (10x)	1 µl (1x)
Restriction enzyme-2	1 µl (10 unit)		T4 DNA ligase	1 µl (10 unit)
Water	Up to 30 µl		Water	Up to 10 µl
Incubate at 37°C for 2hr			Incubate at 16°C overnight and 22°C for 3hrs	

2.13 Routine lab stocks

LB medium -1% Tryptone (10 gm/L) -0.5% yeast extract (5 gm/L) -0.5% NaCl (5 gm/L) Mix in 1 L final volume of water and autoclave for 15 min at 121°C	**LB plate** - 1x sterile LB medium -1.5% Agar (15 gm/L) -Mix well and autoclave/Microwave Let it cool down till ~40°C, add 1x respective antibiotic, pour it to sterile plates and wait till to

	solidify completely.
Coomassie staining solution **(to stain protein SDS-PAGE gel):** -Dissolve 250 mg of Coomassie brilliant blue G250 in 45 ml of Methanol. -Mix well for 15 min. Add 45 ml of distilled water and 10 ml of Acetic acid (100%)	**Coomassie de-staining solution** **(to de-stain protein SDS-PAGE gel):** -200 ml of Methanol (100%) -100 ml Acetic acid (100%) -700 ml of distilled water
1x SDS-PAGE running buffer -25 mM Tris base (MW = 121.14): 3.02 gm -250 mM Glycine (MW = 75.05): 14.4 gm -1% SDS (MW = 288.38) : 1.0 gm Dissolve and adjust the final volume to 1 L using Mili-Q water (~pH 8.8).	**4x protein gel loading dye (Glycerol dye)** -250 mM Tris-Cl, pH 6.8 -5% SDS : 5% -40% Glycerol -5% β-mercaptoethanol -0.04% Bromophenol Blue Add appropriate amount of Mili-Q water
50x TAE buffer (for DNA) -Tris base (MW = 121.14) : 242 gm -Acetate (100%) : 57.1 ml -EDTA: 100 ml 0.5M Na-EDTA (pH 8.0) Adjust the final volume 1 L using Mili-Q water and autoclave the bottle.	**10x TBE buffer (for RNA)** -Tris base (MW = 121.14) : 890mM :(108 gm) -Boric Acid (MW = 61.8) : 890mM : (57.1 ml) -EDTA, pH 8.0 : 20mM : (40ml of 0.5M Na-EDTA) Adjust the final volume 1 L using Mili-Q water and autoclave for 15 min at 121^{o}C.
6x DNA/RNA non-denature gel loading dye (Glycerol dye) -6x TBE buffer -30% Glycerol -0.06% Bromophenol Blue -0.06% Xylene Cyanol FF	**6x RNA/DNA denture gel loading dye (Urea dye)** -10 mM Tris-HCl (pH 7.5) -8 M Urea (MW = 60) -20 mM EDTA -Bromophenol Blue : 0.35 %

Add appropriate amount of nuclease-free water	-Xylene Cyanol FF : 0.35 % Add appropriate amount of nuclease-free water
1x RNA/DNA acrylamide gel staining solution -0.1 % Toluidine blue -10 % Acetic acid Add volume of nuclease-free water.	**1x Agarose gel** -1 x TAE buffer -1 % Agarose - heat it up using a Microwave and 0.5 µg/mL Ethidium bromide (EtBr). Let it solidify.

2.14 RNA migration on a different percentage of PAGE

Concentrations of Acrylamide Giving Optimum Resolution of RNA Fragments Using Denaturing PAGE gel			
Acrylamide (%)	**Fragment sizes separated (bases)**	**Migration of xylene cyanol (bases)**	**Migration of Bromophenol blue (bases)**
20	6 to 100	45	12
15	25 to 150	60	15
12	40 to 200	70	20
8	60 to 400	160	45
5	80 to 500	260	65
3.5	500 to 2000	460	100
From Sambrook J, et al. (2001) Neutral polyacrylamide gel electrophoresis. In: Molecular Cloning. A Laboratory Manual, pp. 5.42, 12.89. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.			

2.15 PAGE (polyacrylamide gel electrophoresis) for RNA/DNA

Native PAGE gel			**Denature PAGE gel**	
Component	**Final**		**Component**	**Final conc.**
TBE (10x)	1.0 x		TBE (10x)	1.0 x
Acrylamide/bis-acrylamide (40%)	X % of gel		Acrylamide/bis-acrylamide (40%)	X % of gel
Ammonium persulfate (10%)	1.0 %		Ammonium persulfate (10%)	1.0 %
Tetramethylethylenediamine (TEMED)	0.1 %		Tetramethylethylenediamine (TEMED)	0.1 %
			Urea	6.5 M
Add appropriate volume of water. Pour into the gel apparatus and wait for 15min			Add the appropriate volume of water. Pour to the gel apparatus and wait for 15 min	

2.16 SDS-PAGE for protein

SDS-PAGE gel					
	Stacking gel (ml)	**Separating gel (ml)**			
Gel percentage	**5%**	**12%**	**15%**	**18%**	**20%**
H2O	3.85	4.3	3.6	2.8	2.3
Tris-base (1.5 M) pH 8.8	-	2.5	2.5	2.5	2.5
Tris-HCl (1.5 M) pH 6.8	0.42	-	-	-	-
Acrylamide and bisacrylamide (40%)	0.625	3.0	3.75	4.5	5
SDS (10%)	0.05	0.1	0.1	0.1	0.1
APS (10%)	0.05	0.1	0.1	0.1	0.1
TEMED	0.005	0.01	0.01	0.01	0.01
Total volume	**5 ml**	**10 ml**	**10 ml**	**10 ml**	**10 ml**
Mix the separating gel, pour it into SDS-PAGE apparatus. Once it solidified, carefully pour stacking gel on top of it and allow it to fully solidify.					

2.17 *In vitro* phosphorylation assay

Protocol for SF1 Phosphorylation		
	Reagent	Concentration
Kinase reaction buffer (1x)	Tris-base Buffer (pH 8.0)	50 mM
	DTT (fresh)	5 mM
	EDTA	1 mM
	MgCl2	30 mM
	PhosStop (Fresh)	1 x
	Glycerol	25 %
	ATP (Fresh)	20 mM
	Stored at -20 °C for later use	
Phosphorylation assay: -Mix SF1 and KIS Kinase in 6:1 molar ratio in Kinase reaction buffer. -Incubate at 30 °C water bath for 5-6 hrs. Stop the reaction by adding an excess of EDTA. -Confirm the phosphorylation status by running 15% SDS-PAGE and mass spectrometry.		

2.18 Sodium phosphate buffer

To prepare 200 ml volume of 100 mM sodium phosphate buffer, the following mixer should be diluted to 200ml final volume with H_2O		
pH at 25˚C	**Volume of 1 M Na_2HPO_4 stock**	**Volume of 1 M NaH_2PO_4 stock**
5.8	1.58 ml	18.42 ml
6.0	2.40 ml	17.60 ml
6.2	3.56 ml	16.44 ml
6.4	5.10 ml	14.90 ml
6.6	7.04 ml	12.96 ml
6.8	9.26 ml	10.74 ml
7.0	11.54 ml	8.46 ml
7.2	13.68 ml	6.32 ml
7.4	15.48 ml	4.52 ml
7.6	16.90 ml	3.10 ml
7.8	17.92 ml	2.08 ml
8.0	18.64 ml	1.36 ml
Note: Use same above mixer to prepare 1 L volume of 20 mM sodium phosphate buffer for NMR sample. (Cold Spring Harbor Protocol, ("Sodium phosphate," 2006).		

2.19 RNA transcription assay

	Reaction mixer	**Volume per reaction (50 µl)**	
1.	Top T7 template (8 µM)	4.0 µl	- Incubate at 37°C for ~3 hrs, run denatured acrylamide gel - purify RNA using long Acrylamide gel or HPLC
2.	DNA template (8 µM)	4.0 µl	
3.	UTP (100 mM)	4.0 µl	
4.	CTP (100 mM)	4.0 µl	
5.	GTP (100 mM)	4.0 µl	
6.	ATP (100 mM)	4.0 µl	
7.	Transcription buffer (20x)	2.5 µl	
8.	PEG (50 %)	5.0 µl	
9.	T7 polymerase	3.0 µl (varies)	
10.	MgCl2 (1 M)	1.0 µl	
11.	DTT (1 M)	0.1 µl	
12.	Nuclease-free water	15 µl	
	Total volume	50 µl	

Methods

2.20 List of protein expression constructs

Different constructs with protein-coding genes were cloned in pETM-11 vector at multiple cloning sites using respective restriction enzymes using standard cloning protocol. The vector comprises an N-terminal 6xHis tag followed by TEV protease sequence and gene with the respective protein constructs. All protein constructs were optimized with a bacterial expression system, and gene sequences were confirmed by the Sanger sequencing method. Several functional versions of subdomain constructs of SF1, U2AF2, KIS kinase, and Npl3 were prepared as listed in the table below:

(a) Splicing factor 1 (SF1) expression constructs

Protein	**Database ID**	
Splicing factor 1	Uniprot ID - Q15637	KEGG ID - hsa:7536
Constructs		
Sub-domain constructs	**Residues boundaries**	**Reference**
SF1	1 to 260 aa	Zhang Y. et al., 2013
SF1-NTD	1 to 145 aa	Zhang Y. et al., 2013
KH-Qua2	135 to 260 aa	Liu Z. et al., 2001
ΔSF1	1 to 260 aa; Δ73-88aa	Current work
ΔSΔN (dSdN)	48 to 260 with deleted Δ73-88aa	Current work
SF1	1 to 320 aa	Current work
SF1-47C, C171N	1 to 260 aa	Current work
SF1-55C, C171N	1 to 260 aa	Current work
SF1-62C, C171N	1 to 260 aa	Current work
SF1-98C, C171N	1 to 260 aa	Current work
SF1-107C, C171N	1 to 260 aa	Current work
SF1-122C, C171N	1 to 260 aa	Current work
SF1-137C, C171N	1 to 260 aa	Current work
SF1-172C, C171N	1 to 260 aa	Current work
SF1-213C, C171N	1 to 260 aa	Current work

(b) KIS kinase (serine/threonine-protein kinase Kist) expression construct

Protein	**Database ID**	
KIS kinase	Uniprot ID - Q8TAS1	KEGG ID - hsa:127933
Construct		
Sub-domain constructs	**Residues boundaries**	**Reference**
KIS kinase	1 to 419 aa	Zhang Y. et al., 2013

(c) U2 small nuclear RNA Auxiliary factor 2 (U2AF2) expression constructs

Protein	**Database ID**	
U2AF2	Uniprot ID - P26368	KEGG ID - hsa:11338
Constructs		
Sub-domain constructs	**Residues boundaries**	**Reference**
RRM1,2-UHM	140 to 475 aa	Current work
RRM1,2	140 to 342 aa	Current work
RRM1,2 extended	140 to 357 aa	Current work
RRM1,2 extended C305S	140 to 357 aa	Current work
UHM-extended	358 to 475 aa	Current work
UHM	370 to 475 aa	Selenko P. et al., 2003

(d) Npl3 expression constructs

Protein	**Database ID**	
Npl3 protein	Uniprot ID - Q01560	KEGG ID - sce:YDR432W
Constructs		
Sub-domain constructs	**Residues boundaries**	**Reference**
RRM1 only	120 to 195 aa	Keil P., Wulf A., Kachariya N, et al, 2022
RRM2 only	196 to 280 aa	As above
Npl3 (RRM1,2)	120 to 280 aa	As above
npl3-D135C & C211S	120 to 280 aa	As above
npl3-E176C & C211S	120 to 280 aa	As above
npl3-N185C & C211S	120 to 280 aa	As above
npl3-D236C & C211S	120 to 280 aa	As above
npl3-RRM1- F162Y	120 to 280 aa	As above
npl3-RRM2- F245Y	120 to 280 aa	As above
npl3-linker P196D & A197D	120 to 280 aa	As above
npl3-W213A	120 to 280 aa	As above

2.21 Protein expression and purification

All the constructs were cloned and expressed in *E. coli* BL21 (DE3) using either Luria broth (LB) or minimal M9 media supplemented with ^{15}N-labeled NH_4Cl and/or ^{13}C-labeled glucose as the sole nitrogen and carbon source, respectively. Protein expression was induced at 0.8 OD_{600} with 0.5 mM isopropyl β-d-1-thiogalactopyranoside (IPTG) and cells were grown at 22 °C overnight. Afterward, the cells were harvested and stored at -20 °C before purification. The cells were re-suspended and lysed using cell lysis buffer (50mM Tris, pH 8, 150mM NaCl, 1mM TCEP, 5mM Imidazole, 1x protease inhibitor cocktail) by either sonication or French press. Insoluble fragments of the lysate were removed using a centrifuge at 7000 rpm for 30 min.

All protein constructs with six histidine tagged were purified using standard affinity chromatography. The cell lysate was passed through the gravity column with 4 ml of Ni-NTA sepharose resins. The resins were then washed with wash buffers (50 mM Tris, 150 mM NaCl, 1 mM TCEP, pH 8) with increasing NaCl concentration from 150 mM to 1500 mM step-wise. The protein was eluted from Ni-NTA resins using 25 ml of elution buffer (50 mM Tris, 50 mM NaCl, 1 mM DTT, 250 mM imidazole, pH 8). The N-terminal His-tag was removed by cleavage with TEV protease in 100x volume of dialysis buffer (20 mM Tris, pH 8, 50 mM NaCl, 1 mM DTT) followed by the reverse Ni-NTA step. To further purify the sample, ion exchange chromatography was done using either a Resource Q or Resource S column, followed by size exclusion chromatography using a Hi-load 16/600 Sepharose S75 column. The purified sample was then exchanged with NMR buffer (containing 20 mM sodium phosphate, pH 6.5, 50 mM NaCl, and 1 mM DTT) until specified. 5 % D_2O was added to the NMR sample to lock the magnetic field,.

For KIS kinase, affinity chromatography was used for purification, followed by buffer exchange using an Amicon with a 10 KDa cut-off membrane. The sample was stored at -80 °C with storage buffer (50 mM Tris, 50 mM NaCl, 10% Glycerol, 5 mM DTT, 1 mM EDTA, 15 mM $MgCl_2$, pH 8).

2.22 *In vitro* SF1 phosphorylation assay

SF1 (1-260 aa) and SF1-NTD (1-128 aa) were phosphorylated *in vitro* using KIS-kinase. For this, KIS kinase was added to SF1 in a 1:6 molar ratio in the kinase buffer. The mixture was

then kept at 37 °C in a water bath for 3-4 hours. After the reaction, the sample was diluted and dialyzed overnight using a dialysis buffer. The phosphorylated SF1 was purified using reverse Ni-NTA affinity chromatography followed by ion exchange and gel filtration chromatography. The final buffer for phosphorylated SF1 was 20 mM MES, 50 mM NaCl, pH 8. The phosphorylation status was confirmed by 15% SDS-PAGE, Electron Spray Ionization (ESI) mass spectrometry and NMR.

2.23 Synthetic and in-house RNA oligonucleotides

Short synthetic RNA molecules that are less than 20 nucleotides long were obtained from either Dharmacon, USA or IBA Lifesceinces, Germany, with a PAGE-purified and desalted purity grade. A library of short, single-stranded DNA oligonucleotides was ordered from eurofins Genomics, Germany, with a HPLC purity grade. Longer RNA molecules that are over 20 nucleotides long were synthesized through an *in vitro* transcription process. This involved mixing a reverse complementary DNA template from Eurofins with a T7 top primer, heating it to 95 °C for 2 minutes, and then cooling it down on the ice for 10 minutes. An in-house prepared polymerase, a 1x transcription buffer, NTPs, MgCl2, PEG, DTT, and nuclease-free water were then added to the mixer. The transcription reaction was carried out for 3 hours at 37 °C and then stopped by flash freezing it at -80 °C. The resulting RNA was further purified through anion exchange chromatography using a DNAPac PA200 HPLC column. The eluted fractions were checked by running a denatured 20 % acrylamide gel, and fractions that belonged to the correct size of RNA were pooled. The fractions were then concentrated using an ethanol precipitation protocol, followed by desalting with NAP-5 columns. The final RNA sample was lyophilized and stored at -20°C.

2.24 Protein backbone chemical shift assignment

Nuclear magnetic resonance (NMR) measurements were conducted on samples placed in either a Shigemi tube or a 3 mm or 5 mm regular NMR tube at 25 °C. These measurements were taken using Bruker spectrometers operating at proton Larmor frequencies of 500, 600, 800, 900, 950, and 1200 MHz, equipped with either room temperature or cryogenic probes. The recorded NMR spectra were processed using a shifted sine-bell window function and zero-filling before Fourier transformation using either Bruker Topspin 3.5pl6 or NMRPipe (Delaglio et al., 1995) software package. Proton chemical shifts were referenced against sodium 2,2-dimethyl-

2-silapentane-5-sulfonate (DSS). All spectra were analyzed using CCPN analysis v2.5 software (Vranken et al., 2005).

The NMR backbone assignments for subdomain constructs for SF1 and U2AF2 were obtained from previously deposited BNMRB databases (ID 18808, 188802, 17623, and 17622). However, the missing assignments for U2AF2 (35 residues long linker between RRM2 and UHM) were assigned using conventional triple resonance experiments, including HNCACB, HNcoCACB, HNN, HNCO, HNcaCO, and HcccoNH. The NMR backbone assignment for subdomain constructs for SF1 and U2AF2 were obtained from previously deposited BNMRB (ID 18808, 188802, 17623 and 17622) database. However, the missing assignments for U2AF2 (35 aa long linker between RRM2 and UHM) were assigned by conventional triple resonance experiments: HNCACB, HNcoCACB, HNN, HNCO, HNcaCO and HcccoNH.

For Npl3, backbone chemical shift assignments for RRM1 and RRM2 were obtained from the BMRB (ID 7382 and 7383), and missing assignments for tandem RRM domains of $Npl3^{120\text{-}280}$, $npl3^{120\text{-}280}$-linker mutants were obtained using similar triple resonance experiments as mentioned above.

2.25 Protein-ligand interaction by NMR titration

For ligand-binding studies, a series of ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectra were recorded using 50 μM of ^{15}N labeled protein with gradually increasing concentration of unlabeled ligands such as protein, single-stranded DNA (ssDNA) or RNA. The synthetic ssDNA and RNA were purchased from Eurofins Genomics and Dharmacon, USA, or Biolegio BV, Germany, respectively. NMR titrations were carried out with 4-fold excess of ligand concentration with respect to protein concentration, and 25 ^{o}C temperature was used for all the measurements. The NMR buffer (20 mM $NaPO_4$, 50 mM NaCl, 1 mM DTT, pH 6.5) was used for these experiments. All the spectra were analyzed by CCPN software tool.

The cumulative chemical shift perturbation upon ligand binding (CSPs, Δδ) were calculated as per the equation: $\Delta\delta = \left[(\Delta\delta 1H)^2 + \frac{(\Delta\delta 15N)^2}{25}\right]^{\frac{1}{2}}$. Also, dissociation constants (K_D) were derived from individual sets of NMR titrations by fitting to the equation:

$\Delta\delta_{obs} = \frac{\Delta\delta_{max}}{2[P]_t}\{([P]_t + [L]_t + K_D) - ([P]_t + [L]_t + K_D)^2 - 4[P]_t[L]_t]^{0.5}\}$, Where $\Delta\delta_{obs}$ is the observed chemical shift difference relative to the free state, $\Delta\delta_{max}$ is the maximum shift change in saturation, $[P]_t$ and $[L]_t$ are the total protein and ligand concentrations, respectively, and K_D is

the dissociation constant (Williamson, 2013). The self-written script was implemented to calculate CSPs, map binding surface area, and extract K_D using the quadratic fit function and final plotting.

2.26 {^{1}H}-^{15}N heteronuclear NOE experiment

For KH-Qua2, SF1, U2AF2, and Npl3 steady state {^{1}H}-^{15}N heteronuclear NOE (hetNOE) spectra were recorded simultaneously with and without proton saturation and spectra acquired using 100ms and 120ms acquisition time for direct ^{1}H and in-direct ^{15}N dimensions, respectively. The inter-scan delay was set to 3 seconds and experiments were recorded with ~100 to 300 µM protein concentration. The measurement temperature was set to 25 °C. After the measurement, the spectra were split by Bruker AU program and further processed. The dynamic information for all residues were extracted from the intensity ratio of with and without saturation spectra. Also, {^{1}H}-^{15}N heteronuclear NOE (hetNOE) experiments were recorded for the protein-RNA complex at 90 µM protein in the presence of 3-fold molar access of RNA.

2.27 ^{15}N R_1 and R_2 relaxation measurements

For KH-Qua2 and U2AF2 proteins, a pseudo-3D version of a ^{15}N transverse (R_2) relaxation experiment was recorded on a 600 MHz spectrometer with CPMG pulse trains at 16.96, 33.92, 67.84, 101.76, 135.68, 169.60, 203.52, 237.44, 271.36, 339.2 and 440.96 ms delays at 25 °C temperature. For each delay, 2D ^{1}H,^{15}N planes were extracted and processed with shifted sine-bell window function and zero-filling before Fourier transformation by Bruker Topspin 3.5pl6 software. Spectra were analyzed by CCPN tool. The signal intensity decay of each amide signal was fitted to an exponential decay function and extracted the R_2 rates for each residue. For Npl3$^{120\text{-}280}$, ^{15}N R_2 relaxation data were measured with CPMG pulse trains at 16.96, 33.92, 50.88, 67.84, 101.76, 135.68, 169.60, 203.52 and 254.40 ms. Signal intensity decay was fitted to exponential decay and extracted 1/T_2 time (means R_2 rates).

For KH-Qua2 and U2AF2 proteins,^{15}N longitudinal relaxation rates (R_1) were measured by pseudo-3D version experiments on a 600MHz spectrometer by sampling the exponential decay function of intensity delays with 20, 60, 100, 200, 400, 600, 800, 1000, 1200 and 1500 ms at 25 °C temperature. For Npl3$^{120\text{-}280}$, R_1 rates were measured by sampling exponential decay function of delays with 60, 100, 160, 200, 280, 400, 600, 800, 1200 and 1800 ms. Similarly, for each delay, 2D ^{1}H,^{15}N planes were extracted, processed with shifted sine-bell

window function and zero-filling before Fourier transformation by Bruker Topspin 3.5pl6 software. The signal intensity decay was fitted to an exponential decay to extract R_1 rates.

From estimated R_1 and R_2 rates for KH-Qua2, U2AF2 and Npl3$^{120\text{-}280}$, the total rotational correlation time (τ_c) for all residues was determined from an equation mentioned below : Total correlation time, $\tau_c = \frac{1}{4\pi\nu_N}\left(6\left(\frac{R_2}{R_1}\right) - 7\right)^{\frac{1}{2}}$, where R_2 is transverse relaxation rates, R_1 is longitudinal relaxation rates, ν_N is ^{15}N Larmor frequency.

2.28 Paramagnetic relaxation enhancement experiment

(a) SF1$^{1\text{-}260}$ protein

To modify the SF1$^{1\text{-}260}$, a single cysteine point mutations were introduced at 47C, 55C, 62C, 107C, 122C, 137C, 172C and 213C and the native Cys171 was changed to Asn. The mutant proteins were purified as described above. Before spin labeling, samples were reduced with 5 mM DTT and dialyzed overnight using PRE buffer (50 mM Tris, pH 8.0, 50 mM NaCl) and then treated with 10x excess of a solution containing 3-(2-Iodoacetanido)-PROXYL (IPSL) spin-label. Afterward, the excess spin-label was removed using a desalting PD-10 column, followed by the buffer exchange using Amicon filter. All steps were performed in the dark to avoid light interference. The labeling efficiency of proteins was confirmed using native Electron Spray Ionization mass spectrometry (ESI-MS). The oxidized state of the samples was measured on 950MHz Bruker spectrometer. Afterward, the sample was reduced by adding 10-fold molar excess of ascorbic acid followed by one hour incubation to ensure complete reduction. ^{1}H-^{15}N HSQC spectra were recorded with the oxidized and reduced samples with a 3 sec inter-scan delay for 7 hrs. Spectra were processed and analyzed.

(b) For U2AF2 and SF1 complex

UHM domain of U2AF2 has five cysteines located in the core region of the globular domain and one cysteine on RRM2. Thus, the segmental isotope labeling (SIL) approach was used. For that, C305 of RRM1,2 (U2AF2) was mutated to serine and single cysteine at 274C, 315C and 326C were introduced. Unlabeled RRM1,2 was purified and SL was attached to respective positions as described above. Also, ^{15}N labeled UHM domain of U2AF2 was purified. The spin-labeled RRM1,2 was then ligated to the UHM domain of U2AF2 using the sortase and

purified using size exclusion chromatography. Spectra were measured in both oxidized and reduced states of the samples, both in the presence and absence of SF1-RNA complex, and further analyzed using CCPN.

(c) For Npl3$^{120-280}$

Similarly, to conduct PRE experiments on the Npl3$^{120-280}$ protein, a single cysteine mutant was introduced at positions 135C, 176C, 185C, and 236C. The native cysteine at position 211 was replaced with serine. Afterward, SL was attached to each mutant at its respective positions, as described above. The samples were then oxidized and reduced, and ^{1}H-^{15}N HSQC spectra were measured. The resulting spectra were analyzed to determine the intensity ratio. PREs with the RNA bound Npl3$^{120-280}$ form were recorded for spin labels at residue 185 and 236 with a 3-fold excess of the “CN--GG” RNA oligo.

2.29 Structural determination using rigid body refinement

(a) SF1$^{1-260}$ protein

For SF1, high-resolution structural coordinates of HH, KH and Qua2 domains were used from the Protein Data Bank (PDB), accession codes 4FXX and 1K1G, respectively. N-terminal ULM and helix of Qua2 were treated as flexible during rigid-body refinement steps, while a degree of freedom was allowed between HH and KH domains during simulation. In the PDB file, cysteine coordinates were replaced at the position of 47C, 55C, 62C, 107C, 122C, 137C, 172C, and 213C, and coordinates of spin label IPSL moieties were attached to the cysteine side chain. A short molecular dynamic (MD) simulation was performed to randomize IPSL moieties, ULM, HH, KH, and Qua2 domains and these coordinates were used as a starting template for modeling. Intra- and inter-domain distance restraints were derived from PRE experiments for individual datasets, and structural refinement was performed using CNS 1.2. At the end of the refinement, 100 models were generated and analyzed. An ensemble of the 20 lowest energy structures were selected and further evaluated. Quality factor (Q-factor) was derived by comparing the back-calculated (from the generated models) and the experimental PRE curves as described. For solvent refinement, IPSL coordinates at cysteine positions were removed, native residues were replaced in the final ensemble structures and a short energy minimized simulation was performed. Final refinement in explicit water was performed using

CNS 1.2. The structural quality of the final ensemble was analyzed by Ramachandran plot using Procheck_NMR 3.5 (Laskowski et al., 1996). PyMol (http://pymol.org/2/) was used for visualizing the protein structures.

(b) $SF1^{1-260}$-$U2AF2^{140-475}$ complex

For ensemble structural calculation of $SF1^{1-260}$ - $U2AF2^{RRM1,2-UHM}$ complex free and bound to RNA comprise BPS^{opt} and PPT^{opt} motif, similar semi-rigid body refinement approach was applied using high-resolution individual domain structures of SF1 and U2AF2. For U2AF2, RRM1,2 inter-domains and with RNA distances were derived from PDB id 5EV1 and 4FXW, while distances between KH-Qua2 to BPS RNA were derived from 1K1G PDB structure. Inter-domain distance restraints between SF1 and U2AF2 were derived from SF1 PRE model of SF1, PRE data from complex and PDB structures. The starting PDB template with randomized domains of SF1, U2AF2 and RNA was generated using CNS1.2. The ensemble description of structure was accessed by the ensemble optimization method (EOM) based on experimental SAXS data.

(c) $Npl3^{120-280}$ protein

A similar, semi-rigid body refinement approach was used for the ensemble structure of tandem RRMs of Npl3. In brief, structural coordinates of the individual RRM domains were taken from the PDB ID 2OSQ and 2OSR for RRM1 and RRM2, respectively. PRE-based experimental distance restraints were derived for individual datasets of PRE and TALOS-N based backbone torsion angle restraints were used. Combining torsion angle and distance restraints, structural refinement was carried out using CNS 1.2. 100 randomized models were generated and analyzed. Quality factor (Q-factor) was derived by comparing the back-calculated PRE curves (from the generated models) and the experimental PRE as described (Simon et al., 2010). An ensemble of the 15 lowest energy structures was selected and further analyzed. For solvent refinement of the final structures, spin label moieties were removed, and native residues were replaced in the final ensemble PRE model. The position of the rest of the atoms was fixed and a short energy minimized simulation was performed. Final refinement in explicit water was performed using Aria 1.2/CNS 1.2. Backbone structural quality of the final ensemble of structures was checked by Ramachandran plot using Procheck_NMR 3.5. To derive a structural model of the protein-RNA complex, PRE experiments were measured with

spin labels attached to 185C (on RRM1) and 236C (on RRM2) positions in the presence of 3-fold excess of “CN--GG” RNA. Protein-RNA distance restraints were obtained from chemical shift perturbations seen in NMR titration and based on the homologous structure (PDB: 2M8D and 5DDR). CNS1.2 was used to generate a pool of 400 models and scored them against experimental SAXS data.

2.30 Small angle X-ray scattering

(a) SF1- U2AF2 in free and RNA bound complex

For $SF1^{1-260}$, $U2AF2^{140-475}$, $SF1^{1-260}$-U2AF2 complex, SAXS data were collected using in-house Rigaku BIOSAXS1000 instrument as well as Deutsches Elektronen-Synchrotron (DESY) in Hamburg and European Synchrotron Radiation Facility (ESRF) in Grenoble. Before the measurements, protein-only, protein-protein complex and protein-protein-RNA complex samples were prepared using size-exclusion chromatography to ensure an equimolar ratio of the complex.

The in-house Rigaku BIOSAXS1000 instrument is mounted to a Rigaku HF007 microfocus rotational anode with a copper target (40 kV, 30 mA). Transmissions were measured with a photodiode beam stop and calibration was carried out with a silver behenate sample (Alpha Aeser). For in-house instrument, samples were measured in 12900 second frames to check beam damage. Before the measurements, all samples were dialyzed overnight using SAXS buffer (20 mM Tris, 50 mM NaCl, 2 mM DTT, pH 7.4). The different concentrations ranging from 2 to 10 mg/ml were measured for each sample dataset at 25 °C to exclude concentration-dependent structure factors. Multiple time buffers were measured in between each run and buffer subtraction was applied using SAXSLab software (v3.02).

Size-exclusion SAXS (SEC-SAXS) measurements were carried out at DESY and ESRF synchrotron and data were analyzed by Chromixs (ATSAS package v3.0.0). The radius of gyration (R_g), normalized pair-distance distribution function *P(r)*, normalized Kratky plot and double logarithmic plots were calculated and analyzed for all samples by using ATSAS software package 3.0.0 (Manalastas-Cantos et al., 2021). CRYSOL (ATSAS package) was used to generate back-calculated theoretical SAXS curves from the structural models generated by experimental PRE data (Franke et al., 2017). The ensemble optimization method (ATSAS package) was used for generating a randomized pool of models and to select the ensemble

representative models (Sagar et al., 2021).

(b) Npl3$^{120-280}$ in free and RNA bound complex

Similarly, for Npl3 protein, SAXS measurements were performed in-house on a Rigaku BIOSAXS1000 instrument with a similar set-up described above. Samples were measured in 12900 second frames to check beam damage. Samples were dialyzed with NMR buffer before measurement, and protein-RNA complex with “CN--GG” RNA was prepared using size exclusion chromatography. To eliminate a concentration-dependent effect, a different concentration range from 2 to 8 mg/ml was measured for each dataset at 4°C. Also, the buffer was measured multiple times in-between each run and applied for buffer subtraction by SAXSLab software (v3.02). Pair-distance distribution function, *P*(r), and double logarithmic plots were calculated using the ATSAS software package 3.0.0. Theoretical SAXS curves from structures were generated by CRYSOL.

2.31 Isothermal titration calorimetry

ITC experiments were conducted on a MicroCal PEAQ-ITC device (Malvern, UK). For Npl3 protein, samples were dialyzed against NMR buffer (20 mM sodium phosphate, pH 6.4, 50 mM NaCl, 1 mM TCEP). For SF1$^{1-260}$, U2AF2$^{140-475}$ and SF1$^{1-260}$- U2AF2$^{140-475}$, all samples were dialyzed against ITC buffer (20 mM Tris, pH 7.4, 50 mM NaCl, 1 mM TCEP). In each case, the ITC cell was filled with a 15 µM concentration of either RNA or protein, while the ITC syringe was filled with the protein sample. The ITC titrations were performed with 39 points of 1 µl injections with a 150 sec interval at 25°C. All measurements were performed in duplicates and analyzed using Malvern’s MicroCal PEAQ-ITC analysis software (v1.0.0.1259). Binding curves were fitted to one-site binding mode and thermodynamic parameters were extracted.

Chapter 3 – Recognition of 3’ splice site RNA by SF1 and U2AF2 complex

Introduction

3.1 Pre-mRNA splicing

Splicing, i.e. the removal of non-coding introns from the pre-mRNA transcripts and joining the exons through the spliceosome is an essential aspect of post-transcriptional gene regulation in the eukaryotic organisms. This process leads to the synthesis of mature mRNA and the translation of a functional protein. Yeast has approximately 5,000 protein-coding genes, with only around 400 containing one intron. On the other hand, humans have 20,000 protein-coding genes, with multiple non-coding introns averaging 1000 bases in length. Nearly 95% of these genes undergo alternative splicing, which creates functional protein isoforms to regulate diverse cellular functions. Therefore, constitutive and alternative splicing events are tightly regulated, and any mutations or alterations in the pre-mRNA processing can disrupt cellular homeostasis and cause various diseases, including cancer, metabolic, and immune disorders in humans.

Pre-mRNA has specific sequence elements called "splice sites" on the intron, which are crucial for initiating the splicing process. These sites can be found on both introns and exon-intron boundaries, including the polypyrimidine tract (PPT), branch point sites (BPS), 5' splice site (ss), and 3' ss, as depicted in **Figure 3.1**. Interestingly, yeast lacks PPT splice sites but has conserved BPS sites, while human and other eukaryotic species have canonical PPT sites but highly degenerative BPS sites. These differences explain the diverse splicing regulation and complexity found in both yeast and human (Xu et al., 2021) (Ren et al., 2021) (Urbanski et al., 2018) (Cartegni et al., 2002) (Padgett, 2012) (De Conti et al., 2013).

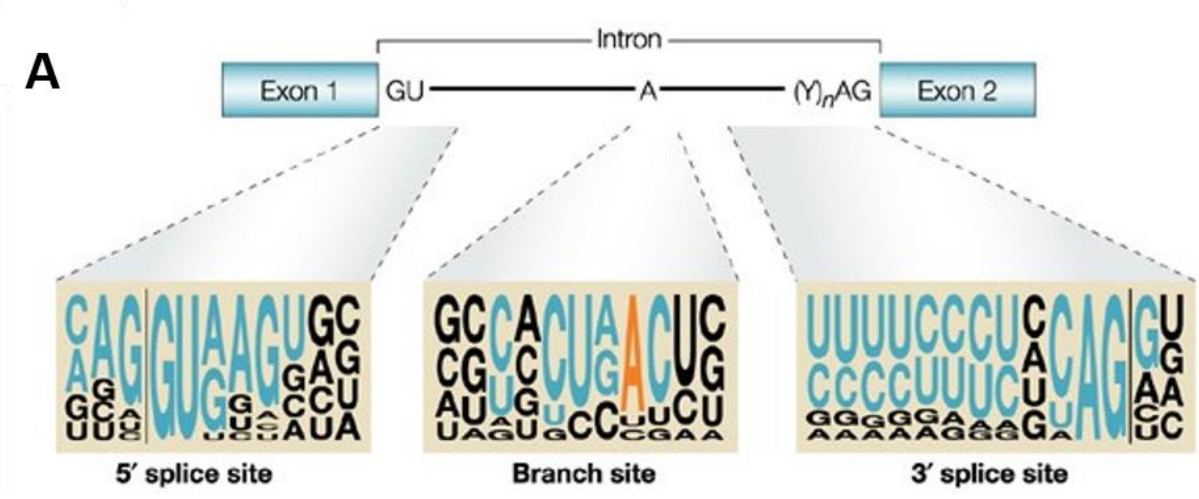


Figure 3.1. Schematic diagram of precursor mRNA sequence highlighted with exon-intron boundaries and consensus splice sites. (A) 5’ GU and 3’ AG dinucleotides at the intron ends, the poly-pyrimidine tract (PPT), and the branch point site (BPS) sequences are shown in a two-exon pre-mRNA (Figure adapted from Cartegni, Chew et al. 2002).

3.2 Alternative splicing

Constitutive splicing produces a single complete transcript from the pre-mRNA, resulting in a full-length mature mRNA that translates into a full-length protein. On the other hand, alternative splicing generates multiple transcript mRNA copies from the same pre-mRNA sequence, leading to the expression of different isoforms of the protein with varying cellular functions. Constitutive splicing is typically driven by strong and consensus sequences that are easily accessible by spliceosome complexes. In contrast, alternative splicing is often observed for weaker consensus splice sites and is regulated by auxiliary splicing regulatory elements such as ESE (exonic splicing enhancer), ESS (exonic splicing silencer), ISE (intronic splicing enhancer), and ISS (intronic splicing silencer).

Various types of alternative splicing patterns have been identified, forming different variants of mature mRNA, as illustrated in **Figure 3.2**. These patterns include exon skipping, intron retention, alternative 3’ splice sites, alternative 5’ splice sites, alternative promoters, Cassette exons, exon scrambling, and mutually exclusive exons. These splice variants are considered as canonical products of alternative splicing and expressed in somatic cells for the normal cell-growth or to assist cell specific functions. Such constitutive and alternative splicing processes are mainly regulated by several cis- and trans-acting splicing factors, which are influenced by RNA binding proteins (RBPs) located near consensus splice sites. These RBPs act as enhancers or repressors for splicing factors. Some of the well-characterized examples include heterogeneous nuclear ribonucleoproteins (hnRNPs), serine/arginine-rich (SR) proteins, poly-pyrimidine track binding proteins (PTB), and others. However, alterations in canonical intron sequences or accessory splicing factors can cause to global splicing deregulation and result in a aberrant splicing products (Kramer, 1992) (Stanley & Abdel-Wahab, 2022) (Ni et al., 2007) (Wright et al., 2022) (Ren et al., 2021) (da Costa et al., 2017) (Chen & Weiss, 2015).

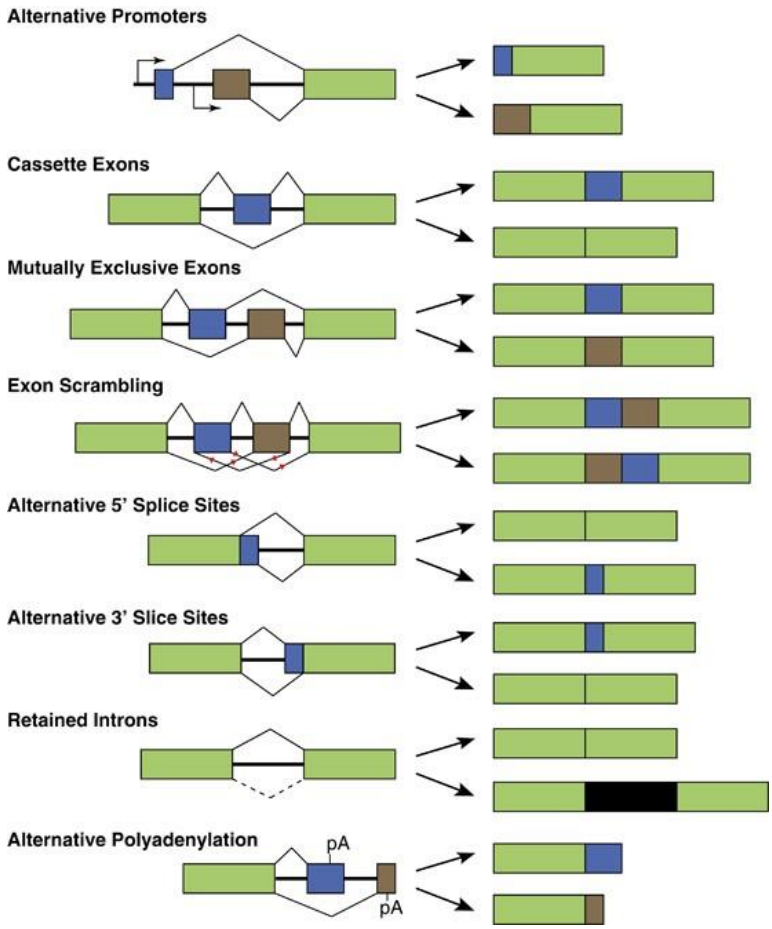


Figure 3.2. Constitutive splicing and alternative splicing of pre-mRNA transcript. The constitutive splice products are shown in green (right panel), while alternative splice products are highlighted in blue, brown or black (right panel). The different processes of alternative splicing can give different mature mRNA transcripts as shown on the right side of the figure (figure is taken from Chen and Weiss 2015).

3.3 Chemistry of splicing catalytic reaction

The spliceosome assembly plays a crucial role in synthesizing mature mRNA from pre-mRNA sequences. This process involves removing intron sequences through two trans-esterification steps, followed by the catalysis reaction that occurs in two main steps: branching and exon ligation, as shown in **Figure 3.3**. During the first branching step, the spliceosome brings the branch point adenosine and 5' splice site close together. The 2' hydroxyl group of the branch point adenosine then attacks the phosphate group at the 5' ss, resulting in the formation of a 3'-lariat intermediate exon. In the second step, the spliceosome undergoes a major conformational rearrangement, bringing the 3' and 5' splice sites together. The 3' hydroxyl group of a nucleotide at the 5' splice site then undergoes a second nucleophilic attack at the phosphate group of the 3' splice site, resulting in exon ligation. At the end of the reaction, the

lariat intron is removed along with the spliceosomes, and two consecutive exons are ligated (Query et al., 1996) (Mikheeva et al., 2000) (Li & Tarn, 2006) (Wilkinson et al., 2020).

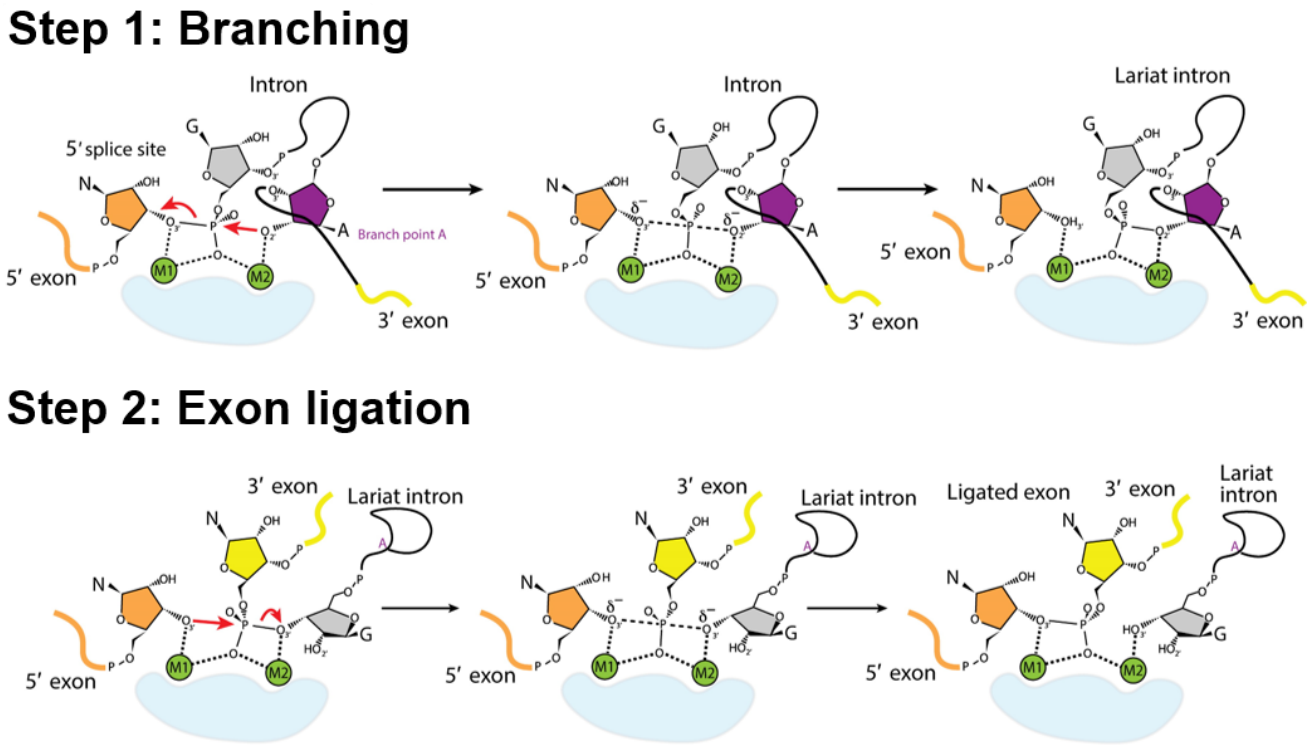


Figure 3.3. Schematic representations of the splicing catalytic reaction. In the branching reaction step, the branch site “A” attacks at phosphorus of the 5’ ss and produces a lariat–3’ exon intermediate. In the exon ligation step, the 5’ and 3’ exons are ligated via the nucleophilic attack of the 5’ exon 3’OH group at the phosphorus atom of the 3’ ss. (Figure is adapted from Wilkinson, Charenton et al. 2020).

3.4 Splice site recognition by spliceosomes

During the splicing cycle, the consensus exon-intron sequence boundaries of pre-mRNA are processed by five small ribonucleoprotein complexes, known as spliceosomes, along with other regulatory proteins. The spliceosomal assembly process begins with U1 snRNPs binding to the 5' splice sites and splicing factors binding to the branch point and 3' splice sites in ATP independent manner, forming the very early-stage **complex E**. This triggers the binding of U2 snRNPs at the branch point sites by replacing splicing factor 1, forming a transient splicing complex called **complex A**. These complexes assist in the assembly of U4-U6 and U5 trimmer snRNPs to form complex B, which undergoes significant conformational rearrangement and becomes active complex B by removing U1 and U4 snRNPs.

Next, the intermediate 3' lariat exon in complex C is formed after the trans-esterification

steps and conformational rearrangement. In the final step, adjacent exons ligate to create the mature mRNA by removing the lariat intron, snRNPs, and non-snRNP protein assemblies. This whole process is dynamic and regulated by the association and dissociation of several spliceosome assemblies and splicing factors, which ultimately regulate the consecutive and alternative splicing of pre-mRNA. Any changes in these splicing proteins or mutations in the consensus splice sites could lead to severe diseases or defects (**Figure 3.4**) (Tanackovic & Kramer, 2005) (De Conti et al., 2013) (Wilkinson et al., 2020) (Wan et al., 2019) (Ren et al., 2021).

The assembly's characteristic features are dynamic conformations of inter-domain arrangements and cooperative and transient interactions. NMR and other biophysical solution techniques would provide better insights into the molecular mechanisms at an atomic level resolution in such a complex system.

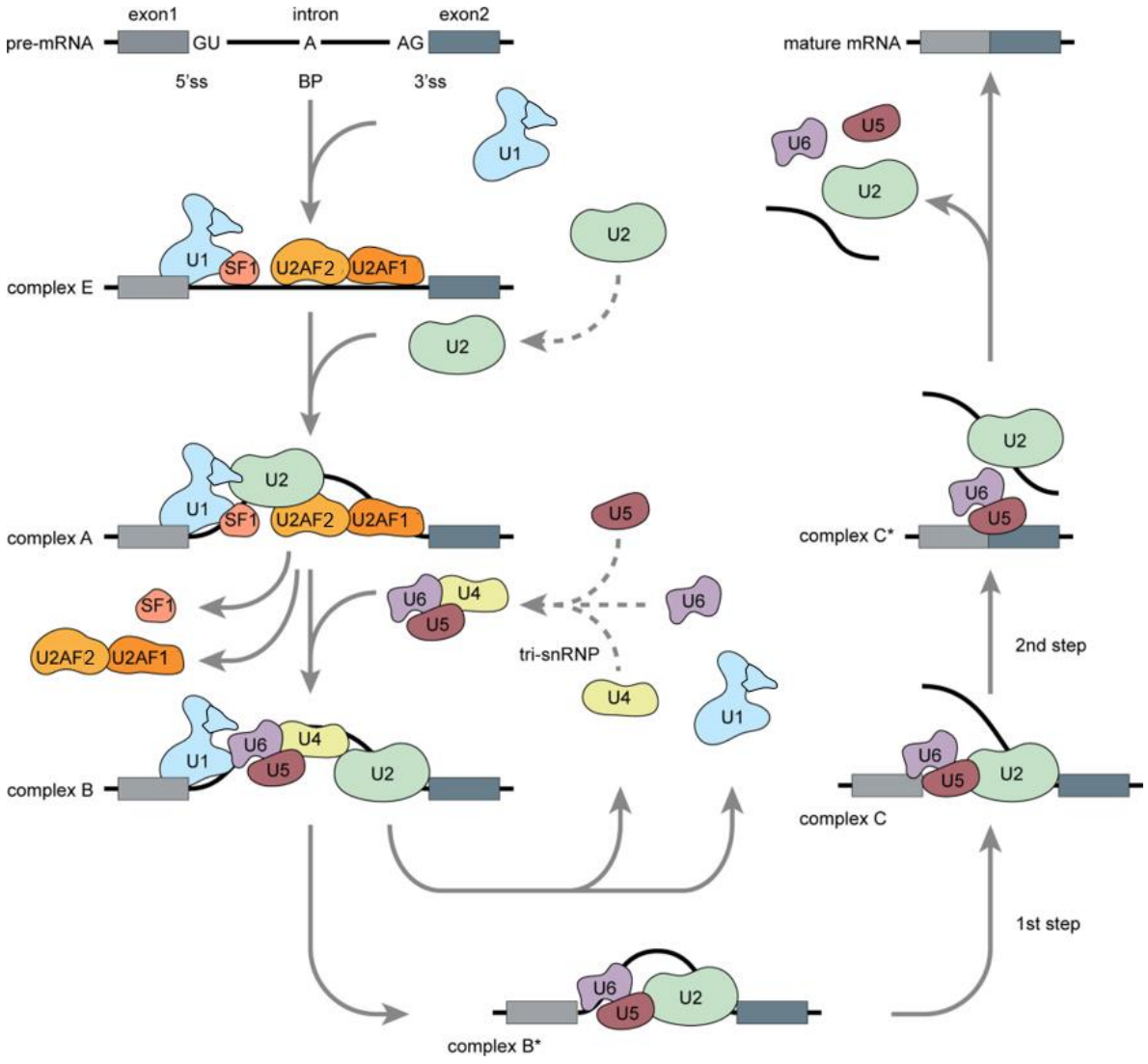


Figure 3.4. Schematic representation of splicing factors, spliceosome assembly, and splice site recognition in humans. A splicing cycle is initiated with the loading of U1 snRNP assembly onto pre-mRNA via 5'ss and splicing factors at 3' splice sites to form complex E. U2 snRNP replaces splicing factor, SF1 at the branch point and forms a complex A. Then U4/U6–U5 tri-snRNP complexes are loaded to form complex B and catalytically active complex B* via the release of U1 and U4. Then complex B* is converted into a final complex C* where U2, U5, and U6 sn-RNPs are released with intron excision, exon ligation, and mature mRNA formation (figure adapted from Ren, Lu et al. 2021 with minor modification).

3.5 Complex E: an early stage of spliceosome assembly at 3' splice site

To begin the splicing cycle, specific splicing factors first bind to the consensus splice sites located on the boundaries between introns and intron-exons. This helps the spliceosomes assemble and initiate the splicing machinery. In humans, U2 auxiliary factors 1 (U2AF1), U2AF2, and splicing factor 1 (SF1) recognize the 3' ss of the intron through the conserved 3' "yAG" ss, polypyrimidine tract (PPT), and branch point site (BPS). This forms the early stage of splicing assembly, also known as complex E (**Figure 3.5**).

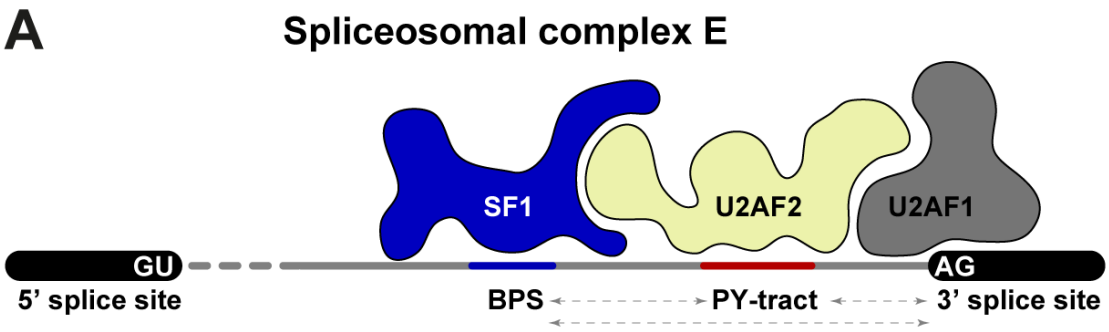


Figure 3.5. The early stage of splicing assembly at the 3' splice site. (A) The branch point site (BPS) and polypyrimidine tract (PPT) and 3' "AG" splice sites are recognized by the corresponding splicing factors SF1, U2AF2, and U2AF1 and form the early stage of splicing complex E.

The complex architecture and protein-protein interactions of the splicing factors that stabilize the splicing assembly and initiate the splicing cycle, are important characteristics of the splicing assembly. **Figure 3.6** shows that U2AF1's UHM (U2AF homology motif) domain binds to U2AF2's ULM (UHM ligand motif) motif, while U2AF2's UHM domain interacts with SF1's ULM and helix-hairpin (HH) domain. U2AF2's RRM1,2 domains bind to consensus PPT sites downstream of 3' "yAG" splice sites, and its N-terminal serine/arginine-rich (SR) region acts as a cis-/trans-regulatory motif, binding to non-specific intronic sequences near PPT sites.

Conversely, SF1 has a multi-domain architecture with an N-terminal ULM, followed by HH, KH (hnRNP K homology), Qua2 domains, and C-terminal proline-rich motifs. The KH and Qua2 domains are responsible for recognizing BPS RNA sites located upstream of PPT (Liu et al., 2001) (Selenko et al., 2003) (Pastuszak et al., 2011) (Loerch & Kielkopf, 2016) (Gupta et al., 2011).

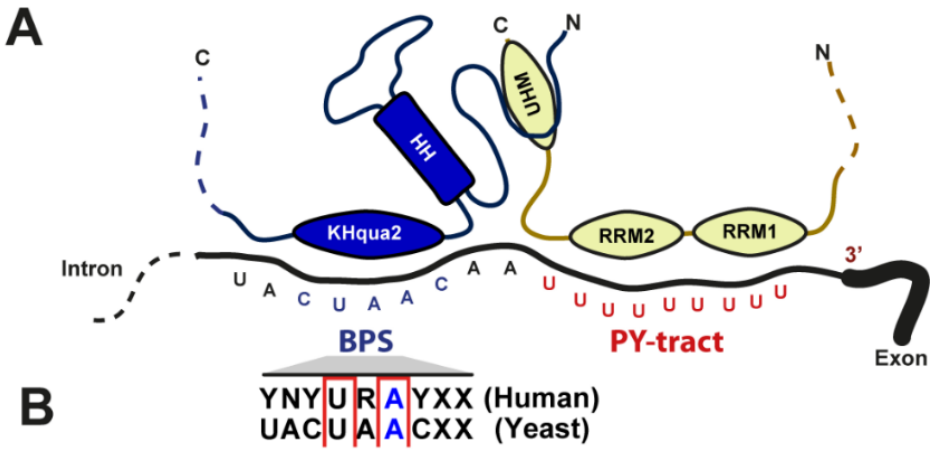


Figure 3.6. Domain arrangement of splicing factors U2AF2 and SF1. (A) RRM1 and RRM2 of U2AF2 bind to PPT sites, while UHM binds to ULM domain of SF1. SF1's KH-Qua2 domain recognizes BPS RNA sites. (B) The sequence alignment of yeast and human BPS sites highlights the conserved sequences.

3.6 Branch point site recognition by SF1

The first step in the splicing catalytic reaction is to identify the diverse sequence type of branch point sites, 5' ss and 3' ss by spliceosome assemblies. Any changes in recognition of the specific splice sites can disrupt the process, leading to exon/intron skipping and aberrant splicing. In humans, the canonical BPS site motif is defined as "YNYUR**A**Y" (A: adenosine branch site, R: any nucleotide, Y: pyrimidine), which is highly degenerative, and only branch site "A" with up-stream "U" at -2 positions are conserved. Due to the degenerative motifs, several hundred branch point sites have been identified in human pre-mRNA introns. In contrast, BPS is nearly invariable as "UACUA**A**C" across different intron sequences in yeast (*S. cerevisiae*), and recognized by BBP (branch point binding protein), which is an ortholog of human SF1. Additionally, yeast introns lack the extended PPT between BPS and 3' splice sites, while PPT is relatively conserved in humans and other higher evolutionary eukaryotes. This explains how the fully conserved BPS recognized by BBP helps in the splicing catalytic reaction. Although human and yeast splice sites differ, both species contain similar functional domains

of SF1 and U2AF2.

Splicing catalytic reactions can still occur in the presence of degenerative BPS motifs, as SF1 and U2 snRNP (in later steps) can locate representative BPS splice sites. However, the point mutation at BPS sites on specific introns can lead to splicing defects and results in various human diseases such as Fish-eye disease, Ehler-Danlos Syndrome type-II, Epidermolysis bullosa with pyloric atresia, Extrapyramidal movement disorder, hyper-triglyceridemia, and cardiovascular diseases. In such cases, spliceosome assemblies unable to locate canonical BPS sites lead to the recognition of nearby cryptic sites resulting in alternative splice products. The three-dimensional structure of the SF1's KH-Qua2 domain with the BPS motif "UACUAACAA" has been known for some time (**Figure 3.7**); however, the structural insight with different disease-causing BPSs needs further study (Berglund et al., 1997) (Mercer et al., 2015).

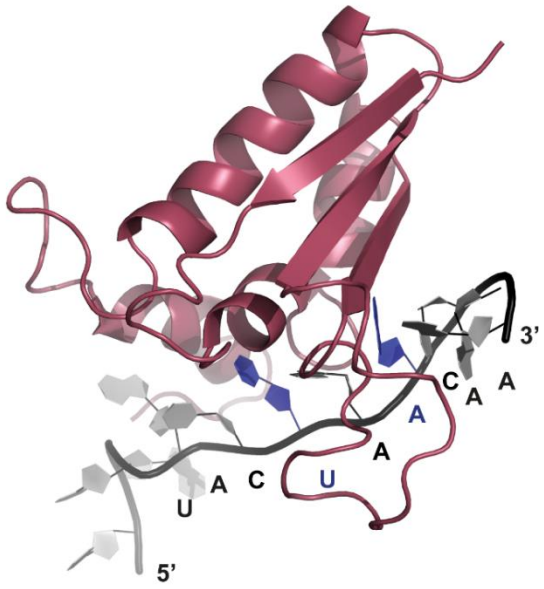


Figure 3.7. Structure of KH-Qua2 domain of human SF1 bound to BPS RNA. The BPS "5'-UACUAACAA -3'" RNA sequence is shown in black. The branch site "A" and +2 upstream "U" nucleotides are highlighted with blue (PDB accession code 1K1G).

3.7 Diverse regulatory functions of SF1

In the early stages of spliceosome assembly, splicing factors SF1 and U2AF2 bind to BPS and PPT sites, respectively. In subsequent steps, the U2 snRNP complex replaces SF1 and binds to the BPS site. Studies have shown that the presence of SF1 is not essential for the splicing cycle, and the branch-site catalytic reaction can still occur in *SF1* knock-out cell lines.

However, the splicing reaction rate slows down without SF1, indicating that SF1 influences the efficiency of splicing kinetics (Tanackovic & Kramer, 2005) (Guth & Valcarcel, 2000) (Rutz & Seraphin, 2000). Also, a recent study revealed that the helix of SF1's C-terminal proline-rich motif interacts with protein-subunits from U2 snRNP assembly via the SURP domain (A. Crisci et al., 2015) (Nameki et al., 2022). Additionally, the *SF1* gene has six protein isoforms that differ in length due to alternative splicing. These isoforms have a varying length of proline-rich tail at the C-terminal of SF1. One of the SF1 isoforms detected at mRNA lacks the first 115 residues of the ULM domain, which is essential for U2AF2 interaction. This may explain the unique splicing or non-splicing function of SF1 in cells. To summarize, all these analyses shed light on the specific role of SF1 in the complex process of the splicing cycle. Additionally, SF1 and U2AF2 are highly expressed in the nucleus and nuclear-paraspeckle, a condensed region within the nucleus that stores splicing-related proteins when not in use. The presence of RNA-free SF1 with U2AF2 in the nuclear paraspeckle suggests a new regulatory role for the SF1-U2AF2 complex that could be independent of splicing (Rino et al., 2008). The long non-coding RNA (lncRNA) Gomafu contains tandem repeats of “UACUAAC” sequences which constitutes SF1 RNA binding sites. Since Gomafu stably accumulates in the nucleus, it may regulate local SF1 concentration and also the splicing cycle (Romero-Barrios et al., 2018) (Tsuiji et al., 2011) (Ishizuka et al., 2014).

Previous studies indicate that genetically modified *SF1+/-* mice have reduced levels of SF1 protein in their tissues, resulting in decreased in intestinal polyp development. Additionally, a recent study suggests that SF1 plays a crucial role in alternative splicing for extending the lifespan of *Caenorhabditis elegans*. Over-expressing the *SF1* gene in *C. elegans* modifies tos-1 splicing, resulting in an increase in lifespan by 15-33%. In Arabidopsis thaliana, AtSF1 (a homolog of human SF1) has been found to regulate temperature-responsive flowering through alternative splicing. The loss-of-function of AtSF1 mutant has shown temperature insensitivity and lower expression levels of FLM-β transcript (Mazroui et al., 1999) (Godavarthi et al., 2020) (Heintz et al., 2017) (Lee et al., 2020) (Ishizuka et al., 2014) (Angela Crisci et al., 2015).

3.8 Phosphorylation of SF1 and SR proteins

The splicing process requires precise regulation through post-translational modifications, specifically phosphorylation and methylation. Certain SR (serine/arginine-rich) proteins undergo reversible phosphorylation and de-phosphorylation, which is crucial for the assembly

of snRNPs and the splicing cycle. These SR proteins facilitate the interactions with other splicing factors and contribute to fine-tuning regulation of the splicing cycle. The phosphorylation of these SR proteins is carried out by several known serine-arginine protein kinase (SRPK) family proteins, such as SRPK1 and SRPK2. Also, multiple phosphorylation sites are found throughout the RS domain of SR proteins and are later dephosphorylated by phosphatase. Also, cyclin-dependent kinase-2 has been reported to target SF3B155, a U2 snRNP component.

In addition, SF1 has three positions that undergo post-translational modification as phosphorylation which are regulated by different kinases. As per the previous literature, the phosphorylated SF1 level was reported in high abundance in prostate cancer patients (Myung & Sadar, 2012). cGMP-dependent protein kinase-1 (PKG-1) phosphorylates the SF1 at serine 20, which impairs U2AF2 interactions (Wang et al., 1999). SF1 also has a conserved "RSPSP" sequence motif on the HH domain which is mainly targeted by KIS kinase and phosphorylates both serines. NMR (Zhang et al., 2012) and crystal study of phosphorylated SF1 (Wang et al., 2013) suggest that phosphorylation of the “RSPSP” motif reduces the loop flexibility at sub-nanosecond timescales and forms a stable salt bridge with neighboring positively charged arginines (**Figure 3.8**). As per previous literature, the phosphorylation of SF1 weakly improves its binding with consensus, optimal, and sub-optimal branch point site RNAs (Long et al., 2019) (Manceau et al., 2006) (Wang et al., 2013) (Zhang et al., 2012) (Lipp et al., 2015). However, it is worth noting that these phosphorylation sites are located far from the RNA binding regions of SF1. Therefore, the specific function and impact of SF1 phosphorylation in splicing remain unclear.

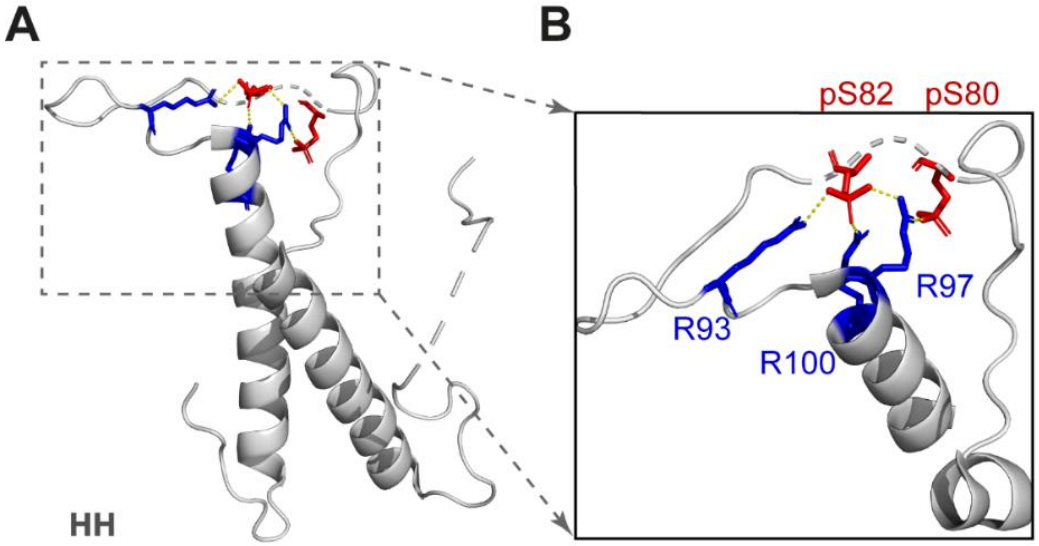


Figure 3.8. SF1 structure with phosphorylated serines. (A) Phosphorylated serines are located on the “RSPSP” conserved motif onto the loop between two helices. (B) Zoom view is shown where phosphorylated serines form the salt bridge with nearby arginines (PDB accession code 4FXW).

3.9 Domain structures of SF1 and U2AF2

From a structural point of view, high-resolution individual sub-domain structures for SF1 and U2AF2 are available in the free or ligand-bound form, as shown in **Figure 3.9**. In U2AF2, its RRM1,2 solution structure in free form acts as a tandem domain, with a dynamic linker connecting the two domains on a sub-nanosecond timescale. The linker's C-terminal residues dynamically interact with RRM2 and exhibit an auto-inhibitory role for the respective RNA targets. A crystal structure of tandem RRM1,2 with polypyrimidine tract (U9) RNA suggests a role of the linker in RNA interactions, which is, however, not well supported by solution NMR data (Kang et al., 2020; Mackereth et al., 2011). In the presence of U9 RNA, the tandem RRM1,2 of U2AF2 adopt an open conformation (Agrawal et al., 2016; Mackereth et al., 2011). NMR solution data have shown that the tandem RRM domains adopt a close arrangement in the absence of RNA, which also samples to some extent the open arrangement observed in the RNA-bound complex. A dynamic equilibrium of an ensemble of open to closed conformations (Huang et al., 2014) is shifted toward the open and active state. This active population is increased depending on the “strength” (i.e., binding affinity) of the RNA ligand (Mackereth et al., 2011). U2AF2's C-terminal UHM domain has an RRM-like fold and interacts with the N-terminal 30-residues long ULM peptide of SF1. The structure of UHM bound to SF1 suggests that the tryptophan 22 of the ULM domain is crucial for forming a stable complex. In addition, several hydrophobic residues of the SF1 protein also support the formation of the complex with U2AF2 (Wang et al., 2013) (Zhang et al., 2012). For SF1, the HH domain next to the N-terminal ULM consists of two anti-parallel helices connected by a 26-residue long linker with a conserved "RSPSP" motif. These helices are connected by multiple hydrophobic contacts (Wang et al., 2013). The SF1 KH-Qua2 domain recognizes the branch point site RNA (5’ UACUAAC-3’) (Liu et al., 2001) is located upstream of the polypyrimidine tract where U2AF2 binds.

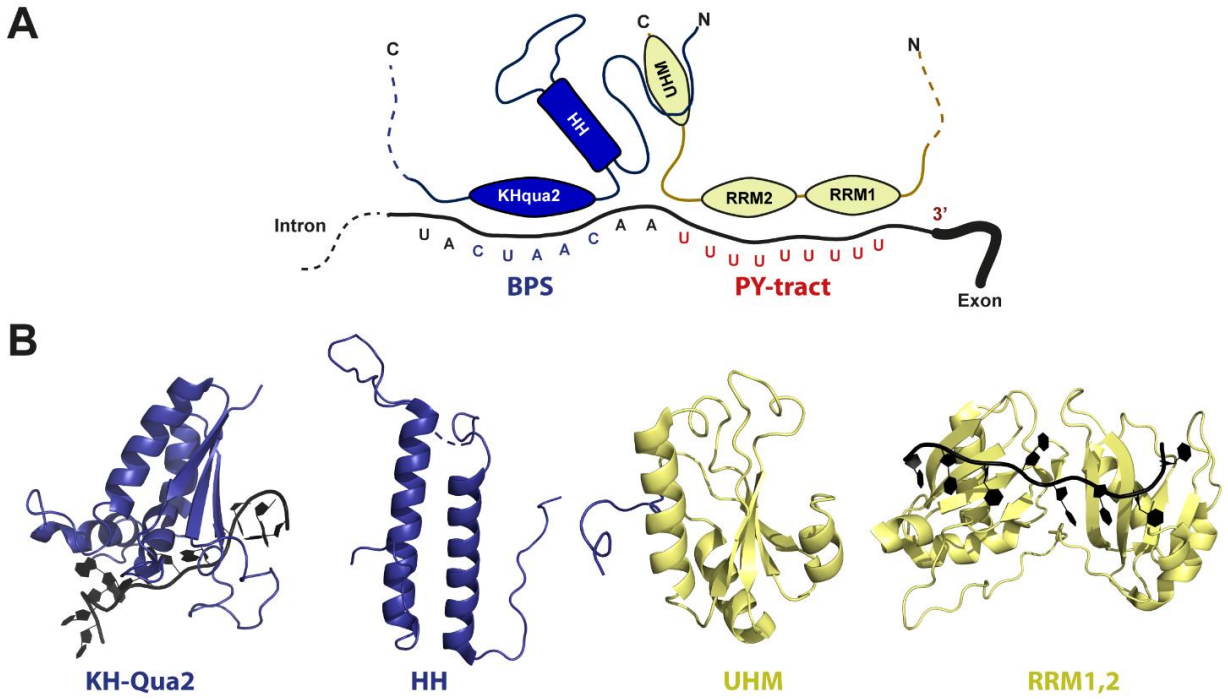


Figure 3.9. Sub-domain structures of SF1-U2AF2 complex. (A) Schematic representation of ternary complex of S1-U2AF2-RNA is shown (B) Individual domain structures are shown. The PDB accession codes for KH-Qua2, HH, UHM-ULM, RRM1,2 are 1K1G, 4FXX, 1OPI and 5EV1, respectively.

While there are high-resolution structural data available for most domains of SF1 and U2AF2, there is a lack of information on the full-length structures of SF1 and U2AF2 proteins in the abscence and bound to intron RNA. A high-resolution structure of the SF1-U2AF2 complex in both RNA-free and RNA-bound states has not been extensively characterized. While recent publications have presented high-resolution cryoEM structures of the yeast complex E, where SF1-U2AF homologo BBP-Mud2 are bound to the 3' splice site (Li et al., 2019), they do not show electron density for BBP (SF1 homolog) and Mud2 (U2AF2 homolog) proteins, which are crucial for understanding the complex's dynamic behavior in the early stage of splicing assembly. Small-angle X-ray scattering (SAXS) data suggest that the functional and minimal version of the SF1-U2AF2 complex undergoes open-to-close conformational changes in the abscence and presence of RNA. However, it still lacks high-resolution structural insights and an understanding of the 3' splice site recognition mode. Therefore, integration of various solution state biophysical methods are necessary to comprehend the complex's dynamic behavior in solution.

Results

3.10 Specificity of branch point site (BPS) recognition by SF1

Understanding the structure and molecular interactions of the recognition of splice sites by splicing factors is crucial for initiating the splicing cycle. Pre-mRNA sequences contain diverse and complex splice sites on the intron, with the sequence conservation of BPS and PPT splice sites varying between intron-poor and intron-rich species. In intron-poor species, the consensus BPS sequences "UACUAAC" are fully conserved. In contrast, BPS in intron-rich species are highly degenerative and represented by "YNYURAYNN" motif, where N, Y, and R stand for any nucleotide, pyrimidine, and purine, respectively. However, both intron-poor and intron-rich species have fully conserved canonical BPS motifs, with "U" +2 upstream and branch point site "A" (**Figure 3.10 A**). The BPS recognition protein SF1 has a well-conserved K-homology domain (KH) and Qua2 domain across the different species (**Figure 3.10 B)**. In humans, the mutation at the branch site "A" or the +2 upstream "U" cause severe splicing defects. However, the specific disease-associated BPS motifs and how they interact with SF1, and their structural details have not been thoroughly studied.

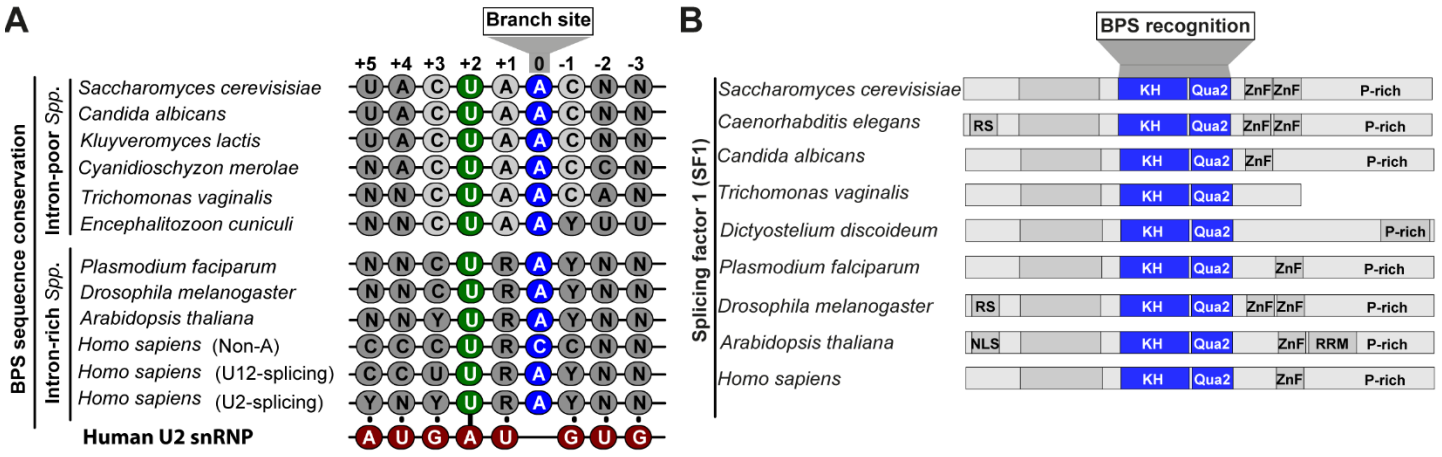


Figure 3.10. Branch point site RNA and SF1 domain sequence conservation. (A) Branch point site sequence alignment with intron poor and intron rich species are highlighted (Irimia & Roy, 2008) (Zhang et al., 2022) (B) SF1 protein domain conservation across the different species is shown.

3.11 SF1's interaction with variations of the BPS RNA

To study BPS recognition by SF1, a range of short oligo-nucleotide sequences were

designed. The optimized yeast BPS oligo-sequence "TACTAACAA" was selected as a control for NMR binding study, and BPS variants were designed by altering the T(U) to C, G, or A nucleotide at the +2 upstream of the branch site (as shown in **Figure 3.11**). Single-stranded DNA oligos were used as a proxy for RNA to screen a large number of sequences. For that, NMR binding experiments were carried out with SF1's KH-Qua2 domain with different oligo sequences (**Figure 3.12 A**). NMR titrations of KH-Qua2 with complementary BPS sequence showed maximum chemical shift perturbations (CSP) at the "GxxG" loop, β2-sheet, α1 and α2 helices of the KH domain, and helix of Qua2 domain, which aligns with previously published results. Next, a KH-Qua2 NMR binding experiment was performed with various oligo motifs, as shown in **Figure 3.11 A**. The results indicate that oligos with T>A, T>G, or T>C nucleotide mutations at the +2 position have similar CSP to the KH domain. However, there was a significant reduction in CSP observed in the α-helix region of the Qua2 domain when compared to optimized-BPS oligo with "T" at the +2 position (as shown in **Figure 3.12 B, C**). This suggests that the "T" at the +2 position is highly selective for the Qua2 domain, and oligos with "T" at +2 position and branch site "A" are the most preferred sites for KH and Qua2 in recognizing highly degenerative BPS sites. Moreover, the "T(U)" at +2 upstream on BPS remains evolutionarily conserved across different species (**Figure 3.10 A)**. However, changing the +3 position C>T does not affect the overall CSPs for KH and Qua2 domain binding (**Figure 3.12 A, B, C bottom panel**). When the branch site is mutated from A>C, KH binding shifts to the +1 "A" position, but no change in NMR shift was observed for the Qua2 domain (**Figure 3.12 A, B, C bottom panel**).

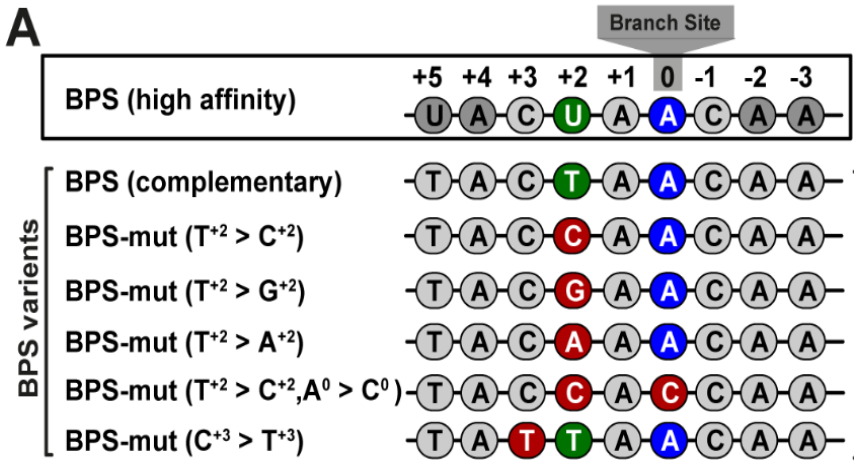
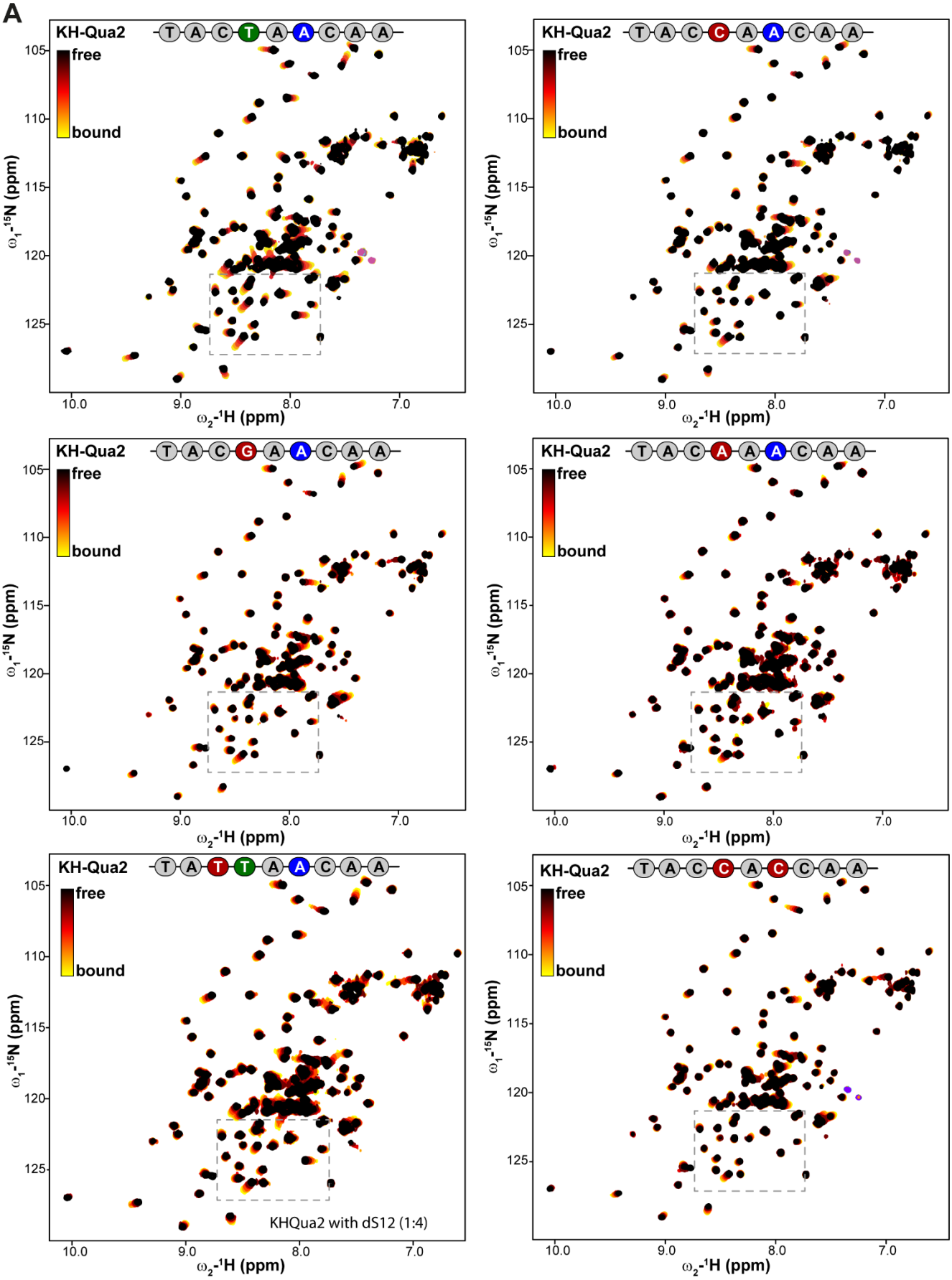


Figure 3.11. Designing of BPS sequence variant for SF1 binding study. (A) BPS optimized sequence is shown on top with relative position from branch site "A". Complementary and BPS variant sequences are shown. Branch site "A", +2 position from branch site, and point mutants are colored in blue, green, and maroon, respectively.


A
KH-Qua2
T A C T A A C A A
free
bound
105
110
115
120
125
ω_1-^{15}N (ppm)
10.0
9.0
8.0
7.0
ω_2-^{1}H (ppm)
KH-Qua2
T A C C A A C A A
free
bound
ω_1-^{15}N (ppm)
ω_2-^{1}H (ppm)
KH-Qua2
T A C G A A C A A
free
bound
ω_1-^{15}N (ppm)
ω_2-^{1}H (ppm)
KH-Qua2
T A C A A A C A A
free
bound
ω_1-^{15}N (ppm)
ω_2-^{1}H (ppm)
KH-Qua2
T A T T A A C A A
free
bound
ω_1-^{15}N (ppm)
ω_2-^{1}H (ppm)
KHQua2 with dS12 (1:4)
KH-Qua2
T A C C A C C A A
free
bound
ω_1-^{15}N (ppm)
ω_2-^{1}H (ppm)

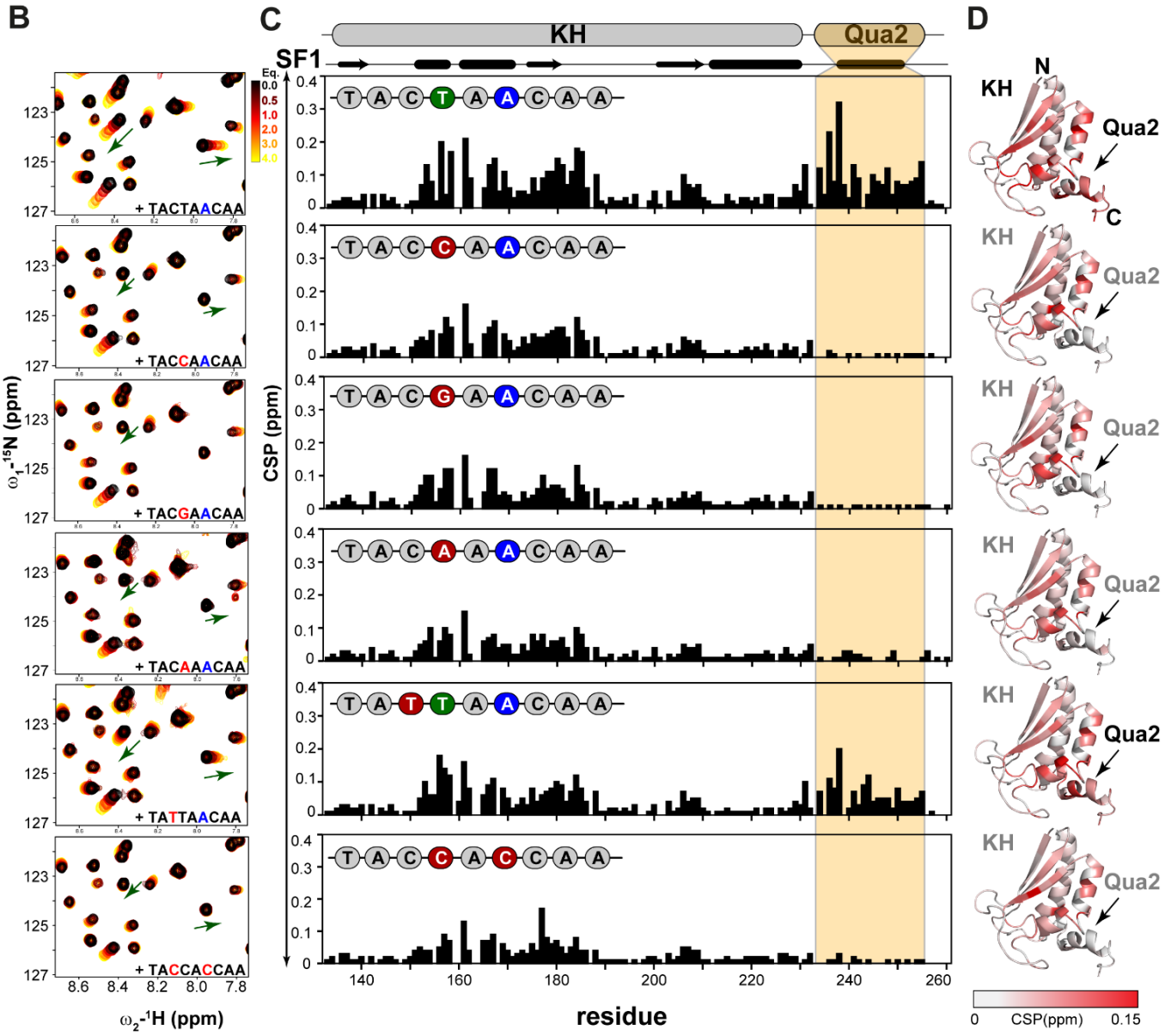


Figure 3.12. NMR binding study of SF1's KH-Qua2 domain with mutated branch point sites. (A) Overlay 1H-15N HSQC spectra of KH-Qua2 domain with increasing concentration of BPS oligo, colored as black, maroon, red, orange, and yellow, respectively. (B) Zoom view of overlay spectra with the series of titration points. (C) Chemical shift perturbation (CSP) plot between free (1:0) and bound (1:4 molar ratio) for different BPS mutants. (D) Mapping of BPS binding surface area on KH-Qua2 structure, colored from gray (no shift) to red (maximum shift).

3.12 SF1 has reduced interactions with disease-associated BPS motifs

Recognizing BPS splice site is the first step in initiating the splicing cycle, while mutations in such a caonical BPS sites can result in human diseases. For example, Fish Eye Disease is caused by a T>C mutation at the +2 upstream of the BPS site "CCCT/CG**A**CCC," leading to

the complete retention of intron-4 on the LCAT (lecithin: cholesterol acyltransferase) gene. Another example is an extrapyramidal movement disorder, where a T>A mutation at the +2 upstream of BPS on intron-11 of the tyrosine hydrolase (TH) gene causes alternative splicing using nearby cryptic splice. Similarly, a T>G mutation at the +2 upstream branch site of intron-32 on the COL5A1 (collagen type V α-1 chain) gene leads to partial exon-33 skipping and causes Ehler-danlos syndrome type II in humans (Burrows et al., 1998). Epidermolysis-bullosa with pyloric atresia disease is caused by a point mutation T>A at +2 upstream of the branch site on the ITGB4 gene, encoding integrin β4 protein, resulting in either complete retention of intron-14 or intron-31 (Masunaga et al., 2015). In Hypertriglyceridemia disease, a point mutation A>G at the branch site of intron-1 of the LIPC gene results in complete retention of intron-1, which encodes the lipolytic serine hydrolase enzyme (Brand et al., 1996). Additionally, the branch site on the FBN2 gene that encodes fibrillin-2 is mutated from A>G in congenital contractual arachnodactyly (Beals) syndrome, leading to a splicing defect of exon-32 (Gupta et al., 2004). A comprehensive list of these mutations and associated diseases are shown in **Figure 3.13**. To gain a better understanding of SF1 interaction with disease-associated BPS, NMR titration experiments were conducted.

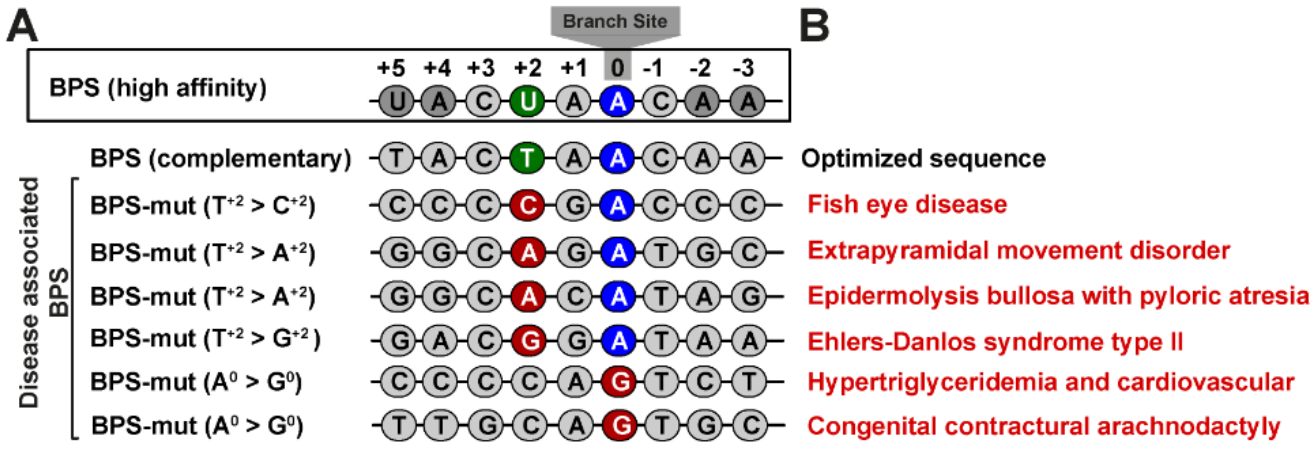


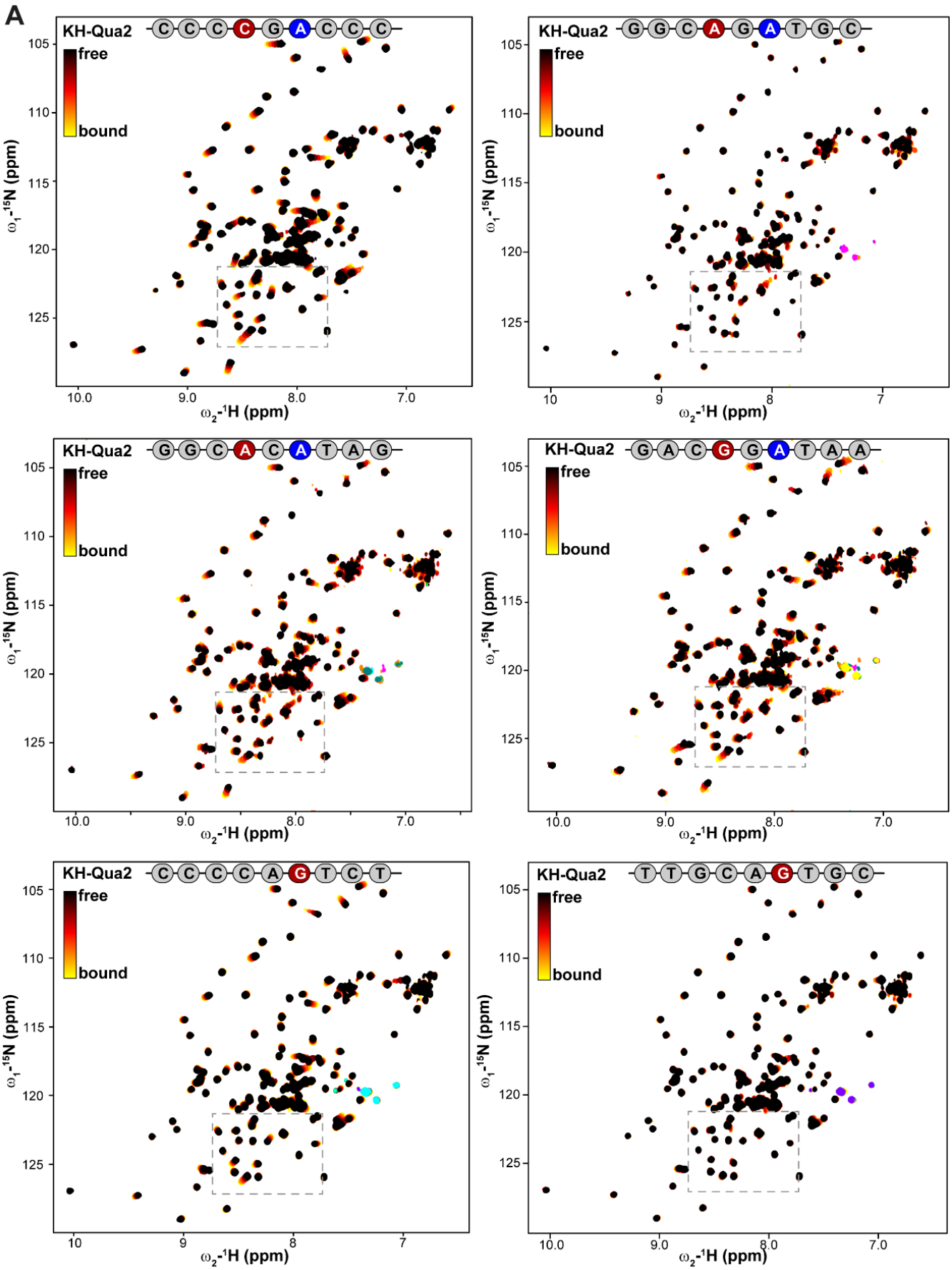
Figure 3.13. Disease-associated BPS mutant sequences. (A) BPS sequences are shown where mutation position highlighted in maroon color. (B) Human diseases associated with BPS mutation are highlighted in red.

In order to investigate how SF1 interacts with disease-associated BPS sequences, the short 9mer ssDNA oligos were designed as listed in **Figure 3.13**. NMR binding experiments were performed using ^{15}N labeled KH-Qua2 construct of SF1. For that, ^{1}H-^{15}N HSQC spectra

were recorded for the KH-Qua2 domain in free and in the presence of 4-fold access of ssDNA oligos (**Figure 3.14 A, B**) and analyzed.

Our analysis indicates that the BPS with mutations in the +2 upstream regions of LCAT and COL5A1 genes exhibit NMR CSPs similar to the optimized BPS for the KH domain. However, no noticeable CSPs are observed for the Qua2 helix (**Figure 3.14 B, C upper three panels**). Similarly, the BPS disease variant for TH and ITGB4 genes show reduced CSP for the KH domain, while no CSP was observed for the Qua2 helix compared to the optimal BPS (**Figure 3.14 B, C upper four and fifth panels**). The mutant branch site "A" of the LIPC gene showed reduced and no observable CSP for the KH and Qua2 domains, respectively. Likewise, the branch site mutant of the FBN2 gene showed no significant CSP for the KH and Qua2 helix (**Figure 3.14 B, C lowere two panels**).

To summarize, the mutations related to the disease in BPS sites have a significant impact on the interaction with the C-terminal Qua2 helix. On the other hand, the KH domain has a similar or weaker binding to the tested sequences. This suggests that the KH domain provides support for binding to BPS sequences, while the Qua2 helix is responsible for selectivity in binding to specific BPS sites. Therefore, any changes in the +2 upstream site "T (U)" or the branch site "A" lead to a loss of selectivity for SF1 in recognizing disease-causing BPS sites. This results in an abnormal splice product of mRNA transcript during the subsequent steps of splicing.


A
KH-Qua2
C C C C G A C C C
free
bound
105
110
115
120
125
ω_1-^{15}N (ppm)
10.0
9.0
8.0
7.0
ω_2-^{1}H (ppm)
KH-Qua2
G G C A G A T G C
free
bound
ω_1-^{15}N (ppm)
10
9
8
7
ω_2-^{1}H (ppm)
KH-Qua2
G G C A C A T A G
free
bound
ω_1-^{15}N (ppm)
ω_2-^{1}H (ppm)
KH-Qua2
G A C G G A T A A
free
bound
ω_1-^{15}N (ppm)
ω_2-^{1}H (ppm)
KH-Qua2
C C C C A G T C T
free
bound
ω_1-^{15}N (ppm)
10
9
8
7
ω_2-^{1}H (ppm)
KH-Qua2
T T G C A G T G C
free
bound
ω_1-^{15}N (ppm)
ω_2-^{1}H (ppm)

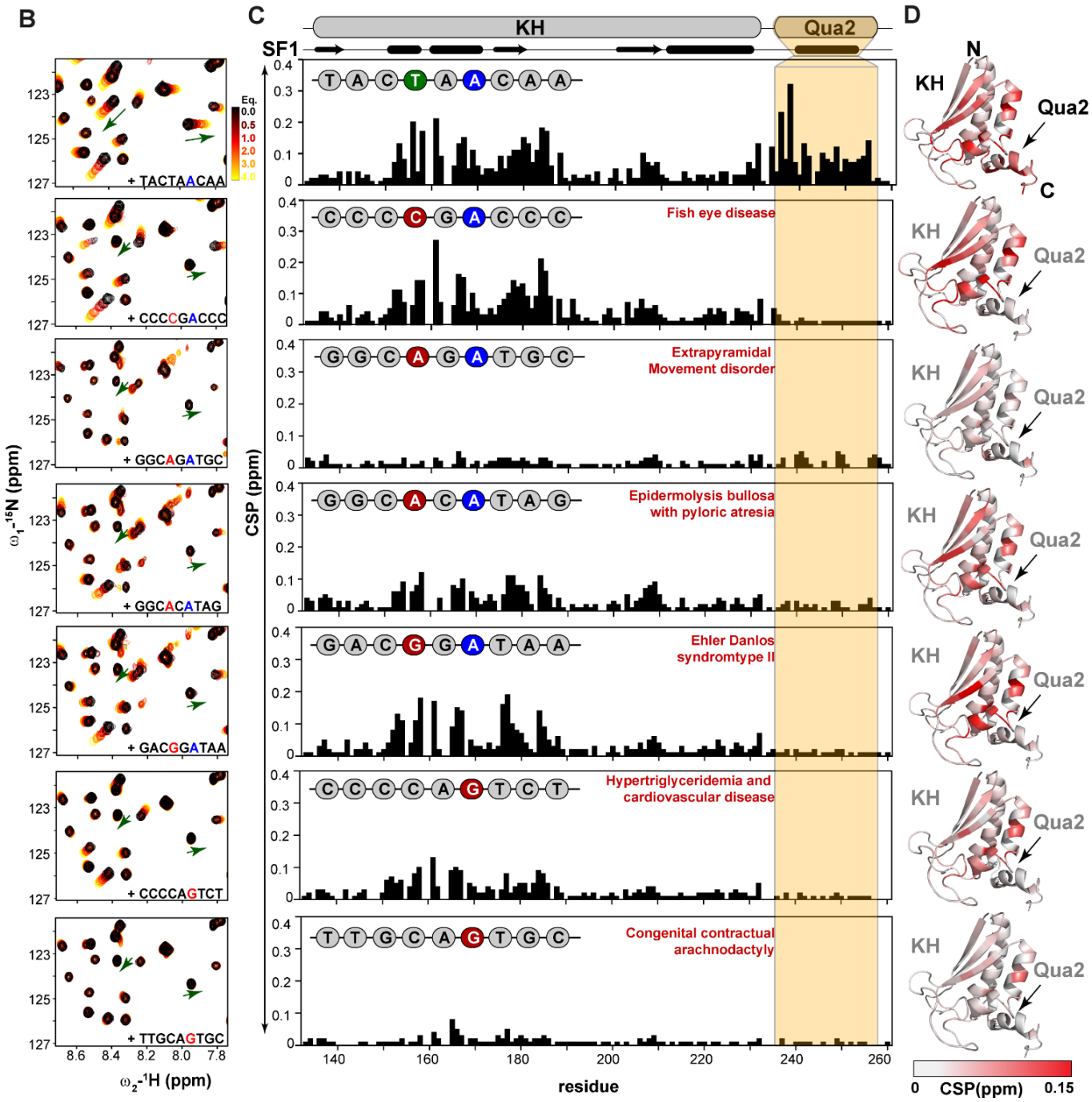


Figure 3.14. NMR binding study of KH-Qua2 domain with disease-associated branch point site sequences. (A) Overlay ^{1}H-^{15}N HSQC spectra of KH-Qua2 in free with increasing concentration of oligos are colored as black (free), red (intermediate) to yellow (bound) gradient. (B) Chemical shift perturbations plotted between free and bound for respective BPS. (C) Mapping of BPS binding surface area on KH-Qua2 structure, colored from gray (no shift) to red (maximum shift).

3.13 SF1's Qua2 is dynamic with respect to the KH domain

In SF1, KH and Qua2 domains recognize the branch point sites of introns. According to the previously published structure, it has been shown that the Qua2 helix tends to adopt a closely packed conformation in the presence of optimized RNA bound to KH-Qua2 (Liu et al., 2001). To investigate the impact of flexibility on RNA recognition, the solution NMR method was utilized to probe the fast-motion dynamics of KH-Qua2. The analysis from the NMR-based $\{^1H\}$-^{15}N hetero-nuclear NOE experiment suggests that the variable loop in KH and Qua2 exhibit high flexibility on a sub-nanosecond time scale in the RNA-free state (**Figure 3.15**, **Figure 3.16**). Additionally, the total correlation time (τ_c) calculated from longitudinal and transverse relaxation experiments indicates distinct average local correlation times, $\tau_C \sim 9$ ns and $\tau_C \sim 6$ ns, for the KH and Qua2 regions, respectively (**Figure 3.15**).

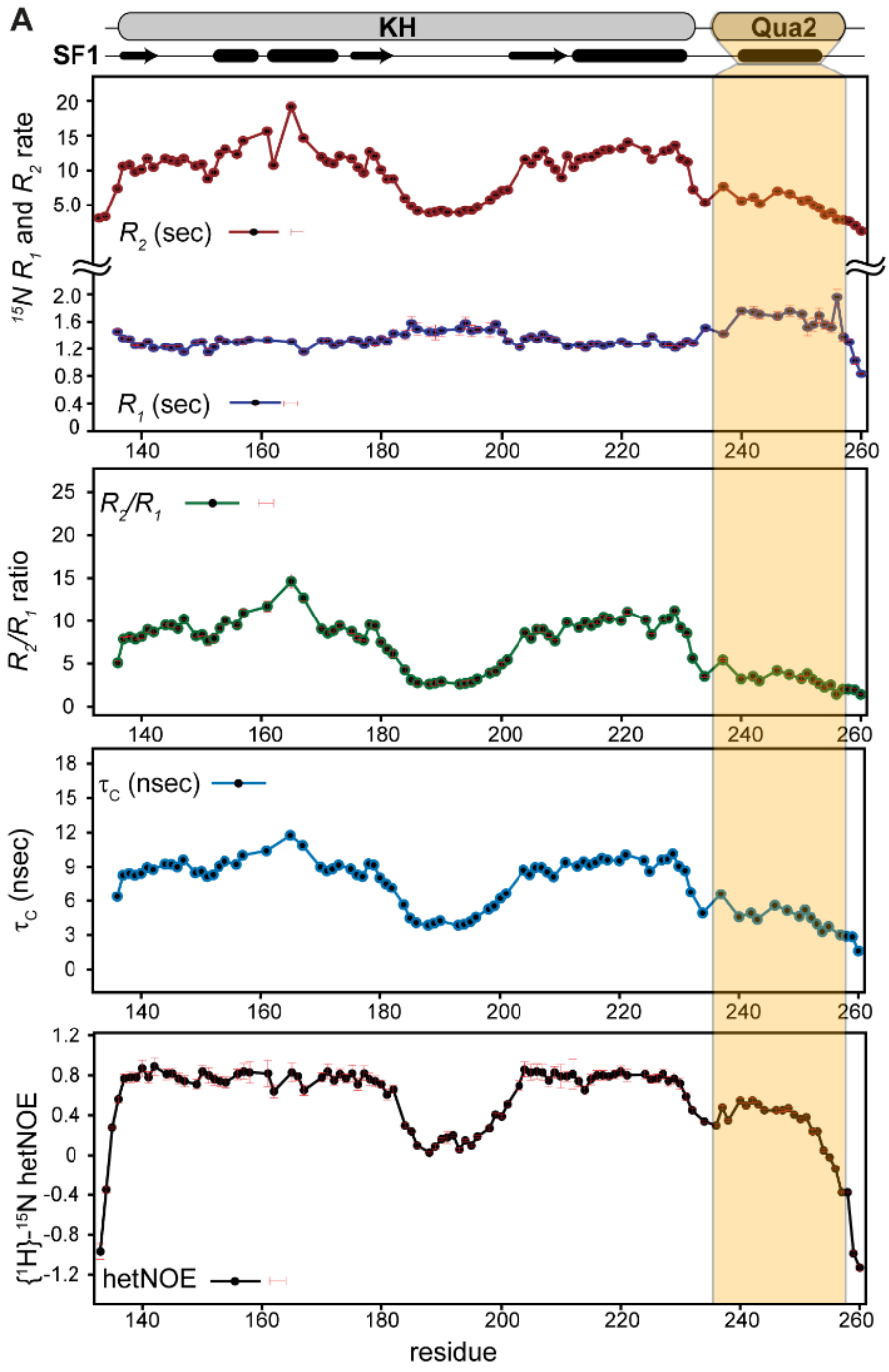


Figure 3.15. Fast scale dynamics measurements for KH and Qua2 domains of SF1. (A)(B) ^{15}N R_1 and R_2 rates are measured at 600MHz spectrometer and R_2/R_1 ratio is extracted. (C) Total correlation time for each residue was calculated from R_1 and R_2 rates. (D) $\{^1H\}$-^{15}N heteronuclear NOE experiment is shown. The Qua2 domain is highlighted in yellow.

Due to the flexible nature of the Qua2 domain, an ensemble optimization method (EOM) was implemented to derive the ensemble model of KH-Qua2 consistent with experimental SAXS data. EOM analysis revealed that around 85% of the selected structures in the ensembles have a radius of gyration (R_g) of 20 Å, while the remaining 15% have an R_g of 25 Å. This was out of the 10000 randomized pool of unbiased structures. The theoretical SAXS curve generated from the ensemble fit well with the experimental data, with a χ^2 of 1.3 (**Figure 3.16 B, C**). This indicates that the flexible Qua2 helix of SF1 is rather flexible and not well packed against the KH domain in the RNA-free state compared to the RNA-bound state (PDB 1K1G). This is consistent with a previous analysis of the SF1 homolog from *Xenopus laevis* (Maguire et al., 2005). Such flexibility of Qua2 is presumably important to enabling binding to diverse BPS RNA motifs. Additionally, in the extended version of $SF1^{1-260}$, including ULM, HH and KH-Qua2 domains, the ULM, linker connecting two helices of HH, variable loop on KH and Qua2 domains showed high flexibility at a sub-nano-second time scale, while KH and HH domains stayed rigid in solution (**Figure 3.17 A**).

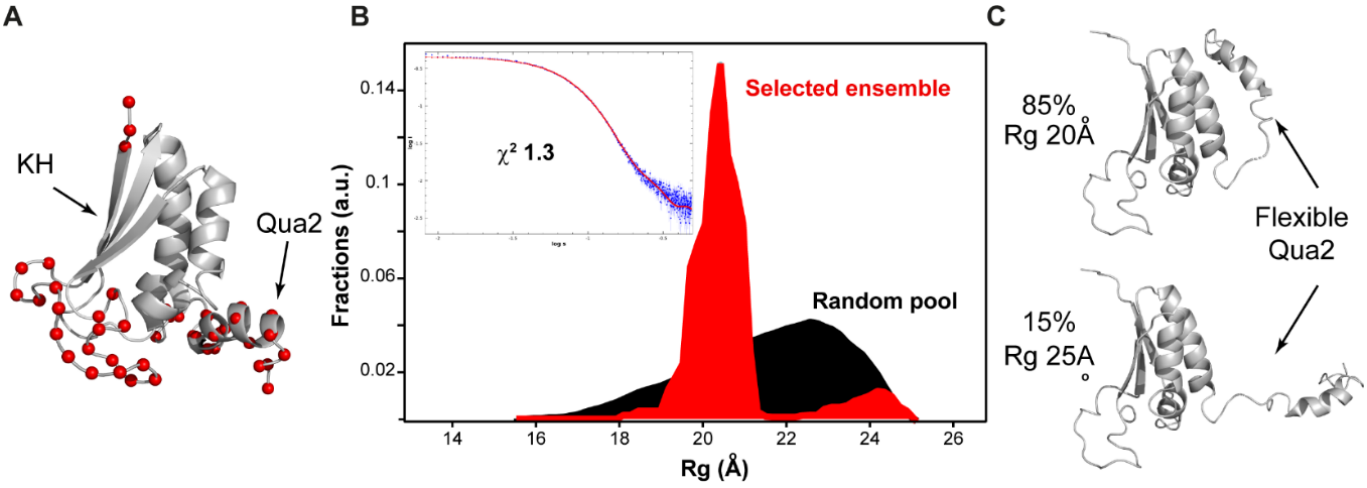


Figure 3.16. Ensemble structure description of KH-Qua2 domain. (A) Flexible residues (based on hetNOE) of the KH-Qua2 domain are highlighted with red spheres. (B) R_g distribution of the random pool and a selected ensemble of structures are shown in black and red, respectively. Experimental and theoretical SAXS fit curves for selected ensemble structures are shown in blue and red. (C) The population of selected ensemble models is shown.

3.14 Structure of N-terminal segment of SF1

The individual domain structures of SF1 have been previously identified, but information about how these domains are oriented relative to each other was missing. To address this, paramagnetic relaxation enhancement (PRE) NMR experiments were conducted where a

single cysteine residue was introduced at different positions of KH and HH domains and attached nitroxide spin labels. The PRE data were analyzed to derive long-range intra- and inter-domain distance restrictions (**Figure 3.17**), which were then used for semi-rigid body structure calculation of the complete N-terminal region of SF1, comprising residues 1 to 260 aa. Experimental details for the data analysis and structure calculation are described in the methods section (chapter 2, section 2.28, 2.29). Heteronuclear NOE analysis showed that the HH and KH domains are rigid, while the ULM, Qua2, and connecting loops exhibit sub-nanosecond scale flexibility. Therefore, cysteine positions for spin labeling were selected based on the rigid region of the HH and KH domain for structure refinement.

To determine the structure of SF1, PRE-derived inter-domain restraints were utilized to determine the structure of SF1 through a rigid body refinement protocol (see method section for more information). After refinement, the 20 lowest energy structures were selected and further analyzed. The derived structure indicates that the HH domain remains close to the KH domain, while the flexible ULM and Qua2 domains are free in solution. The derived structures align well with experimental PRE data, with a quality factor (Q-factor) of 2.5 and a 2 Å root mean square deviation between HH and KH domains (as depicted in **Figure 3.18 A**). The Q-factor has good agreement with 47C, 55C, 122C, 137C, 172C and 213C PRE datasets, however, back-calculation PRE curves for 62C and 107C show slightly different compared to the experimental data.

ULM and Qua2 helix also show high flexibility as demonstrated by $\{^{1}H\}$-^{15}N hetNOE. Hence, SAXS-based EOM method was utilized for the ensemble description of $SF1^{1-260}$ (as shown in **Figure 3.18 B**). The selected EOM ensemble structures suggest that SF1 adopts extended conformations with respect to the pool of the randomly generated structures, with a radius of gyration ranging from 29 to 31 Å compared to the randomized pool of structures. This is mainly due to the high flexibility of the N-terminal ULM peptide and the C-terminal Qua2 helix, while, the HH and KH domains have reduced flexibility, as indicated by PRE.

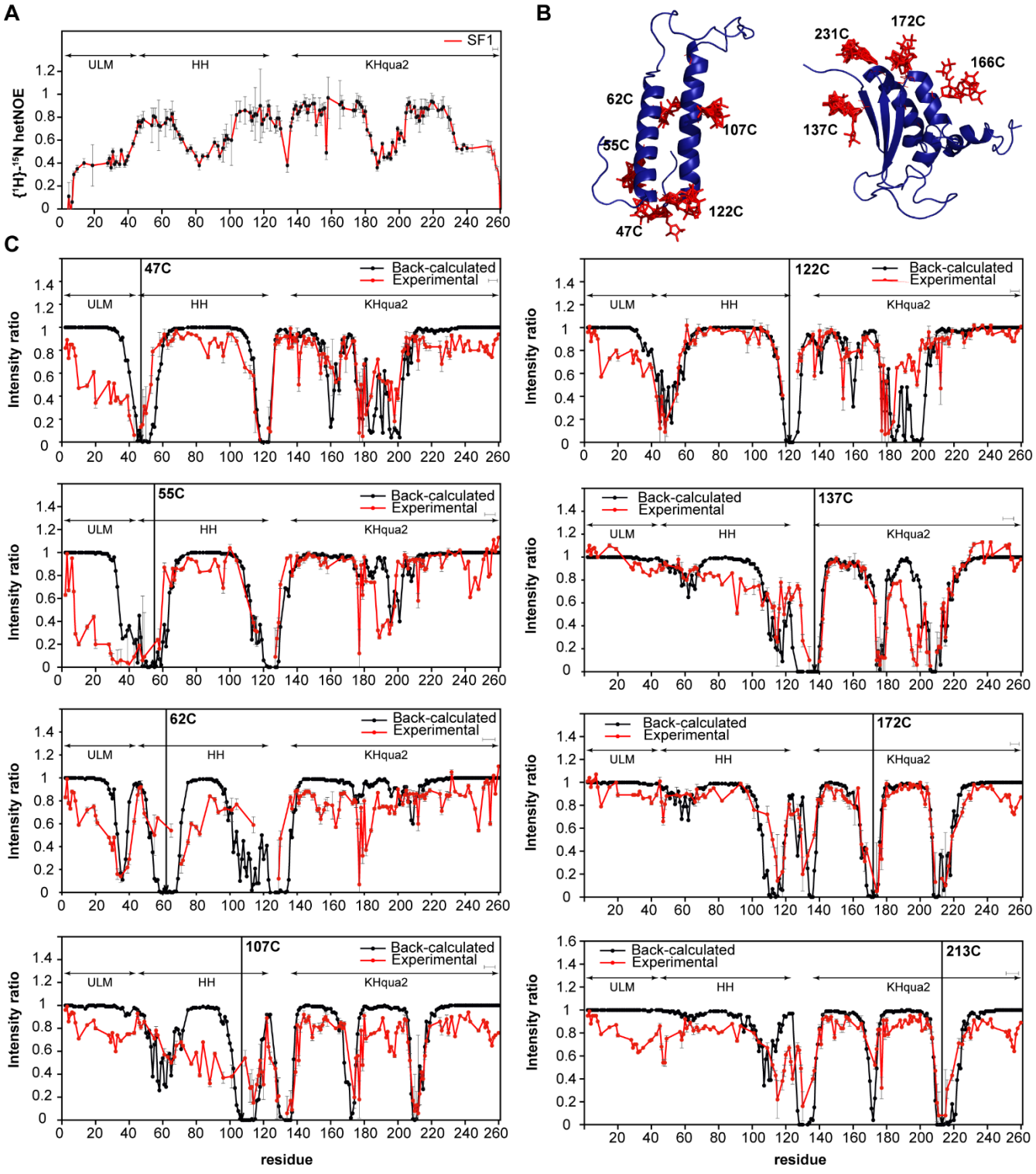


Figure 3.17. Paramagnetic relaxation enhancement (PRE) measurement for SF1. (A) {^{1}H}-^{15}N heteronuclear NOE plot for the SF1 (1-260 aa) is shown. (B) The position of cysteine mutations on HH (PDB 4fxx) and KH-Qua2 (PDB 1K1G) domains are shown with sticks. (C) Intensity ratio (of oxidized and reduced sample spectra) plots are shown. Red and black are experimental PRE and back-calculated PRE (see next session) plots, respectively.

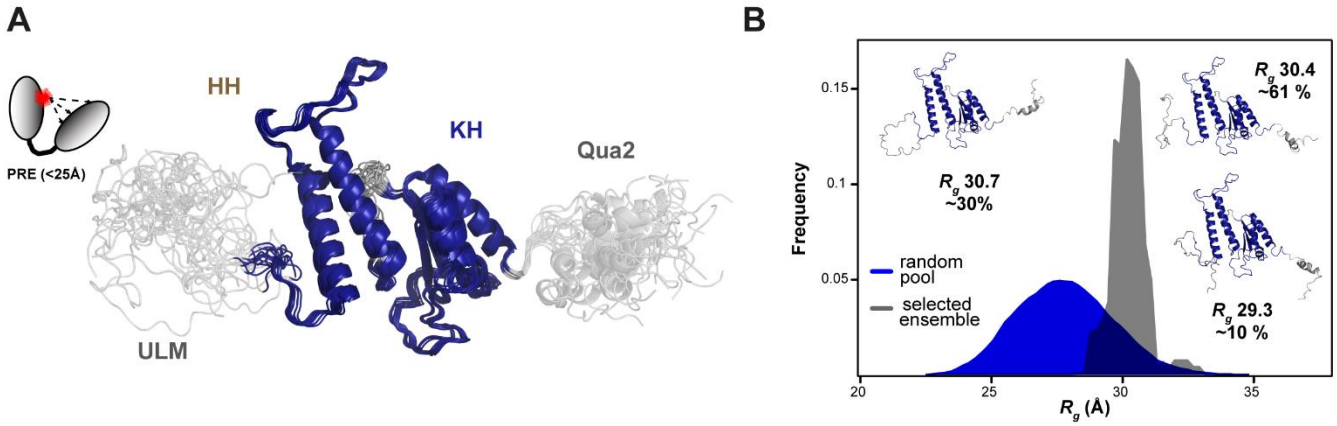


Figure 3.18. Ensemble structures of SF1. (A) PRE-derived overlay of 15 ensemble structures of SF1 are shown, where HH, KH domains in dark blue and ULM, Qua2 domains in gray. (B) Different populations of SF1 models are shown derived from SAXS. Based EOM method. The probability of R_g distribution plots are shown where black and red are randomized pools and selected ensemble structures are shown.

Table 3.1. Structural statistics of SF1 (1-260 aa)

Structure statistics of PRE analysis (rigid-body refinement)	
SF1 (N-terminal fragment)	ULM, HH, KH and Qua2 from 1 to 260 aa
Ensemble structure	20 structures
Intra domain restraints	133
Inter domain restraints	107
Distance violations	< 4 Å
RMSD (20 aligned structures)	2.3 Å (for 48-228 aa)
Quality factor	0.35
Ramachandran plot	90.7 % most favoured region 6.20 % additioanlly allowed regions 2.10 % generously allowed regions 1.00 % disallowed region
EOM analysis (SAXS)	
Number of structures in Random pool	10000
Selected ensembles	30.3 %, R_g 30.7 Å 60.6 %, R_g 30.4 Å 10.1 %, R_g 29.3 Å
Rflex/*Rσ*	59.26 % (~ 87.18%) / 0.32
SAXS fitting	χ^2 3.9

3.15 Large scale RNA binding analysis of SF1 and U2AF2

To better understand SF1's role in the recognition of canonical 3' splice sites, next, *in vitro* iCLIP (individual-nucleotide resolution cross-linking and immuno-precipitation) experiments were performed with U2AF2 in the absent and presence of SF1 using a gene library containing various types of branch point (BPS) and polypyrimidine tract (PPT) sites. The data indicates that U2AF2 alone binds to PPT sites nearby 3' splice sites as expected (**Figure 3.19 control panel**). However, significant U2AF2 binding at PPT was observed in the presence of SF1, which suggests that SF1 stabilizes U2AF2, allowing it to recognize canonical 3' splice sites as shown in **Figure 3.19 A, B**. Similarly, U2AF2 stabilization effect upon SF1 was observed while categorizing splice sites with varying strengths, ranging from strong to weak PPT and BPS sites. The overall finding indicates that SF1 plays a crucial role in stabilizing U2AF2 for recognizing a range of 3' splice sites.

According to the literature, SF1 has two phosphorylation sites on a loop connecting two helices of the HH domain. These sites are on a conserved "RSGSG" motif, and serine phosphorylation has been shown to weakly facilitate RNA binding with sub-optimal branch sites (Lipp et al., 2015). However, the specific role of SF1 phosphorylation with U2AF2 and the variable strength of splice sites has not been fully explored. To investigate this, iCLIP experiments were conducted using a gene library with varying BPS and PPT splice site strength, both in the presence of non-phosphorylated and phosphorylated SF1. The analysis suggests that the SF1-U2AF2 complex binds similarly to splice sites with varying strength, regardless of non-phosphorylated or phosphorylated SF1 (as shown in **Figure 3.19 A, B and Figure 3.20 A, B, C**).

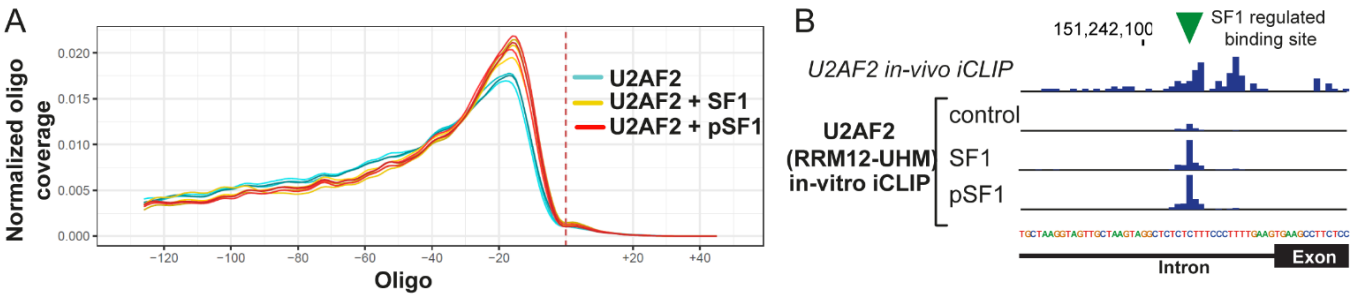


Figure 3.19. U2AF2 in-vitro iCLIP analysis for 3' splice site recognition. (A) Overall RNA-map oligonucleotide profiles for U2AF2 alone and with SF1 complex are shown. (B) The relative intensity profile of oligo binding is shown for all oligos (Experiments were performed and analyzed by Stefani Ebersberger and Julian König, IMB Mainz).

NMR ^{31}P experiments of phosphorylated SF1 showed no significant shifts, except for a weak peak shift for one of the phosphorylated serines, indicating possible transient interactions in the presence of U2AF2 and short oligosnucleotide containing optimized BPS and PPT sites (**Figure 3.20 D**). Additionally, SAXS analysis shows a slight increase in compactness upon phosphorylation of SF1, both alone and in complex with U2AF2. However, a similar compactness was observed upon deleting residues on the phosphorylated loop (Δ73-88 aa), suggesting a re-arrangement of the linker (**Figure 3.20 F**). In addition, to study how RNAs of different strengths affect the structure of non- and phosphorylated SF1-U2AF2 complexes, SAXS experiments were conducted. For that, two short RNA oligos were designed: (i) strong PPT RNA (5'- UACUAACAAUUUUUUUUU -3') with an optimized BPS and PPT motif (ii) weak PPT RNA (5' UACUAACAAUAAAAAAAA -3') with an optimized BPS, but weak strength of PPT motif (**Figure 3.20 G**). The SAXS derived pairwise distance distribution plots revealed that both the non-phosphorylated and phosphorylated SF1-U2AF2 complexes bound to strong PPT RNAs were similarly compact. Likewise, both complexes showed identical open conformations with weak PPT RNA. Additionally, the binding affinity by ITC was similar for both phosphorylated and non-phosphorylated SF1 to U2AF2 **(Figure 3.20 E**).

The results of iCLIP, ^{31}P NMR, ITC, and SAXS suggest that phosphorylation of the complex has the equal contribution as the non-phosphorylated complex in the 3' splice site recognition. Therefore, SF1 phosphorylation may play a role in unknown regulatory pathways in splicing.

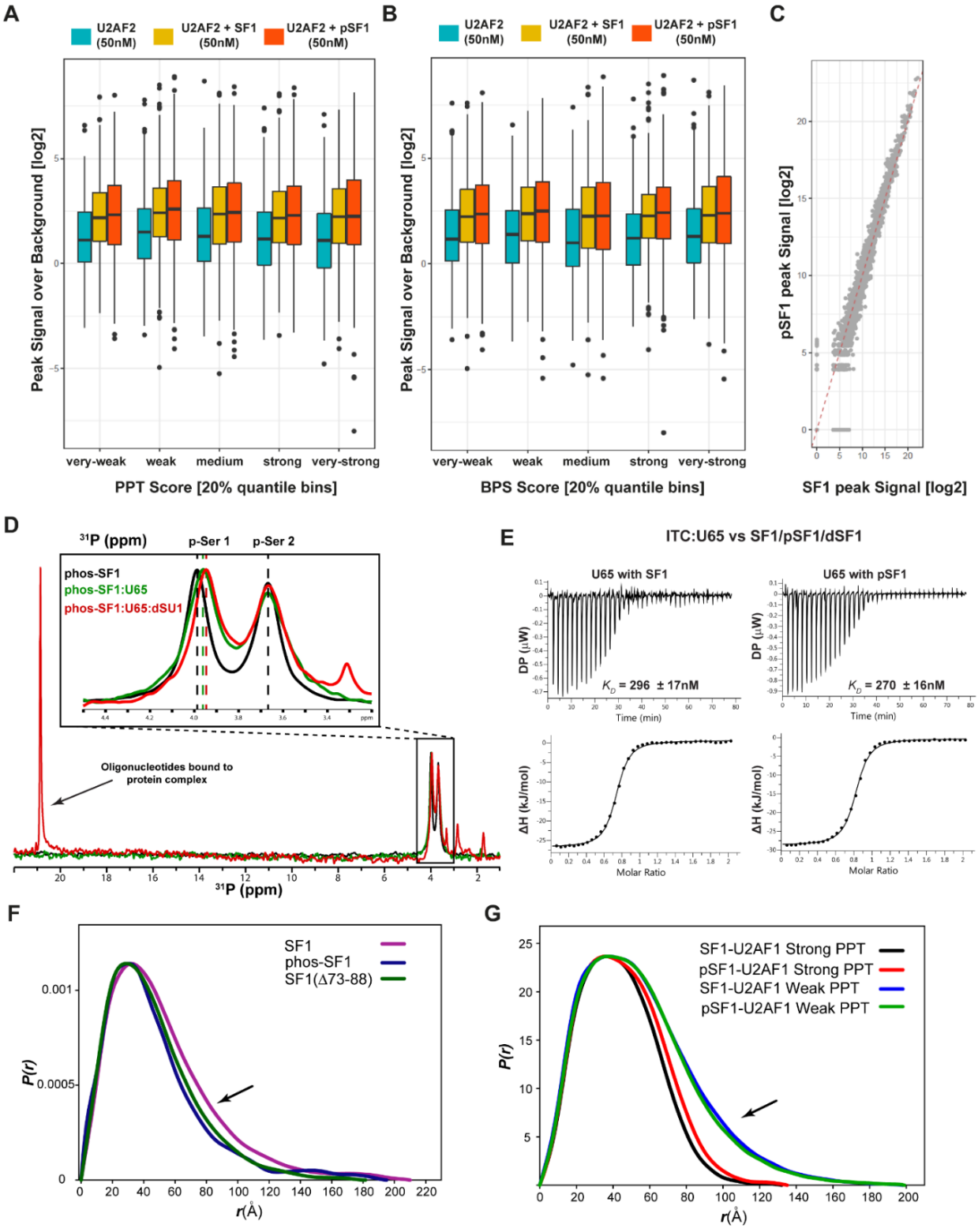


Figure 3.20. iCLIP, ^{31}P NMR and ITC analysis of non- and phosphorylated SF1-U2AF2. iCLIP oligos binding profiles are shown for U2AF2 alone, with non- and phosphorylated SF1 with variable strength of PPT (A) and BPS (B) motifs. (C) The correlation oligonucleotide binding curve is plotted for non- and phosphorylated SF1. (D) ^{31}P NMR spectra for phosphorylated SF1 (black), with U2AF2 (green) and with oligonucleotide (red) are highlighted by zoom area. (E)

ITC curves for U2AF2 without and with phosphorylated SF1 are plotted. (F) Pair distance distribution [*P(r)*] plots for $SF1^{1-260}$, phosphorylated $SF1^{1-260}$ and $SF1^{1-260}$ (Δ73-88) are shown. (G) *P(r)* plots for without and with phosphorylated SF1-U2AF2 complex with RNAs with strong PPT and weak PPT are shown.

3.16 Structural analysis of the SF1-U2AF2 complex with variable RNA ligands

Genome-wide analysis shows the SF1-U2AF2 complex recognizes a wide range of intronic sequences which comprise well diverse strength of branch point sites and polypyrimidine tract motifs. Hence, to understand the structure of the SF1-U2AF2 complex to recognize RNAs having varying strength of splice sites, the SAXS experiments were conducted and extracted biophysical parameters, such as the shape, size, radius of gyration, and relative compactness (**Figure 3.21 A; Table 3.2**). The data analysis revealed that the SF1-U2AF2 complex without RNA has an extended size (D_{max}≈ 200 Å) and ≈ 42 Å radius of gyration (R_g), representing an open conformation (**Figure 3.21 B black curve**), consistent with previous findings (Zhang et al., 2012).

Interestingly, the SF1-U2AF2 complex bound to RNA having BPS^{opt}, $BPS^{sub\text{-}opt}$ or BPS^{mut} with PPT^{opt} has a much smaller D_{max} ≈ 110-130 Å, indicating a more compact conformation compared to the RNA-free complex. In this compact conformation, the RRM1,2 of U2AF2 binds to optimized PPT (“U9”) and positions SF1’s KH-Qua2 to bind to strong or mutated BPS sites located nearby. On the other hand, when binding to RNA with BPS^{alt} along with either PPT^{opt} or PPT^{alt}, the SF1-U2AF2 complex has a D_{max} size of about 18 Å, indicating a semi-compact conformation. In this case, SF1 does not bind to the altered BPS but rather RRM1,2 binds to "U9" or "C9" RNA motif. However, when BPS^{opt} and PPT^{alt} are present, SF1-U2AF2 adapts an open conformation with a size of approximately 200 Å, allowing for interaction with optimized BPS and preventing RRM1,2 of U2AF2 from binding to weak "A9" RNA (**Figure 3.21**). The Kratky plots overlay of RNA-free and various strengths of BPS and PPT with SF1-U2AF2 complex display a unique non-parabolic shape with a non-zero tail at lower and higher scattering angles (q). This represents a dynamic multi-domain protein with a flexible linker (see **Figure 3.21 C; Table 3.2**).

In conclusion, the SF1-U2AF2 complex selects different compact, semi-compact, or open conformations when RNA with variable strengths of BPS and PPT splice sites are added (**Figure 3.21 G**). These dynamic characteristics of multi-conformations of SF1-U2AF2 help to recognize the canonical splice sites that are nearby and initiate the splicing machinery.

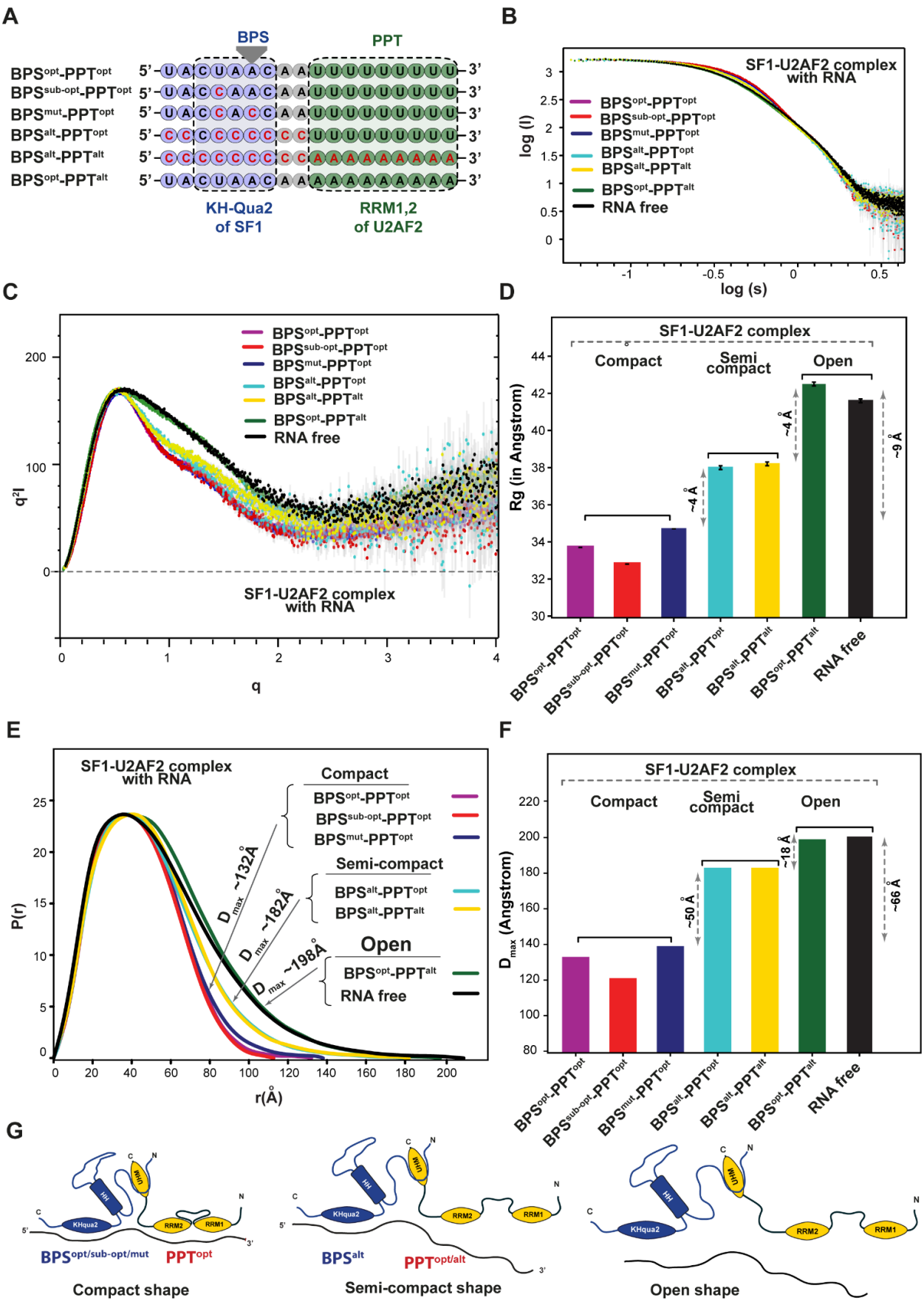

A
BPS
PPT
BPSopt-PPTopt
BPSsub-opt-PPTopt
BPSmut-PPTopt
BPSalt-PPTopt
BPSalt-PPTalt
BPSopt-PPTalt
KH-Qua2 of SF1
RRM1,2 of U2AF2
B
SF1-U2AF2 complex with RNA
log (I)
log (s)
RNA free
C
q²I
q
SF1-U2AF2 complex with RNA
D
SF1-U2AF2 complex
Compact
Semi compact
Open
Rg (in Angstrom)
~4 Å
~4 Å
~9 Å
RNA free
E
SF1-U2AF2 complex with RNA
Compact
Semi-compact
Open
P(r)
r(Å)
D max ~132Å
D max ~182Å
D max ~198Å
F
SF1-U2AF2 complex
Dmax (Angstrom)
~50 Å
~18 Å
~66 Å
G
KHqua2
RRM2
RRM1
HH
CHM
BPSopt/sub-opt/mut
PPTopt
BPSalt
PPTopt/alt
Compact shape
Semi-compact shape
Open shape

Figure 3.21. Shape analysis of SF1-U2AF2 complex with variable strength of BPS and PPT sites. (A) A list of oligonucleotide RNA sequences are highlighted with variable strength of BPS and PPT sites. Double logarithmic (B) and Kratky plot (C) are plotted for SF1-U2AF2 with different RNA oligos. (D) R_g plot is shown for SF1-U2AF2 complex with respective RNA. (E) Pair distance distribution *P(r)* vs r (in Angstrom) plot (E) and D_{max} bar plot (F) are plotted for SF1-U2AF2 complex with different RNA oligonucleotides. (G) Schematic models of SF1-U2AF2 with different conformations are shown.

3.17 SF1-U2AF2 complex with increasing spacing between BPS and PPT

In vitro iCLIP analysis of over 2,000 transcripts revealed that the stretch of canonical PPT sites located near 6 to 15 nucleotides from “AG” 3' splice sites of introns. The distance of canonical BPS sites also varies between 2 to 140 nucleotides from 3’ "AG" splice sites (Pastuszak et al., 2011). Studies have shown that BPS sites located far from PPT sites can reduce the efficiency of splice site recognition and lead to intron retention. Therefore, optimal distances between BPS and PPT sites are crucial for the splicing assembly.

To better understand how splicing factors recognize splice sites at varying distances, the SAXS experiments were performed on the SF1-U2AF2 complex in the presence of RNA with variable distances between splice sites. For that, RNAs oligos were designed for having the optimized BPS (“UACUAAC”) and optimized PPT ("UUUUUUUUU") motifs with increasing the spacing in between (**Figure 3.22 A-C; Table 3.2**). The analysis demonstrated that the complex of SF1-U2AF2 bound to RNA with BPS^{opt} and PPT^{opt} motifs spaced by 2 nucleotides (nt) had a compact conformation, with a size of ~160 Å and R_g of ~37 Å. However, the complex with RNA having 10 nt spacing showed a well-extended conformation, with a D_{max} and R_g of ~195 Å and ~46 Å. On the other hand, the SF1-U2AF2 complex with RNA having 5 nt spacing showed a semi-compact conformation, with a D_{max} of ~180 Å and R_g of ~42 Å **(Figure 3.22 D-F**). The Kratky plot extracted from SAXS data suggested a shape for a folded single or multiple domains protein with a flexible region at a higher scattering angle (**Figure 3.22 C; Table 3.2**).

Based on these data, the SF1-U2AF2 complex can adopt different conformations (compact, semi-compact, or open) depending on the distance between BPS and PPT splice sites. This allows the complex to dynamically locate nearby strong splice sites on introns (**Figure 3.22 G**). Thus, multi-domain splicing factors with dynamic conformations are crucial for recognizing a wide range of intron sequences.

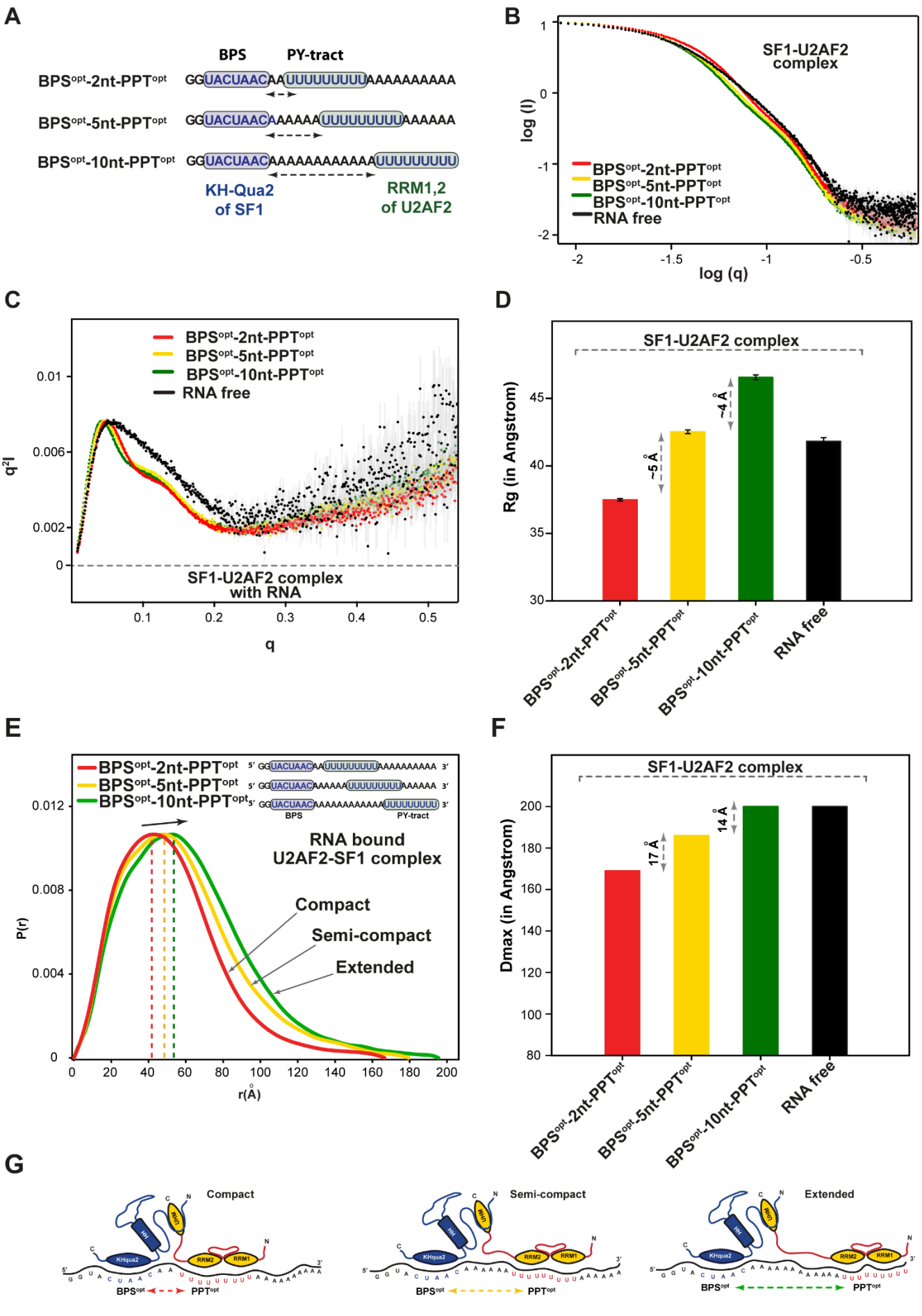

A
BPS
PY-tract
BPSopt-2nt-PPTopt
GGUACUAACAAUUUUUUUUUAAAAAAAAAA
BPSopt-5nt-PPTopt
GGUACUAACAAAAAAUUUUUUUUUAAAAAA
BPSopt-10nt-PPTopt
GGUACUAACAAAAAAAAAAAUUUUUUUUU
KH-Qua2 of SF1
RRM1,2 of U2AF2
B
SF1-U2AF2 complex
log (I)
log (q)
BPSopt-2nt-PPTopt
BPSopt-5nt-PPTopt
BPSopt-10nt-PPTopt
RNA free
C
q²I
q
SF1-U2AF2 complex with RNA
D
SF1-U2AF2 complex
Rg (in Angstrom)
~5 Å
~4 Å
E
P(r)
r(Å)
RNA bound U2AF2-SF1 complex
Compact
Semi-compact
Extended
F
SF1-U2AF2 complex
Dmax (in Angstrom)
17 Å
14 Å
G
Compact
Semi-compact
Extended
KHqua2
RRM2
RRM1
BPSopt
PPTopt

Figure 3.22. Shape analysis of SF1-U2AF2 complex with a variable position of the splice site. (A) RNA sequences are shown where BPS and PPT sites are highlighted. A double logarithmic plot (B) and Kratky plot (C) are shown for the SF1-U2AF2 complex with respective RNAs. (D) The radius of gyrations plotted for SF1-U2AF2 with respective RNAs. Pair distance distribution function (E) and respective Dmax (F) plots are highlighted for the SF1-U2AF2 complex with increasing spacing between BPS and PPT sites. (G) Schematic models of different conformations of SF1-U2AF2 are shown.

3.18 SAXS analysis of SF1-U2AF2 interactions with multiple splice signals

BPS sequences are highly degenerative in humans, meaning some intron sequences may contain more than one BPS-like motif. In cases of the disease, a point mutation on the BPS site can alter the recognition of the RNA splice site, causing splicing factors and subsequent spliceosome assembly to bind to nearby BPS-like (also known as cryptic) splice sites. This can alter the recognition of the constitutive splice site. SAXS studies were carried out to determine how the SF1-U2AF2 complex identifies such RNAs with several accessible splice sites. For that, RNA oligos were designed with two BPS sites and a PPT stretch (UUUUUUUUU) at respective distances (**Figure 3.23 A-C; Table 3.2**).

Results showed that when the SF1-U2AF2 complex was present with RNA comprising the first and second optimized BPS (UCACUAAC) motifs separated by 2 nt and 17 nt from PPT stretch, the D_{max} was approximately 157 Å and 210 Å, and R_g was around 37 Å and 42 Å, respectively (**Figure 3.23 D-F yellow, magenta**). This represents the compact and extended conformation of the SF1-U2AF2 complex, respectively. This suggests that despite having two BPS sites, the SF1-U2AF2 complex prefers to bind to the nearest splice sites with respect to the PPT site. Moreover, when RNA with the optimized first BPS sites and mutated the far BPS site from PPT, the SF1-U2AF2 complex presented a similar D_{max} of around 215 Å and R_g of approximately 42 Å, indicating an extended conformation (**Figure 3.23 blue; Table 3.2**). A similar extended conformation was also observed for RNA with both optimized BPS and altered PPT.

In conclusion, the SF1-U2AF2 complex can adjust its shape to find the BPS site located near PPT, even other similar BPS-like splice sites are on the intron (**Figure 3.23 G**). In case of mutation on canonical BPS, the splicing factors may recognize nearby sites that alter the splicing products and cause disease.

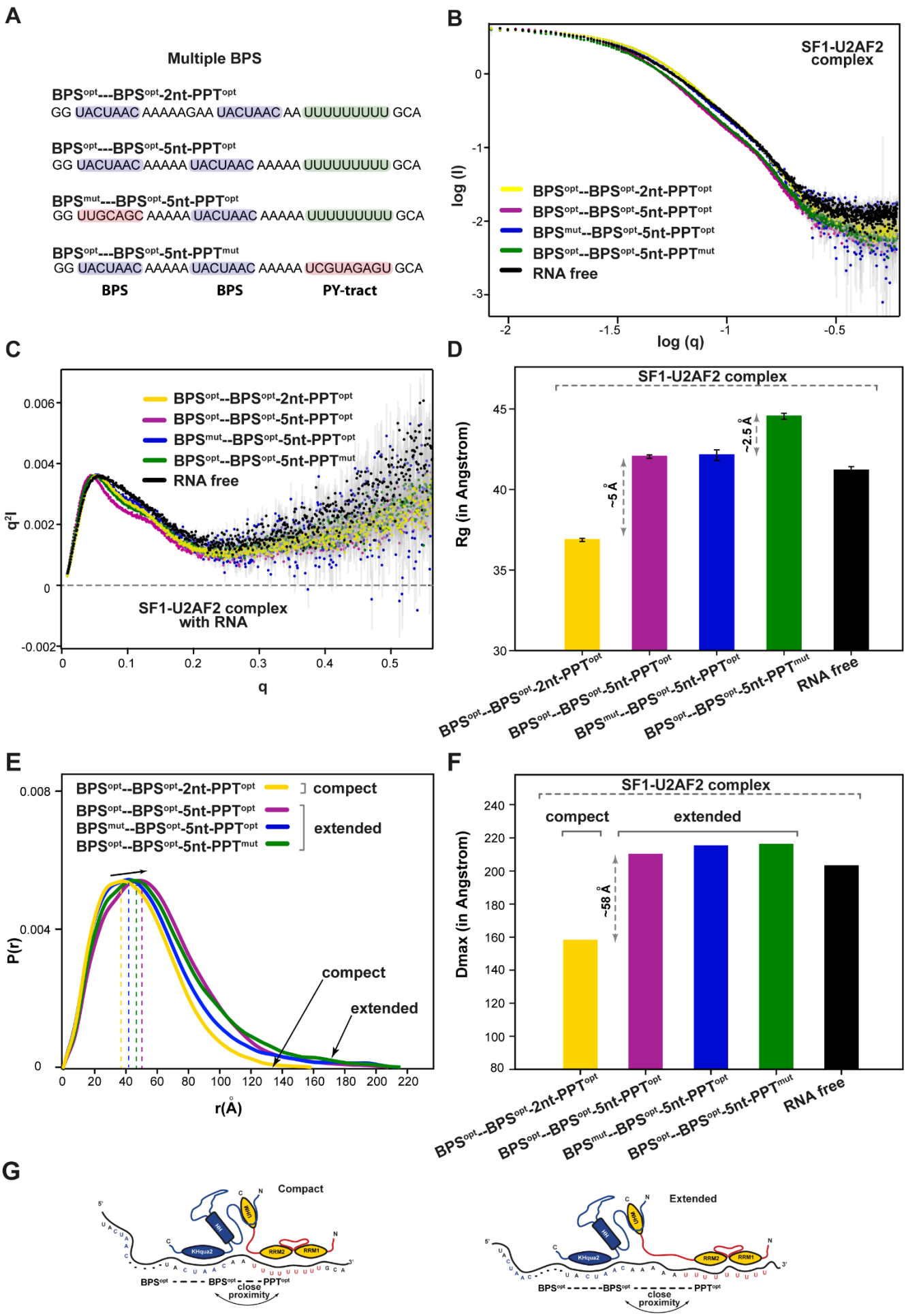

A
Multiple BPS
BPSopt---BPSopt-2nt-PPTopt
GG UACUAAC AAAAAGAA UACUAAC AA UUUUUUUUU GCA
BPSopt---BPSopt-5nt-PPTopt
GG UACUAAC AAAAA UACUAAC AAAAA UUUUUUUUU GCA
BPSmut---BPSopt-5nt-PPTopt
GG UUGCAGC AAAAA UACUAAC AAAAA UUUUUUUUU GCA
BPSopt---BPSopt-5nt-PPTmut
GG UACUAAC AAAAA UACUAAC AAAAA UCGUAGAGU GCA
BPS
BPS
PY-tract
B
SF1-U2AF2 complex
log (I)
log (q)
BPSopt--BPSopt-2nt-PPTopt
BPSopt--BPSopt-5nt-PPTopt
BPSmut--BPSopt-5nt-PPTopt
BPSopt--BPSopt-5nt-PPTmut
RNA free
C
q²I
q
SF1-U2AF2 complex with RNA
D
SF1-U2AF2 complex
Rg (in Angstrom)
~5 Å
~2.5 Å
E
compect
extended
P(r)
r(Å)
F
SF1-U2AF2 complex
Dmax (in Angstrom)
~58 Å
G
Compact
Extended
close proximity

Figure 3.23. Shape analysis of SF1-U2AF2 complex with RNA having multiple BPS splice sites. (A) RNA oligos having multiple BPS and PPT sites are highlighted. A double logarithmic plot (B) and a Kratky plot (C) are shown. (D) The radius of gyrations plotted for SF1-U2AF2 complex in the presence of respective RNAs. Pair distance distribution function (E) and respective D_{max} (F) plots are shown. (G) Schematic models for compact and extended conformations of RNA-bound SF1-U2AF2 are shown.

Table 3.2. SAXS analysis for SF1-U2AF2 complex in the presence of various RNAs.

Protein	RNA	RNA sequence	Rg (Å)	Dmax (Å)
	Complex with RNA having various BPS and PPT			
$SF1^{1-260}$ - $U2AF2^{140-475}$ Complex	RNA free	N/A	41.6 ± 0.1	200
	BPS^{opt}-PPT^{opt}	UACUA**A**CAAUUUUUUUUU	33.7 ± 0.0	132
	$BPS^{sub-opt}$-PPT^{opt}	UACCA**A**CAAUUUUUUUUU	32.8 ± 0.0	120
	BPS^{mut}-PPT^{opt}	UACCACCAAUUUUUUUUU	34.7 ± 0.0	138
	BPS^{alt}-PPT^{opt}	CCCCCCCCCUUUUUUUUU	38.0 ± 0.1	182
	BPS^{alt}-PPT^{alt}	CCCCCCCCCAAAAAAAAA	38.2 ± 0.1	182
	BPS^{opt}-PPT^{alt}	UACUA**A**CAAUAAAAAAAA	42.5 ± 0.1	198
	Complex with RNA having increasing distance between BPS and PPT			
	BPS-2nt-PPT	GGUACUA**A**CAAUUUUUUUUU-AAAAAAAAAA	37.5 ± 0.1	169
	BPS-5nt-PPT	GGUACUA**A**CAAAAAAUUUUUUUUUUAAAAAA	42.5 ± 0.1	186
	BPS-10nt-PPT	GGUACUA**A**CAAAAAAAAAAAA-UUUUUUUUU	46.6 ± 0.2	200
	Complex with RNA having multiple BPS sites and PPT			
	BPS^{opt}-BPS^{opt}-2nt-PPT^{opt}	GGUACUA**A**CAAAAAGAA UACUA**A**CAAUUUUUUUUUGCA	36.9 ± 0.1	157.9
	BPS^{opt}-BPS^{opt}-5nt-PPT^{opt}	GGUACUA**A**CAAAAAUACUA**A**C AAAAAUUUUUUUUUGCA	42.0 ± 0.1	210
	BPS^{mut}-BPS^{opt}-5nt-PPT^{opt}	GGUUGCAGCAAAAAUACUA**A**C AAAAAUUUUUUUUUGCA	42.1 ± 0.3	215
	BPS^{opt}-BPS^{opt}-5nt-PPT^{mut}	GGUACUA**A**CAAAAAUACUA**A**C AAAAAUCGUAGAGUGCA	44.5 ± 0.2	216

3.19 Ensemble structure of U2AF2 complex in free and RNA bound states

Based on the analysis of the SF1-U2AF2 protein complex, it appears that the shape of the complex can vary depending on the RNA it interacts with, as well as the nearby BPS and PPT splice sites. To better understand this structure and dynamic behavior, NMR chemical shifts based on secondary structure propensity (CSI) were calculated for $U2AF2^{140-475}$ using TALOS-N (Shen & Bax, 2013). The analysis revealed that the linkers between RRM1 to RRM2 and RRM2 to UHM domains are largely unstructured, while RRM1, RRM2, and UHM domains follow the known canonical RRM "β1-α1-β2-β3-α2-β4" folds (**Figure 3.24 A, Table 3.3**).

To further investigate the dynamics of the linkers NMR {1H}-15N heteronuclear NOE and ^{15}N R_1, R_2 relaxation rates were measured. The data showed that both linkers connecting RRM1 to RRM2 and RRM2 to UHM stay fully flexible in the sub-nanosecond (ns) time scale, while RRM1, RRM2 and UHM domains remain rigid with respect to the linker (**Figure 3.24 A**). The *P(r)* plot extracted from SAXS data also shows that the shape of U2AF2 in RNA-free form adapts the elongated conformation with D_{max} 167 Å and R_g of 35 Å. Even with the addition of PPT^{opt} ("5'-UUUUUUUUU-3'") RNA to U2AF2, the shape remains very similar with D_{max} 160 Å and R_g 35 Å (**Figure 3.24 C, D, Table 3.3**). These results demonstrate that an elongated flexible linker connecting RRM2 to UHM provides flexibility to U2AF2.

To further explore the structure, an ensemble modeling approach was used for U2AF2 alone and with “5’-UUUUUUUUU-3’” (U9) RNA. For RNA-free U2AF2, RRM1, RRM2, and UHM domains were treated as rigid structures, while the linkers between the domains were treated as flexible. The best-fit (χ^2 1.2) four conformers of U2AF2 have an average R_g of 35 Å and D_{max} of 100 Å. *Rflex* was 87.7% for selected ensembles compared to the random pool, which was 86.16, and *Rσ* was 1.17, suggesting several conformers in the solution (**Figure 3.24 E, F, Table 3.3**). A similar analysis was also performed for U2AF2 bound U9-RNA by considering UHM and U9 bound RRM1,2 complex (crystal structure) as a rigid structure and linker between RRM2 to UHM as flexible. The best-fitted five ensemble conformers were selected, which show an average R_g and D_{max} of 31 Å and 95 Å, respectively. The *Rflex* was 80.23% for ensembles out of 86.40% of the random pool (with *Rσ* of 0.95), predicting multiple conformers in solution (**Figure 3.24 G, H, Table 3.3**). In conclusion, the flexibility of the linker between RRM2 to UHM plays a vital role by allowing the U2AF2 to sample a wide range of conformations.

Table 3.3. SAXS analysis of U2AF2 in free and bound to "U9" RNA.

Structural analysis of U2AF2 in free and U9 RNA bound to U2AF2		
U2AF2$^{140-475}$ free	U2AF2	RRM1,2-UHM (140-475 aa)
	R_g	35.02 ± 0.3 Å
	D_{max}	167.0 Å
	EOM statistics	
	Random pool	10000 structures from Rg 21.37 Å to 60.83 Å
	Population of Selected ensemble	Four ensemble structures 50.6 %, Rg 30.50 Å, Dmax 100.01 Å 33.4 %, Rg 42.80 A, Dmax 135.37 Å 08.1 %, Rg 36.51 A, Dmax 104.92 Å 08.1 %, Rg 28.90 A, Dmax 091.93 Å
	Final ensemble	Rg 35.00 Å, Dmax 111.53 Å
	Rflex	87.73 % (from ~86.18%)
	Rσ	1.17
	Ensemble models fitting with experimental data	χ^2 = 1.14
U2AF2$^{140-475}$ with 5'UUUUUUUUU3' RNA	U2AF2	RRM1,2-UHM (140-475 aa)
	RNA	U9 (5' UUUUUUUUU 3')
	R_g	35.24 ± 0.2 Å
	D_{max}	160.0 Å
	EOM statistics	
	Random pool	10000 structures from Rg 20.27 Å to 52.76 Å
	Population of Selected ensemble	Five ensemble structures 11.1 %, Rg 26.33 Å, Dmax 83.27 Å 11.1 %, Rg 29.01 Å, Dmax 96.25 Å 33.3 %, Rg 34.12 Å, Dmax 107.81 Å 22.2 %, Rg 26.79 Å, Dmax 77.78 Å 22.2 %, Rg 38.38 Å, Dmax 114.11 Å
	Final ensemble	Rg 32.0 Å, Dmax 98.53 Å
	Rflex	80.23 % (from ~86.40 %)
	Rσ	0.95
	Ensemble models fitting with experimental data	χ^2 = 2.87

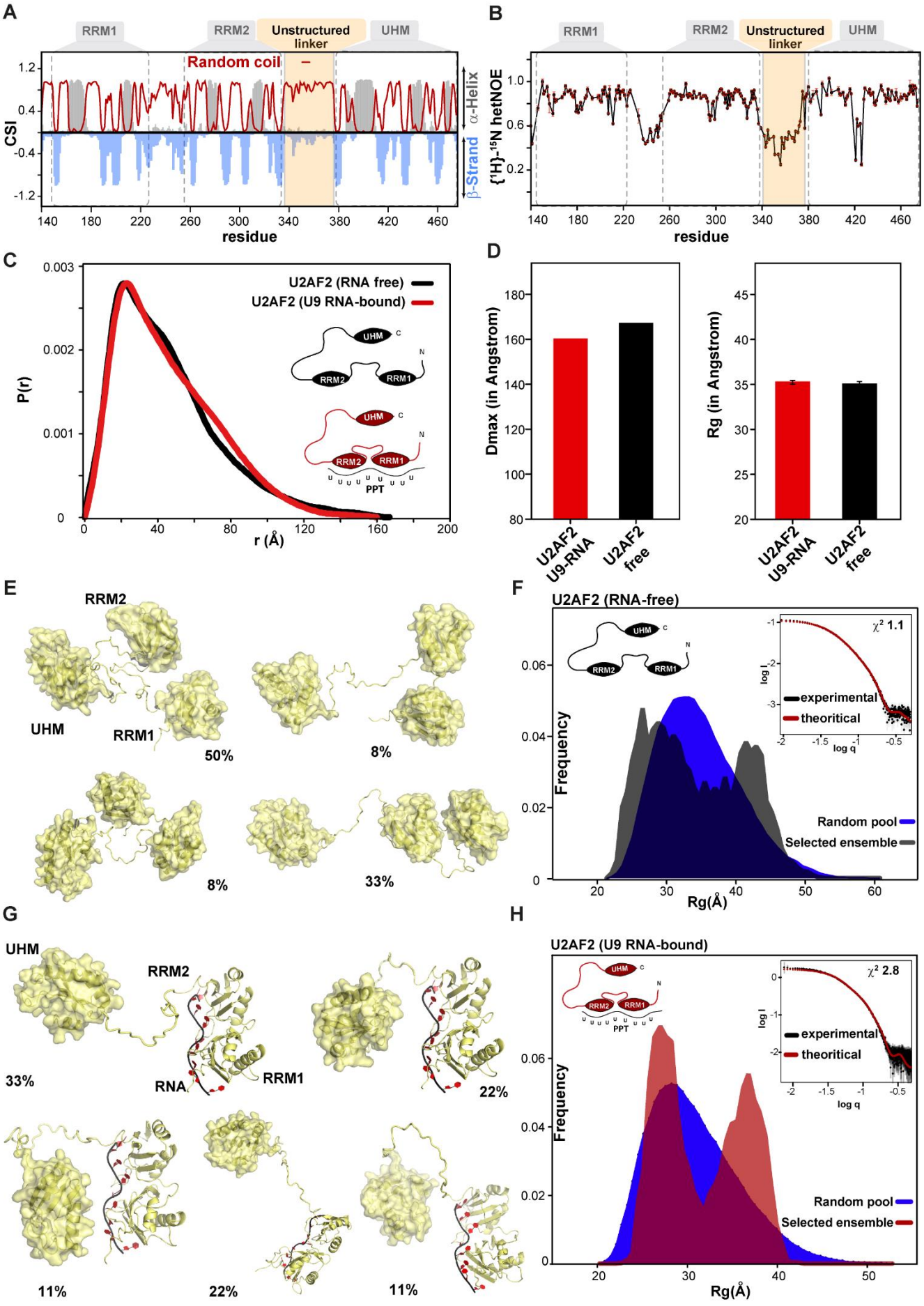

A
RRM1
RRM2
Unstructured linker
UHM
Random coil
CSI
α-Helix
β-Strand
residue
B
{1H}-15N hetNOE
residue
C
U2AF2 (RNA free)
U2AF2 (U9 RNA-bound)
P(r)
r (Å)
UHM
RRM2
RRM1
PPT
D
Dmax (in Angstrom)
Rg (in Angstrom)
U2AF2 U9-RNA
U2AF2 free
E
RRM2
UHM
RRM1
50%
8%
8%
33%
F
U2AF2 (RNA-free)
χ2 1.1
experimental
theoritical
log I
log q
Frequency
Random pool
Selected ensemble
Rg(Å)
G
UHM
RRM2
RNA
RRM1
33%
22%
11%
22%
11%
H
U2AF2 (U9 RNA-bound)
χ2 2.8
experimental
theoritical
log I
log q
Frequency
Random pool
Selected ensemble
Rg(Å)

Figure 3.24. Secondary structure and flexibility analysis of U2AF2$^{140-475}$. (A) Secondary structure propensity (chemical shift index) is plotted for RRM1,2-UHM construct of U2AF2 (top). RRM1, RRM2, and UHM region are in gray, while the linker between RRM2-UHM is highlighted with light orange. Helix and strands are shown in gray and light blue, respectively. (B) {^{1}H}-^{15}N-HetNOE plot for U2AF2 is shown in blue and maroon, respectively (lower panel). (C) SAXS-based pair distance distribution plots are shown for U2AF2 RNA-free and U9-RNA bound forms. (D) D_{max} and R_g are plotted on the left and right panels, respectively. (E) The fraction of ensemble models and (F) R_g distribution are shown for the random pool (in blue) and ensemble models (gray) based on the ensemble optimization method for U2AF2$^{140-475}$ in RNA-free form. (G) The population of ensemble models and (H) R_g distribution are shown for the random pool (blue) and the selected ensemble (maroon) for U2AF2$^{140-475}$ in the presence of "U9" RNA.

3.20 Structure of the SF1-U2AF2 complex in free and bound to RNA

As mentioned above, the linker regions between RRMs and RRM2-UHM domains in U2AF2 are found to be flexible, according to heteronuclear NOE data. This flexibility remains unchanged even in the presence of SF1, which suggests that U2AF2 domains behave independently in both free and SF1-bound forms (**Figure 3.25 A**). SAXS experiments were then conducted to gain a better understanding of the structures and dynamics of the SF1-U2AF2 complex. The results showed that the SF1-U2AF2 complex in RNA-free form and 18mer RNA have D_{max} ≈ 200 Å and 132 Å, respectively (**Figure 3.25 B-D, black and red**). On the other hand, the SF1-U2AF2 complex having BPS (5'-UACUAACAA-3') bound SF1 and PPT (5'-UUUUUUUUU-3') bound U2AF2 suggest the D_{max} and R_g of ≈169 Å and ≈ 39 Å, respectively, which shows intermediate compactness. This is due to the fact that BPS RNA reduces the flexibility of KH-Qua2 domains for SF1 and the PPT motif reduces the flexibility of RRM1,2 of U2AF2 (**Figure 3.25 B-D, cyan**). This analysis revealed that the significant contribution for the flexibility in this complex is due to the 35 residues flexible linker between RRM2-UHM and the minor contribution for eight residues short linker between HH-KH of SF1. NMR CSP showed no direct interactions between KHQua2 and RRM1,2 or UHM of U2AF2 (**Figure 3.25 E**). Thus, KHQua2 and RRM1,2 act as independent domains, and the SF1-U2AF2 complex adapts open conformation without RNA, while RNA motifs bring KHQua2 and RRM1,2 domains in proximity, forming a compact conformation.

To better understand the structure of the SF1-U2AF2 complex, ensemble modeling was used to identify the ensembles of conformations best fitted to SAXS. For that, BPS (UACUAAC) bound KHQua2 domain (from SF1), PPT bound RRM1,2 (from U2AF2) and UHM-ULM known structures were treated as rigid, while the linker between RRM2 and UHM was treated as

flexible to generate a pool of structures (**Figure 3.26 A**). From the pool, the best fitting χ^2 of 1.3 was obtained by combining five conformations of SF1-U2AF2 with an averaged R_g and D_{max} of ≈ 38 Å and ≈ 123 Å (**Figure 3.26 A-C)**. *Rflex* was higher for the ensemble of conformers coexisting in solution fitting to SAXS data than for the random pool of conformers covering the available conformational space (86.71% and 85.73%, respectively), and *Rσ* was >1 (1.17). These results predicted the co-existence of several flexible conformers of SF1-U2AF2 complex in solution with a 1.2-fold variation in the D_{max} value, from 117 to 148 Å in size. In short, the SF1-U2AF2 complex has multiple conformations in solution, which assist the complex in recognizing nearby splice sites.

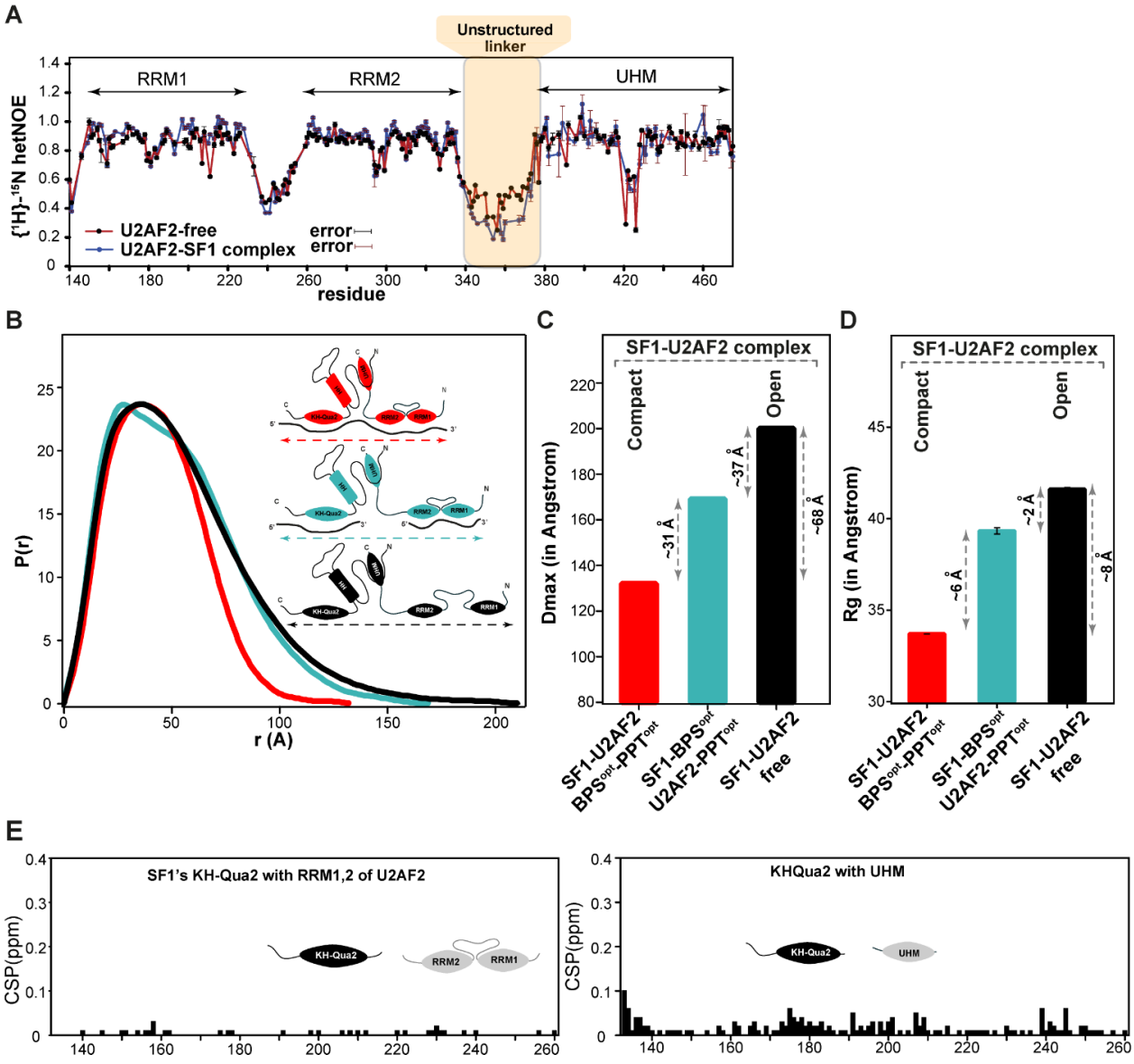


Figure 3.25. Flexibility and ensemble structure of SF1-U2AF2 structure. (A) Overlay spectra of {^{1}H}-^{15}N hetero-nuclear NOE analysis of U2AF2 alone (red) and bound to SF1 (blue). (B) Pair distance distribution analysis of SF1-U2AF2 complex in RNA-free (black), with two RNAs (BPS and PPT motif; cyan) and bound to RNA (BPS^{opt}-PPT^{opt}). Bar plot for D_{max} (C) and R_g (D) of SF1-U2AF2 complex. (E) CSP plot shown for NMR titration of ^{15}N labeled KH-Qua2 with RRM1,2 of U2AF2 (left panel) and UHM (right panel) are shown.

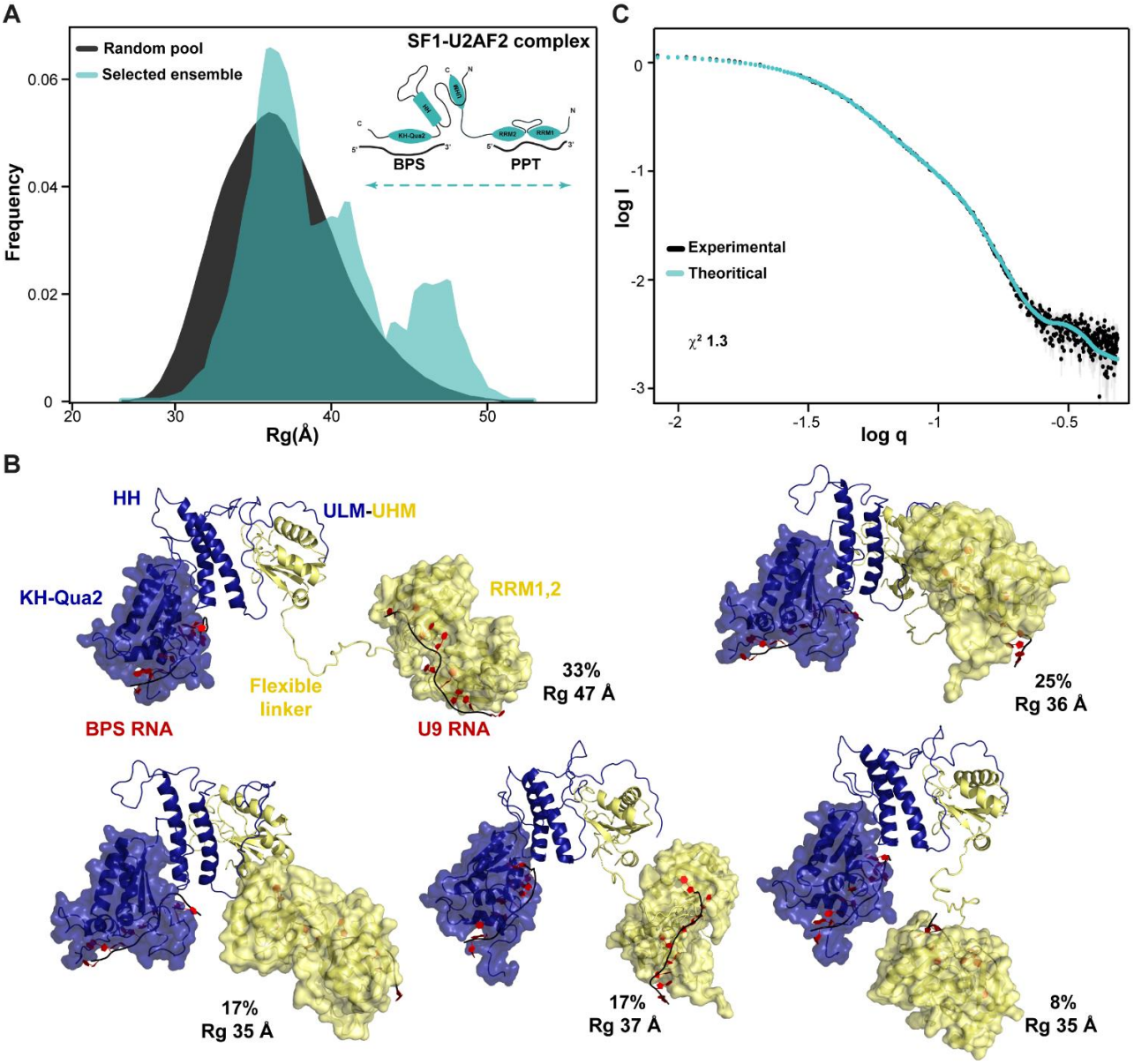


Figure 3.26. Ensemble structural models of SF1$^{1-260}$-U2AF2$^{140-475}$ bound to two independent fragments of RNAs. (A) EOM-based R_g distribution is shown for SF1 (bound to "5'-UACUAAC-3" RNA)-U2AF2 (bound to "5'-UUUUUUUUU 3' " RNA) complex, where a random pool and selected ensembles are colored in black and cyan, respectively. (B) The population of selected ensembles are shown, where SF1 and U2AF2 are in dark blue and yellow, respectively. (C) The agreement of back-calculated and experimental SAXS is plotted in cyan and black, respectively.

Table 3.4. Structural statistics for SF1-U2AF2 complex bound to RNA

Complex $SF1^{1-260}$ bound to "5' UACUAAC 3' " and $U2AF2^{140-475}$ bound to "5' UUUUUUUUU 3' " RNA		
SF1-U2AF2 Complex bound to two RNA fragments	$SF1^{1-260}$ bound to 5' UACUAAC 3' RNA and $U2AF2^{140-475}$ bound to 5'- UUUUUUUUU 3' RNA	
	R_g	39.33 ± 0.17 Å
	D_{max}	169.0 Å
	EOM statistics	
	Random pool	10000 structures from Rg 26.59 Å to 52.96 Å
	Population of Selected ensemble	**Five ensemble structure** 25.3 %, Rg 36.26 Å, Dmax 122.75 Å 17.2 %, Rg 37.55 Å, Dmax 117.53 Å 17.2 %, Rg 35.31 Å, Dmax 112.53 Å 33.4 %, Rg, 46.62 Å, Dmax 147.99 Å 08.1 %, Rg 34.97 Å, Dmax 117.01 Å
	Final ensemble	Rg 39.66 Å, Dmax 128.11 Å
	Rflex	86.71 % (from ~85.71 %)
	Rσ	1.17
	Ensemble model fitting with experimental data	$\chi^2 = 2.87$

3.21 Structure of SF1-U2AF2 bound to BPS^{opt}-PPT^{opt} RNA

The structural information of the SF1-U2AF2 complex in the presence of RNA is not well understood. To better understand this, a semi-rigid body refinement method was used to derive structural models of the complex with RNA that had optimized BPS (5'-UACUAAC-3') and PPT (5'-UUUUUUUUU-3') sites, with a spacing of 2 nucleotides ("AA") in-between. For that, refined structures of individual domains of SF1 and U2AF2 were used from previous studies and derived inter-domain orientation and distance restraints from crystal (PDB 5EV1, 4FXW) and NMR-based structures. Also, protein-RNA distance restraints were derived from sub-domain structures (5EV1, 1K1G) for the structure refinement. The lowest energy structural models were sorted and scored using experimental SAXS data and derived best-fitted ensemble models of RNA-bound SF1-U2AF2 complex (**Figure 3.27**).

The derived structural model of the SF1-U2AF2-RNA complex (**Figure 3.27 A; Table**

3.5) indicates that the optimized PPT and BPS play a crucial role in bringing the RRM1,2 of U2AF2 closer to KHQua2 of SF1, leading to a compact structure. The KHQua2 domain binds to the BPS site, while the N-terminal of ULM interacts with the UHM of U2AF2. The RRM1,2 domains of U2AF2 are fixed when bound to U9 RNA, as seen in the crystal structure (5Ev1), the 35-residue linker connecting RRM2-UHM is unstructured, allowing for flexibility in inter-domain movement, which results in different conformations depending on RNA targets and spacing between PPT to BPS. Additionally, RNA with a spacer between BPS and PPT motifs have flexible backbone torsion angles α, β, γ, δ, ε, and ζ, which also contribute to the flexibility and shape of the complex from linear to inverted "V" shapes.

To sum up, RRM1,2 (of U2AF2) and KHQua2 (of SF1) do not interact directly. However, the position of the PPT and BPS motif on RNA affects the shape of the SF1-U2AF2 structure, which can change from a close to semi-compact to extended conformation based on RNA targets. These flexible shapes help the SF1-U2AF2 complex recognize the nearest canonical splice site, initiating the process of spliceosome assembly (**Figure 3.27**).

Table 3.5. Statistics for structural model of SF1-U2AF2 complex bound to RNA

Structure statistics (Rigid-body refinement)	
SF1 (N-terminal fragment)	ULM, HH, KH and Qua2 (1 to 260 aa) • ULM and HH (PDB 2m09) • KH-Qua2 (PDB 1k1g)
U2AF2 (C-terminal fragment)	RRM1,2, UHM (140 to 475 aa) • RRM1 and RRM2 (PDB 2yh0) • UHM (PDB 4fxw)
RNA (18mer)	5' UACUAACAAUUUUUUUUU 3' RNA
Total inter-domain protein distance restraints	705
• U2AF2 RRM1 and RRM2	230 (PDB 5ev1)
• SF1-U2AF2 (ULM-UHM)	344 (PDB 4fxw; 2m0g)
• SF1 HH to KHQua2	131 (PDB 4fxw + NMR SF1 structure)
Total inter protein-RNA distance restraints	2063
• U2AF2: RRM1-RRM2 to U9 RNA	1498 (PDB 5ev1)
• SF1: KH-Qua2 to BPS RNA	565 (PDB 1k1g)
Distance violations	< 3 Å
Ensemble models scotred against SAXS	20 structures
SAXS agreement	$\chi^2 < 6$

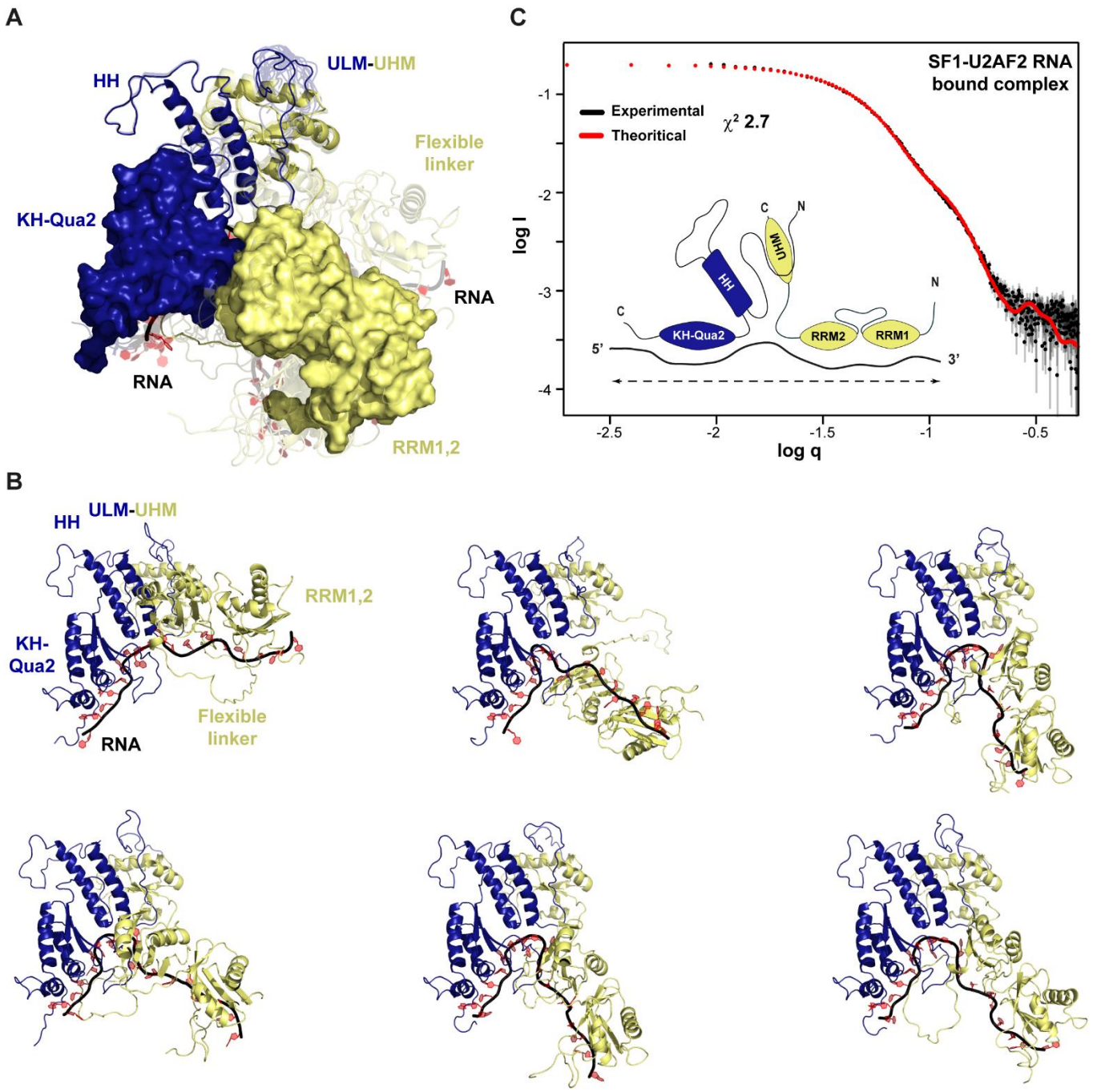


Figure 3.27. Models of SF1$^{1-260}$-U2AF2$^{140-475}$ complex bound to the RNA (5' UACUAACAAUUUUUUUUU 3") (A) Aligned 15 ensemble structures of SF1-U2AF2 RNA bound complex. The blue and yellow surface represents KHQua2 (of SF1) and RRM1,2 (of U2AF2), respectively. (B) SF1-U2AF2 structures with different shapes are shown. (C) Theoretical and experimental SAXS comparison for RNA-bound SF1-U2AF2 complex shown with red and black, respectively.

Discussion

In the early stages of spliceosome assembly, splicing factors SF1 and U2AF2 recognize the branch point site (BPS) and polypyrimidine tract (PPT) at the 3' splice site of the intron. In humans, the polypyrimidine tract sites are conserved across all introns, while the branch point sites (BPS) are highly degenerative, with only the branch site “A” and “U” +2 upstream positions being conserved in most introns. Any mutation at these positions can alter consecutive splicing, resulting in altered splice product that could cause severe human diseases. This study highlights how changes in splice sites sequence can affect protein recognition in the early stage of splicing. The current work demonstrates that an alteration at the +2 upstream position from the branch site abolishes SF1’s Qua2 domain interactions and reduces binding with the KH domain. Therefore, in the cases disease-associated BPS sequences, binding specificity and selectivity for SF1 are alterd, leading to alternative or abrupt splicing products. To gain deeper understanding of SF1’s structure, the paramagnetic relaxation enhancement (PRE), dynamics and small-angle X-ray scattering (SAXS) experiments were carried out for N-terminal soluble fragment comprising ULM, HH, KH and Qua2 domains. The PRE-derived SF1 structure indicates that the structured parts of HH and KH domains are relatively rigid in solution and oriented in proximity with respect to each other. However, the ULM and Qua2 regions show significant flexibility at sub-nanosecond time scale, facilitating U2AF2 and BPS RNA binding, respectively.

According to the literature, SF1’s presence is not essential for the splicing catalytic reaction, but it does speed up the splicing process (Tanackovic & Kramer, 2005). This explains the significant role of SF1 in the early stage of splicing assembly. To understand in detail, the iCLIP analysis was explored in the current study with over 2000 genes having 3’ intronic splice sites. The result shows that SF1 has a significant impact on stabilizing the U2AF2 to recognize polypyrimidine tract sites of transcripts. This demonstrates how precisely SF1 and U2AF2 regulates the kinetics of splicing reactions in the early stage of spliceosomal assembly. Also, reports show that phosphorylation of SF1 has a minor effect on RNA binding *in vitro* (Lipp et al., 2015) (Manceau et al., 2006). However, the iCLIP analysis of the current study found no major difference in RNA binding for non- and phosphorylated SF1-U2AF2 complex with variable splice sites. Means, both non- and phosphorylated complexes equally contribute for RNA binds

to the splice sites. However, a minor change in NMR chemical shifts observed upon phosphorylation of SF1-U2AF2 complex suggest the weak or transient interactions that may not differentiate the recognition of canonical splice sites. Thus, phosphorylation may have other unknown regulatory function in the cell.

The intricate composition of intron sequences in humans contains varying levels of complexity, including diverse strength and distances for BPS and PPT splice sites. Interestingly, some introns even have multiple BPS-like sequences. Despite this complexity, splicing factors accurately recognize the spice sites and initiate the splicing cycle. In this study investigated into how the SF1 and U2AF2 complex recognizes such diverse spice sites. The findings indicate that the SF1-U2AF2 complex can adapt to varying levels of structural compactness depending on the strength of the BPS and PPT sites. For stronger and weaker strengths of splice sites, the complex domains reorganize to compact to intermediate conformations, respectively. Additionally, a larger distance between splice sites results in an even more extended shape for the SF1-U2AF2 complex. Furthermore, the SF1-U2AF2 complex has a preference to bind nearby splice sites if several similar types of splice sites are present on the same intron. These dynamic regulations exhibited by the SF1-U2AF2 complex elucidate that protein complexes and inter-domain flexibility both play important roles in splicing regulation.

To understand the early stage of splicing assembly, it is important to have a structural understanding of SF1 and U2AF2. While the individual domain structures have been well characterized in previously (Agrawal et al., 2016; Mackereth et al., 2011; Wang et al., 2013; Zhang et al., 2012), the structure of the entire complex of SF1-U2AF2 with the 3' splice site RNA is not well understood. This study provides a detailed structural analysis of the complex of SF1-U2AF2 bound to the 3' splice site. The representative ensemble structure of the SF1-U2AF2 complex shows flexible elongated conformations. Without RNA, the SF1's KH-Qua2 and U2AF2's RRM1,2 domains remain apart. However, with the presence of RNA containing strong BPS and PPT sites, the KH-Qua2 domain moves closer to the RRM1,2 domains, resulting in a compact conformation of SF1-U2AF2, even without direct protein-protein interaction between these domains. In conclusion, this study emphasizes the important role that the multi-domain architecture and equilibrium between open and closed conformations of the SF1-U2AF2 complex play in facilitating efficient splice site recognition in the early stages of

splicing (**Figure 3.28**).

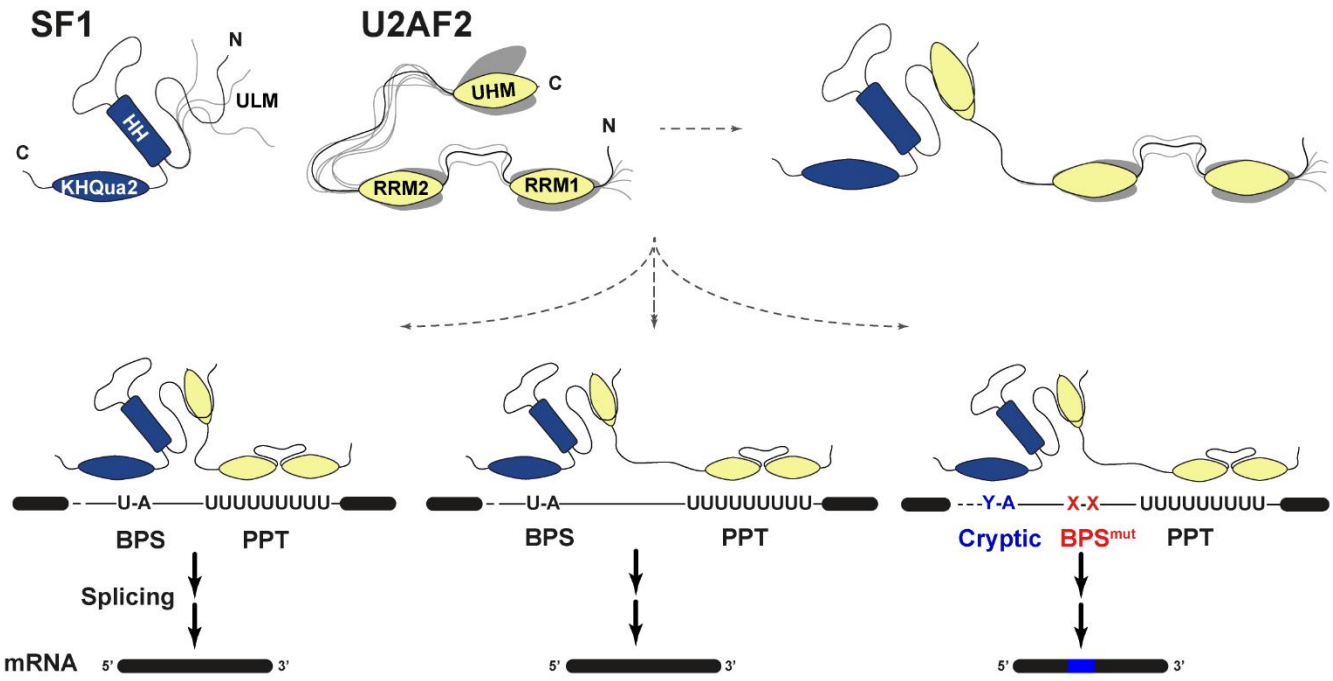


Figure 3.28. Summary of recognition of the 3' splice sites by SF1-U2AF2 complex. SF1 and U2AF2 have flexible inter-domain linkers, forming a dynamic and open complex. Both proteins are recruited to intron splice sites, adapt to different conformations based on the strength of splice sites and initiate the splicing cycle.

Chapter 4 – Structure, dynamics and function of yeast Npl3 in mRNP assembly

This part of the work has been published and permission has been obtained from the journal editor for using figures and text.

Philipp Keil[#], Alexander Wulf[#], **Nitin Kachariya**[#], Samira Reuscher, Kristin Hühn, Ivan Silbern, Janine Altmüller, Mario Keller, Ralf Stehle, Kathi Zarnack, **Michael Sattler***, Henning Urlaub*, Katja Sträßer*. Npl3 functions in mRNP assembly by recruitment of mRNP components to the transcription site and their transfer onto the mRNA*. Nucleic Acids Research, Volume 51, Issue 2, 25 January 2023, Pages 831–851 (doi:10.1093/nar/gkac1206).*

*# Equal contribution; *Corresponding authors.*

Introduction

4.1 mRNA biogenesis

In eukaryotes, the nucleus compartment contains most of the cell's nucleic acids and is separated from the rest of the cell by a nuclear membrane. This separation enables more efficient gene regulation, including the synthesizing messenger RNA (mRNA) and its metabolism. mRNA synthesis begins with the generation of precursor-mRNA (pre-mRNA) by RNA polymerase II and loading onto messenger ribonucleoprotein particles (mRNPs). These mRNPs are then processed through various steps, including 5' capping, intron splicing, 3' polyadenylation, and 3' end cleavage. While the mRNA metabolism process involves the export, translation, cellular localization, and degradation of mature mRNA. Throughout the entire process of mRNA biogenesis, distinct sets of trans-acting factors are involved, including small RNAs and ribonucleoproteins (RNPs), which form a distinct mRNP code. This code determines the fate of the mRNA and regulates gene expression, co- and post-transcription machinery. (Scott et al., 2019) (Jeong, 2017) (Linder et al., 2015). Any alterations to the canonical mRNP code through mutations in RBPs, RNA sequence, their recruitment or remodeling factors can lead to cell growth defects and/or human disease. Therefore, mRNA life-span is highly controlled and undergoes processes that either modulate its existence or promote its degradation (**Figure 4.1**).

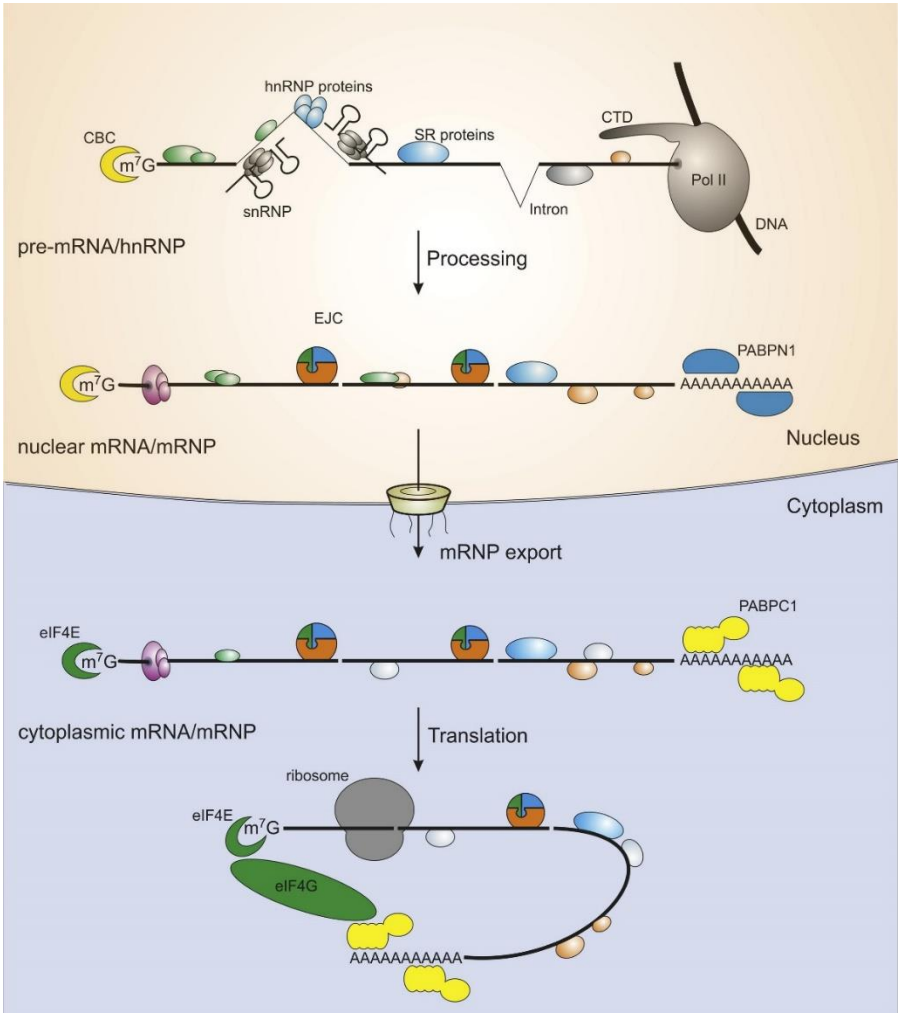


Figure 4.1. mRNA life-cycle. RNA-binding proteins are associated with the mRNA during co and post-transcription. Processed mRNA transport to the cytoplasm where the mRNA is translated and degraded (Figure adapted from Linder, Fischer et al. 2015).

4.2 Role of SR proteins in RNA processing

During RNA metabolism, mRNA transcription in the nucleus goes through multiple steps of processing, which include the pre-mRNA splicing, 5'-capping, 3'-end processing, polyadenylation, mRNA export, nonsense-mediated mRNA decay, and mRNA translation. The Serine-Arginine (SR) family proteins and other protein factors are crucial for mRNA biogenesis, as highlighted in **Figure 4.2**. Typically, an SR protein has one or multiple RNA recognition motifs (RRMs) and a serine-arginine rich region located at either the N or C-terminus of the protein. The RRM domain recognizes RNA and provides binding specificity, while the serine-arginine unstructured region facilitates protein-protein and transient protein-RNA interactions. Additionally, it is associated with the nuclear localization signal (Kramer, 2021) (Wagner & Frye, 2021).

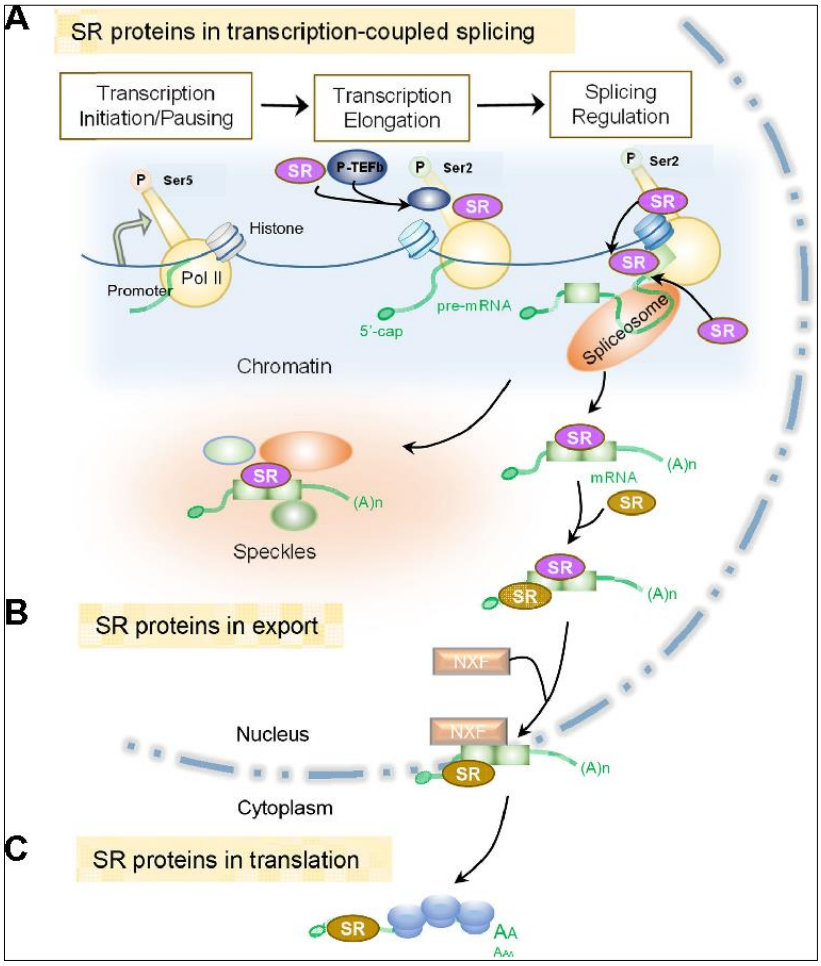


Figure 4.2. Schematic representation of cellular functions of SR proteins. (A) SR protein's regulation in transcription, translation, and splicing in the nucleus are shown. (B) mRNA cytoplasmic export regulated by SR proteins. (C) Translational regulation of SR proteins in the cytoplasm. (figure is adapted from Jeong 2017).

The number of SR family proteins varies across species and is closely linked to the complexity of alternative splicing in eukaryotes. The number of SR family proteins also varies from species to species and correlate well with the increasing complexity of alternative splicing

in eukaryote. Animal, plant, and metazoan species have relatively high numbers of SR proteins, while different types of fungi tend to have limited SR proteins, typically 1 to 3. In humans, 12 SR family proteins (SRSF1 to SRSF12) and several SR-like splicing factors have been extensively studied. *Schizosaccharomyces pombe* has Srp1 and Srp2 SR proteins, and *Saccharomyces cerevisiae* has Gbp2, Hrb1, and Npl3 SR proteins (**Figure 4.3**). Among these proteins, the SR-like Npl3 protein is essential for a cell viability in *Saccharomyces cerevisiae* and contributes to transcription, splicing, mRNP assembly, mRNA processing, and nuclear mRNA export. (Rima Sandhu et al., 2021) (Zhang et al., 2020) (Plass et al., 2008) (Long & Caceres, 2009) (Busch & Hertel, 2012) (Scott et al., 2019) (Wagner & Frye, 2021).

In short, SR proteins have a significant role in RNA processing and metabolism. Any mutations or dysregulation in these proteins can alter the constitutive RNA regulation, resulting in developmental defects or cell death.

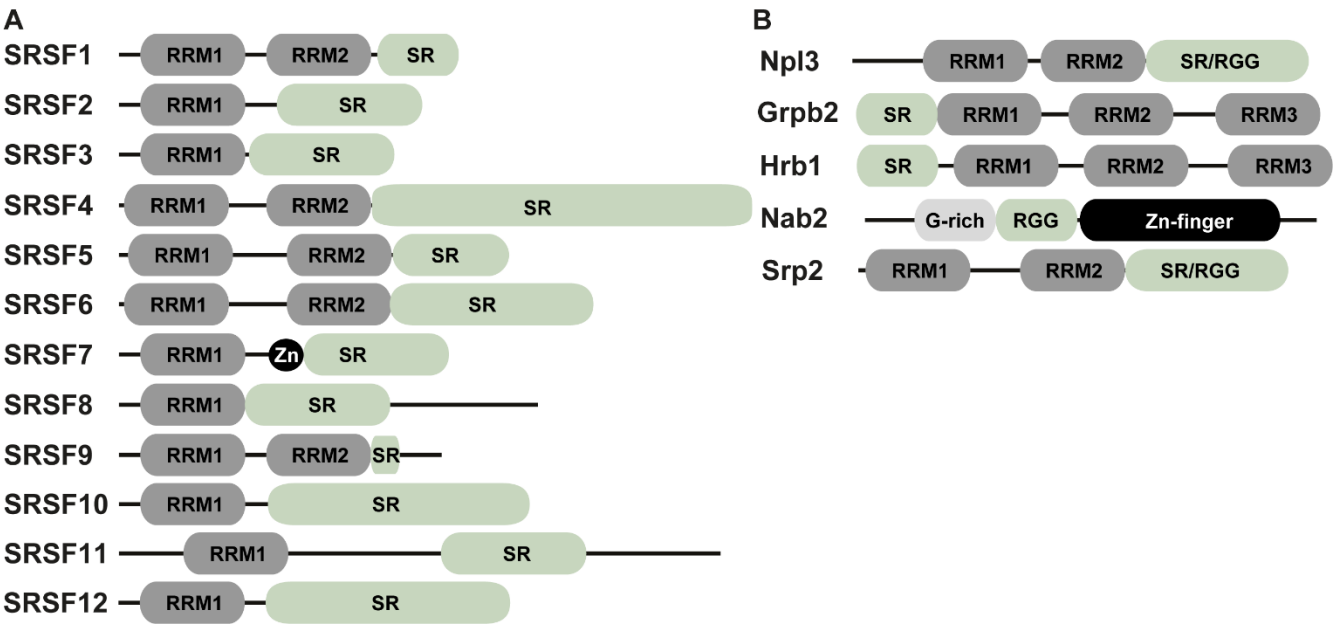


Figure 4.3. Domain architecture of SR proteins. (A) Schematics of Human SRSF1-12 proteins are shown (B) Domain structures of yeast SR proteins are shown. RRM and SR domains are highlighted with gray and light green, respectively.

4.3 Function of Npl3 SR-like protein in budding yeast

Npl3 is a highly abundant RNA-binding protein in budding yeast that plays a crucial role in various regulatory pathways. It shuttles between the nucleus and cytoplasm, but is primarily a nuclear protein. During co-transcription process, Npl3 interacts with RNA polymerase-II

through its C-terminal tail and is recruited to transcribing genes. It is also involved in early splicing machinery and binds to both 5' and 3'-ends of transcripts, indicating its role in early and late mRNA maturation events. Deleting *Npl3* gene causes significant changes in mRNP formation and developmental defects in yeast. In the nucleus, Npl3 along with other RBPs such as Nab2, Yra1, Sub2 Hrb1, Pab1, and Gbp2 proteins load to the nascent mRNA and exported to the nuclear pore complex with the help of Mex67-Mtr2 receptor complex. Npl3 has also been shown to facilitate the export of large ribosomal subunits from the nucleus to the cytoplasm. In the cytoplasm, Sky1 phosphorylates Npl3 which facilitates the dissociation of Npl3 from mRNA and binds to Mtr10 protein. This Npl3-Mtr10 shuttles back to the nucleus via the nuclear pore complex, where Glc7 phosphatase release phosphate from Npl3 to restart the cycle (**Figure 4.4**).

Npl3 deficiency causes extensive alterations in mRNA export and splicing, including an effect on ribosomal protein genes. It regulates meiotic splicing machinery, allowing proper execution of meiotic cell division in yeast. Npl3 also recruits splicing factors to chromatin-associated transcripts, indicating cross-talk between the spliceosome and chromatin modification (McBride et al., 2005) (Hackmann et al., 2011) (Rima Sandhu et al., 2021) (Moehle et al., 2012) (Kress et al., 2008) (Wan et al., 2022) (Lukong & Richard, 2004).

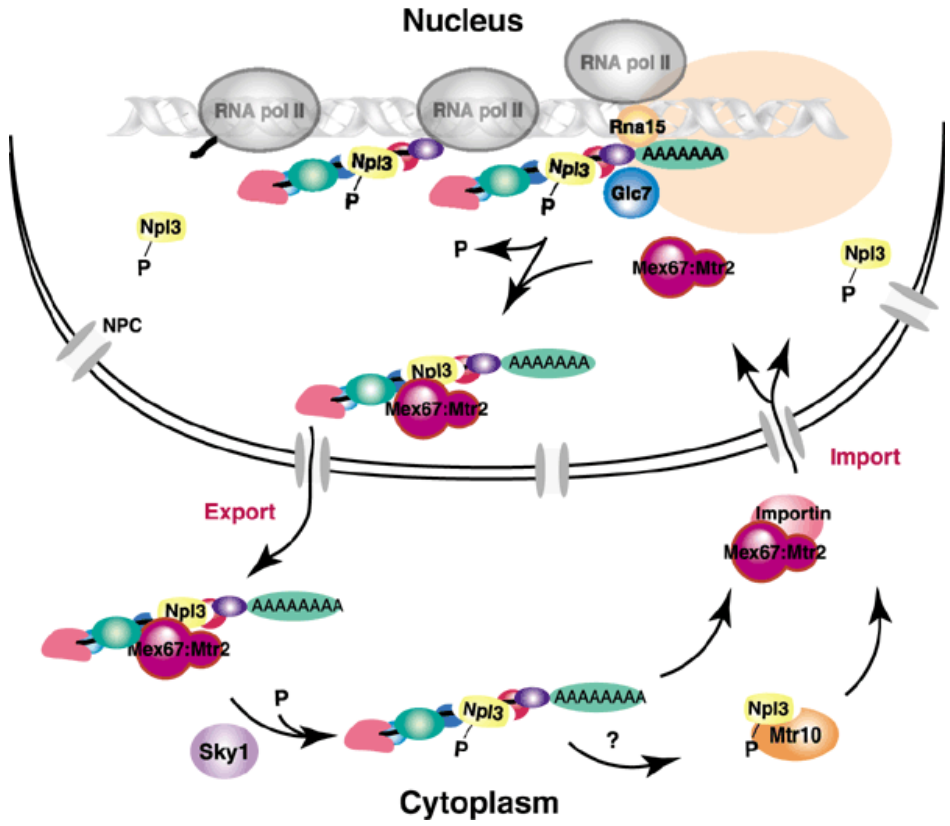


Figure 4.4. Schematic representation of Npl3 regulation. Npl3 with other RNA binding proteins are recruited during the transcription process to the nascent mRNAs. After that mRNP assembly, 3′ end-processing and association of Mex67p-Mtr2p complex result in the retention of the transcripts at the site of transcription (orange area). After cytoplasmic translocation, the cytoplasmic Sky1p kinase phosphorylates Npl3p, promoting the dissociation of Mex67p-Mtr2p and Npl3p shuttle back to the nucleus (Figure is adapted from Lukong and Richard 2004).

4.4 Post-translational modification of Npl3

Npl3 undergoes post-translational modifications, including methylation and phosphorylation, much like other SR proteins. The Npl3's phosphorylation is carried out by Sky1 kinase, which is similar to mammalian SR protein kinase-1. It facilitates the shuttling of Npl3 between the nucleus and the cytoplasm. However, the loss of phosphorylation sites on Npl3 has been shown to disrupt the import-export system and halt transcription processes without affecting splicing in yeast. On the other hand, arginine methylation affects its nuclear export, self-association, and interaction with Tho2 binding partner. Mutating arginine residues weaken the interactions between nuclear proteins and ultimately altering the mRNA export (McBride et al., 2005). Npl3 methylation is also linked to the splicing of meiosis-specific Mer1-dependent transcript, highlighting its role in transcript-specific splicing regulation in yeast (R. Sandhu et al., 2021).

4.5 Npl3 structure and its RNA recognition

Similar to known SR proteins, Npl3 has two RNA recognition motifs (RRMs) located between 120 to 280 residues. The N-terminus region has acidic residues such as Gln, Glu, and Pro, while the C-terminal tail has Arg, Ser, and Gly-rich repeats between 281 to 414 aa. The C-terminus tail is also considered as an "RGG" motif as it comprises 15 "RGG" repeats (**Figure 4.5 A**). Regarding the structure, the structure of individual RRMs is known from solution NMR. RRM1 is 70 residues longer and adapts to the canonical RRM fold. However, RRM2 is slightly larger, with 90 residues in size, and has a canonical RRM fold with "β-α-β-β-α-β-β" (**Figure 4.5 B**). RRM1 has two RNP sites, RNP1 "N-G-F-A-F-V-E-F," and RNP2 "L-F-V-R-P-F," for RNA interactions like other known RRMs. In contrast, RRM2 lacks the canonical RNP motifs despite having an RRM fold and is thus considered as pseudo-RRM or non-canonical RRM. A similar pseudo-RRM has also been identified in the human homolog SRSF1. Literature suggests that Npl3 RRMs bind to U+G RNA sequences (Holmes et al., 2015) (Deka et al., 2008). However, the RNA binding specificity and function of each RRM1 and pseudo-RRM, the structural details

of the tandem RRMs, the role of the linker connecting both domains, and the importance of residues from core sites of RRMs in mRNA export remain unexplored.

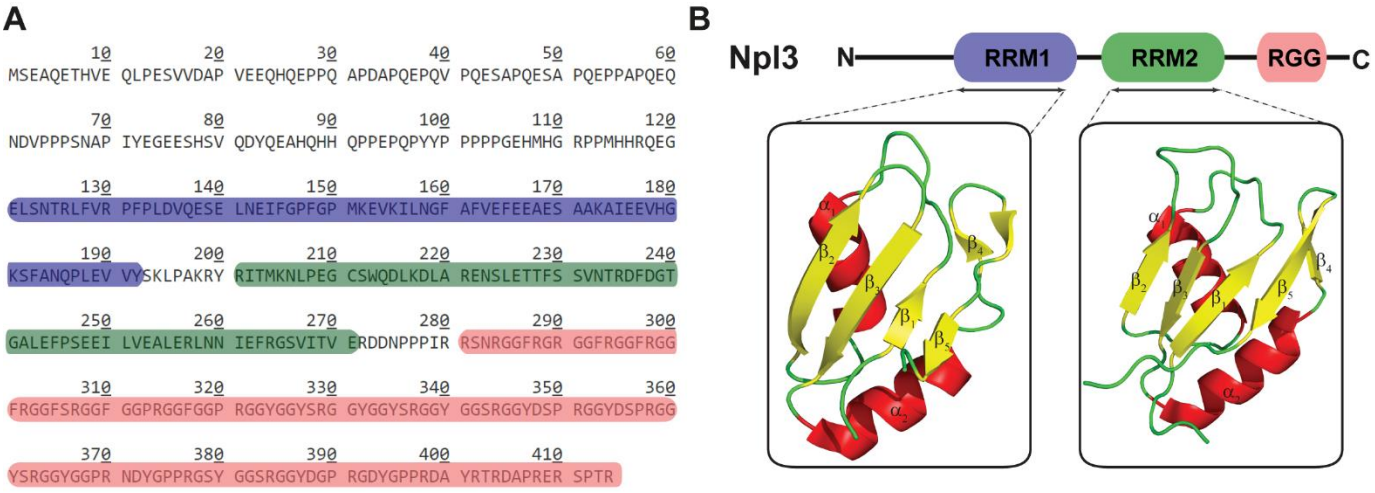


Figure 4.5. Sequence and domain structure of Npl3. (A) A full-length sequence of Npl3 is shown. RRM1, RRM2, and RGG domains are colored in blue, green, and pink. (B) Solution structure of RRM1 (left) and RRM2 (right) are shown where helix, sheet, and loops are highlighted in red, yellow, and green. The PDB accession code for RRM1 and RRM2 are 2OSQ and 2OSR.

Results

4.6 Npl3 domain architecture and protein sequence conservation

RNA-binding proteins in nuclear mRNP components play a crucial role in post-transcriptional gene regulation. To understand their functions better, the UV cross-linking combined with Mass spectrometry methods were used to identify the RBPs that directly interact with RNA. From analysis, over 100 cross-linked peptides were found in 23 nuclear mRNP components and co-purifying splicing factors, including Npl3, Nab2, Tho1, Mex67-Mtr2, and components of the TREX complex (Keil et al., 2023) (experiments were performed in collaboration with Prof. Henning Urlaub's and Prof. Katja Strasse's group). From them, Npl3, a serine-arginine (SR) family protein, was selected for further analysis due to its involvement in various post-transcriptional gene regulation functions in yeast. The analysis shows that yeast cells with *NPL3* gene deletions showed severe growth defects at reduced and elevated temperatures (**Figure 4.6 A-B**), indicating the essential role of Npl3. From protein sequence analysis, Npl3 has an N-terminal acidic domain, followed by two RNA recognition motif (RRM) domains connected by a short linker, and an RGG domain. It is homologous to many known eukaryotic SR-like proteins, where RRM1 of Npl3 shows sequence conservation for consensus ribonucleoprotein (RNP) 1 and RNP2 motifs across different species. Also, the structure of Npl3 shows the canonical RRM fold for RRM1 and consists of RNP1 as "N-G-F-A-F-V-E-F" and RNP2 as "L-F-V-R-P-F" motifs for RNA recognition. In contrast, RRM2 shares the highly conserved motif "S-W-Q-D-L-K-D" located on the α1 helix of the structure and lacks consensus RNP motifs similar to other known non-canonical RRMs (**Figure 4.6 A, C**). Thus, RRM1 and RRM2 of Np3 are considered as canonical RRM and non-canonical or pseudo-RRM, respectively.

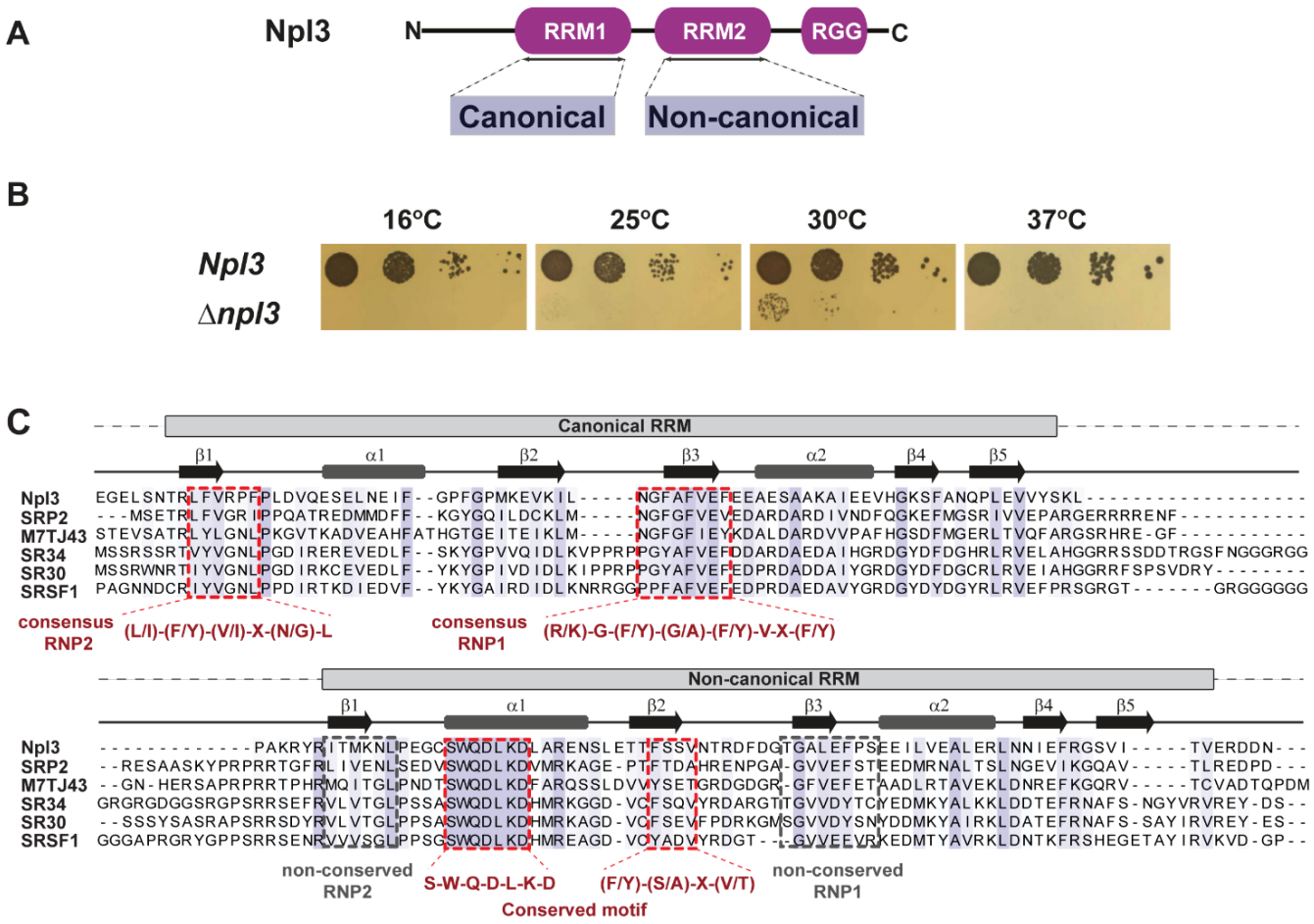


Figure 4.6. Domain architecture and sequence conservation of Npl3 RRM1,2. (A) Domain organization of full-length Npl3 is shown. (B) Deletion of the *Npl3* gene causes a growth defect in yeast. 10-fold serial dilutions of wild-type (wt, NPL3) were spotted onto YPD plates and incubated for 2–3 days at the indicated temperatures (experiments were performed by Philipp Keil). (C) Multiple sequence alignment of RRM domains from SR-like proteins across the various species using Clustal omega and Jalview tools. The red dashed line highlights the alignment of consensus RNP motifs in RRM1 and the non-canonical binding region in RRM2.

4.7 Dynamics analysis of tandem RRMs of Npl3

Structurally, the individual RRM domains of Npl3 have been previously reported without RNA (Deka et al., 2008). However, there is a poor understanding of how each RRM recognizes RNA, the arrangement of the tandem RRM domain (RRM1,2) when free or bound to RNA, and the RNA-binding surface. To better comprehend the structure and dynamics of the tandem RRMs of Npl3 (Npl3$^{120\text{-}280}$), {^{1}H}-^{15}N steady state heteronuclear NOE, ^{15}N R_2 and R_1 relaxation experiments were measured by NMR (**Figure 4.7 A-C**).

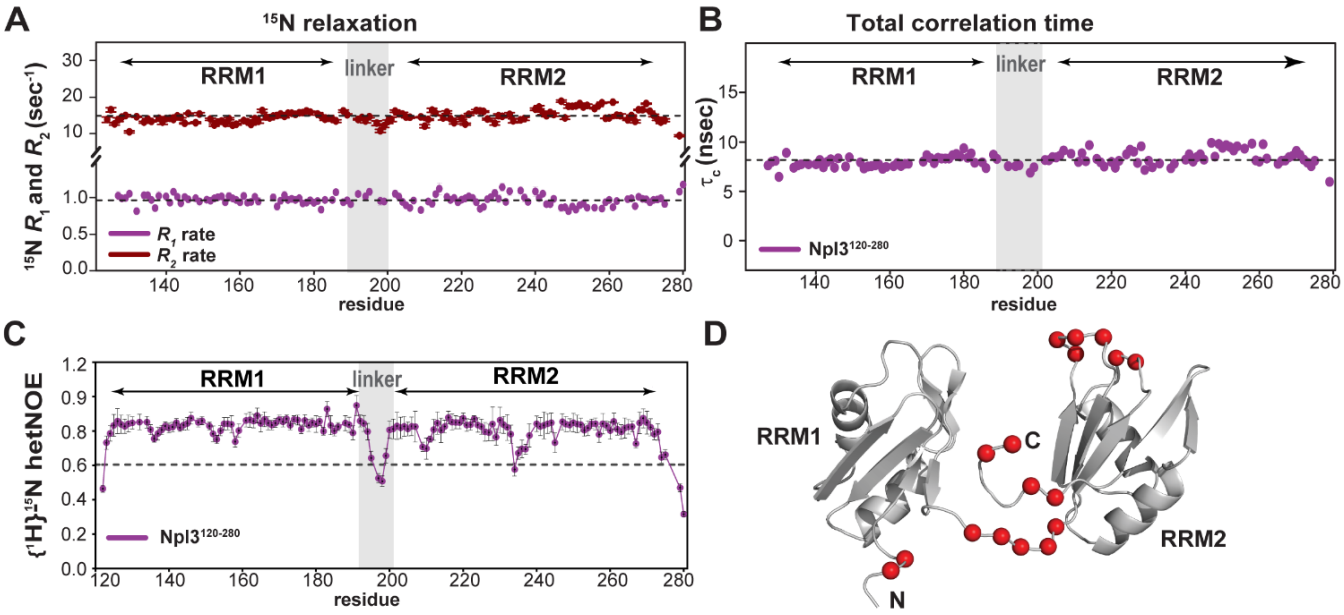


Figure 4.7. Dynamic characterization of Npl3 RRM1,2. (A) NMR ^{15}N longitudinal (R_1) and transverse (R_2) relaxation rates at 900 MHz proton Larmor frequency. Average R_1 and R_2 rates are 0.97 sec^{-1} and 14.8 sec^{-1}, respectively. (B) Residue-specific correlation times (τ_c) calculated from R_1 and R_2 rates. The average correlation time is 8.1 ns. (C) Comparison of (1H)-15N heteronuclear NOE of wild type RRM1,2. Linker between RRM domains shown with gray box. (D) Residues with fast scale dynamics are highlighted on structure with red spheres.

For ^{15}N relaxation rates, R_1 and R_2 experiments were measured by sampling the exponential decay function of delays (more details section 2.27) and the signal intensity decay was fitted to an exponential decay to extract the rate parameters for tandem RRMs of Npl3 (more details section 2.27). The extracted R_1 and R_2 rates suggest the average value of 0.97 sec^{-1} and 14.80 sec^{-1}, respectively for tandem RRs of Npl3 (**Figure 4.7 A**). Combining R_1 and R_2 rates of tandem RRMs, the total correlation times (τ_c) were derived for each residue, indicating an average value of 8.1 nanosec (**Figure 4.7 B**). This correlation time is equivalent to the similar size of single domain protein, suggesting that both RRMs tumble independently. In addition, $\{^1H\}$-^{15}N heteronuclear NOE experiments suggest that Both RRMs are rigid in solutions, except for a few flexible regions on structure, while N- and C-terminal regions are fully flexible. Also, the reduced flexibility observed for the linker connecting the RRMs suggest RRMs oriented in proximity (**Figure 4.7 C, D**).

4.8 Structure of tandem RRMs of Npl3 in free RNA form

To understand the RRM and RNA recognition and regulation, the structural insight of tandem RRMs is essential. Previous studies have reported the three-dimensional structures of Npl3's individual RRMs in the absence of RNA (Deka et al., 2008) (Skrisovska & Allain, 2008). However, the structure of both RRMs together, relative orientation of each domain and dynamic equilibrium of both domains are not explored details. Hence, to determine the structural tandem RRMs domains of Npl3, PRE and SAXS experiments were perfomed. For that, single cysteine point mutations were introduced into the tandem RRMs construct of Npl3 at four specific positions (D135C, E176C, N185C, and D236C) while native Cys211 was mutated to serine. Cys mutants were purified and then attached with an IPSL (3-(2-Iodoacetanido)-PROXYL) spin-label via a stable thioester bond (see method for more details) and spectra were recorded and analyzed between oxidized and reduced state of samples (**Figure 4.8 A, B**) (**Table 4.1**). From these, the intra- and inter-domain distance restraints were derived for structure calculation.

To calculate the structure of the RRM domains, inter-domain restraints were used with a semi-rigid body refinement approach. The individual RRM domains' structural coordinates were obtained from the Protein Data Bank (PDB) accession codes 2OSQ and 2OSR for RRM1 and RRM2, respectively. N- and C-terminal flexible linkers were removed from the structures, and only the rigid core of domains were kept for refinement. Cysteine side-chain coordinates were replaced at specific positions, and spin label IPSL moieties were attached to the cysteine side-chain at position of 135, 176, 185 and 236 residues. A short molecular dynamic (MD) simulation was performed to randomize the N-terminal, C-terminal and linker between two RRM domains, and these coordinates were used as a starting template for structure calculation. Backbone torsion angle restraints were generated using TALOS-N based on secondary chemical shifts, and experimental distance restraints were derived for individual datasets of the PRE experiment. Using these restraints, CNS 1.2 was used for structural refinement, resulting in 100 randomized models. The 15 lowest energy structures were analyzed and compared to experimental PRE and SAXS data, resulting in a final structure with an RMSD of 0.93 Å and an average PRE Q-factor of 0.1 (**Figure 4.9 A, B, C; Figure 4.10 A**). The structure was validated using experimental SAXS data with a chi^2 of 2.7 and a Ramachandran plot, showing 82.8% residues in allowed, 14.8% additionally allowed, 1.6% generously allowed and 0.8% in dis-allowed regions (**Table 4.1**).

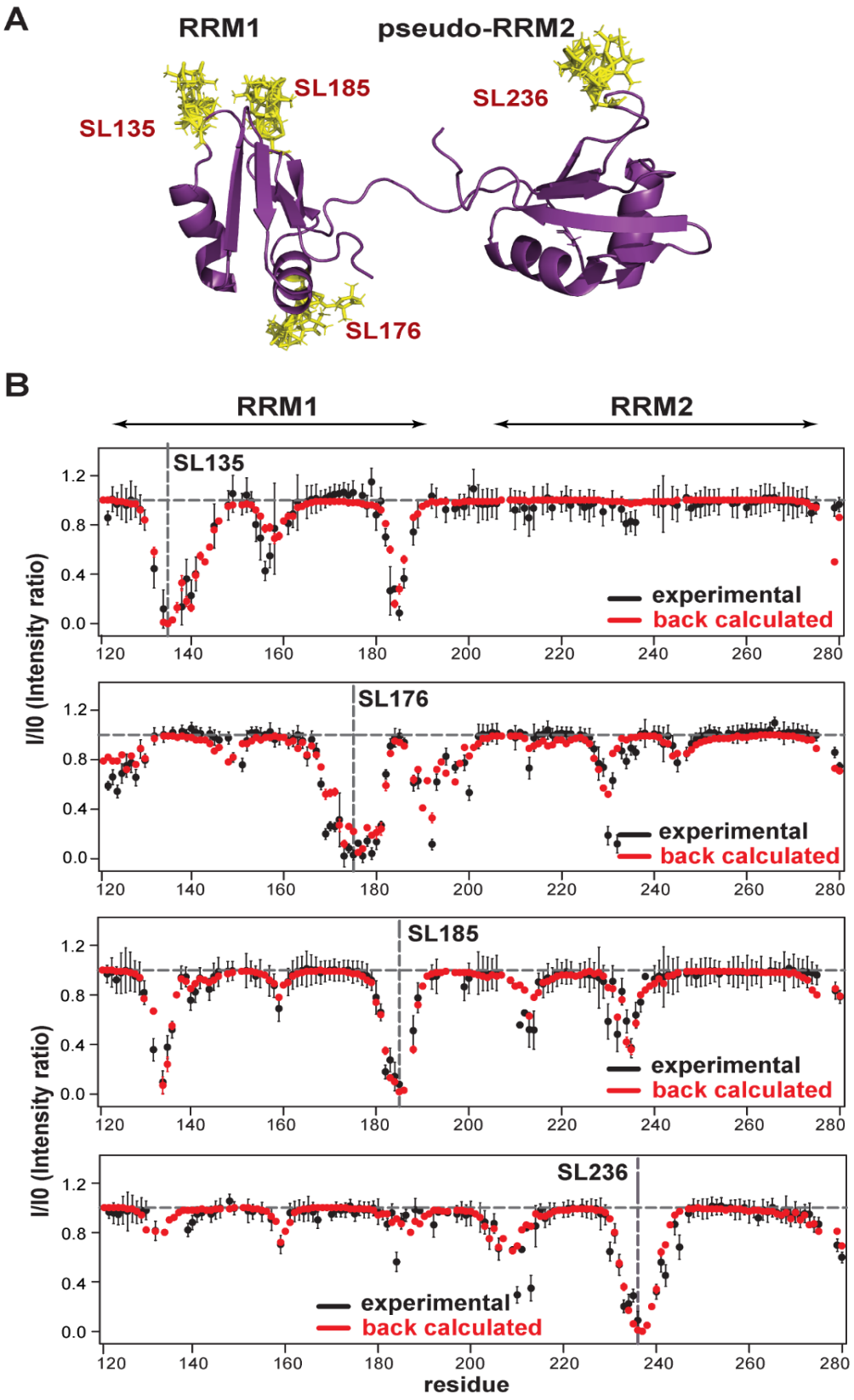


Figure 4.8. PRE analysis of the Npl3 tandem RRM domains. (A) Starting template structure is shown with an ensemble of four copies for each spin label site indicated by yellow sticks. (B)

Paramagnetic relaxation enhancements (PREs) from amide signal intensities in the oxidized and reduced state of the spin-labeled protein (black) vs. PREs back-calculated from the final model (red). Four positions were spin-labeled individually by introducing mono-Cys variants for residue 135, 176,185 and 236.

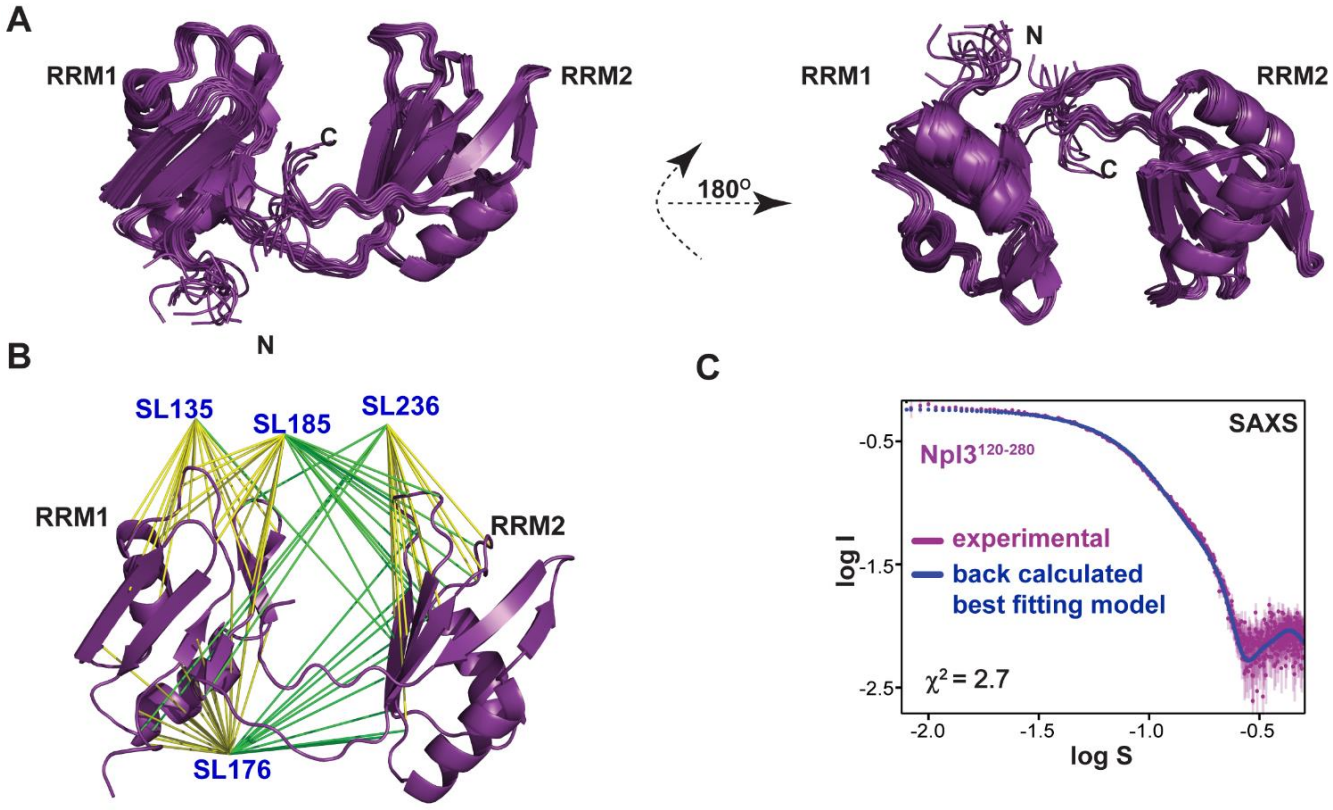


Figure 4.9. Structure of the tandem RRMs of Npl3. (A) Structural model of Npl3 RRM1,2 based on intra- (yellow lines) and inter-RRM (green lines) distance restraints derived from the experimental PRE data. (B) Superposition of the ensemble of 15 lowest energy structures for the tandem RRM1,2 in two different views. (C) Comparison of experimental small angle X-ray scattering (SAXS) data with those back-calculated from the final structure.

The calculated ensemble structural models show that the β-sheet surfaces of the two RRMs face toward each other, forming a positively charged surface (**Figure 4.10 A, B**). The two domains are arranged in close proximity, and the linker connecting the two RRMs has only reduced flexibility (**Figure 4.7**). The positive surface charges assist RRMs together with connecting linkers to recognize the RNA.

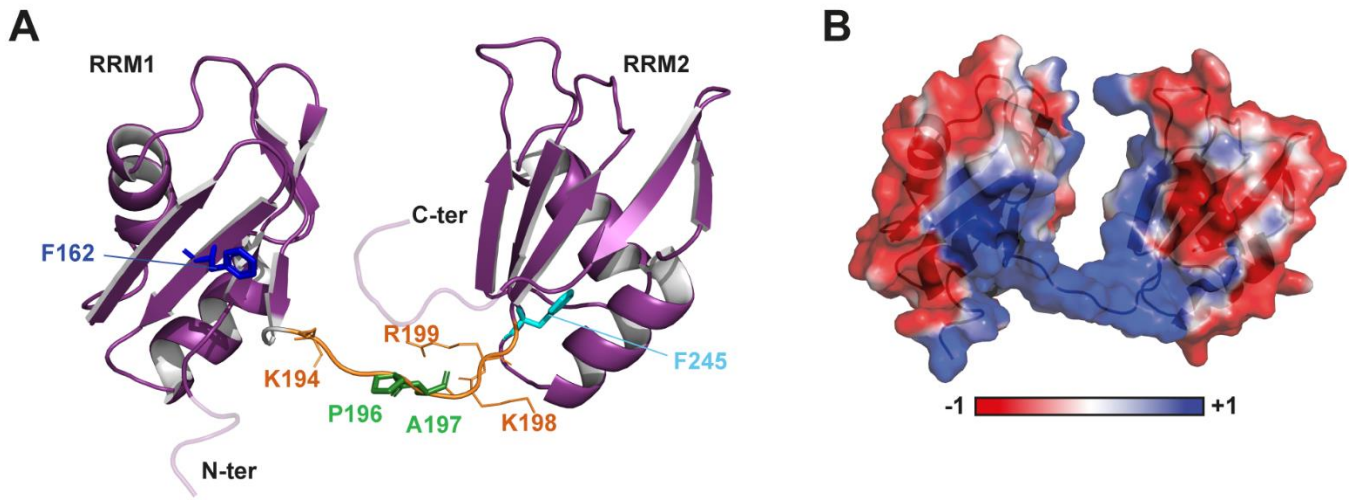


Figure 4.10. Structural analysis of Npl3. (A) Cartoon representation of the NMR-derived structure of the Npl3 tandem RRM domains. The linker connecting the two RRMs is highlighted in orange. (B) Surface representation of the structure colored by electrostatic potential (generated using APBS tool 2.1), blue and red for positive and negative surface charges, respectively.

Table 4.1 Structural statistics

Npl3$^{120-280}$ RRM1,2 free form	**Statistics**
Intra-domain PRE restraints	69
Inter-domain PRE restraints	31
Distance violations	4 (<3 Å)
Average RMSD for 15 lowest energy structures[a]	0.93 Å
PRE quality factor[b]	0.10
Agreement with SAXS data (χ_2)	2.70
Backbone structural quality (Ramachandran plot)	82.8% (allowed) 14.8% (additionally allowed), 1.6 % (generously allowed) 0.8% (dis-allowed)
[a] out of 100 structures calculated [b] see Methods	

4.9 RNA binding specificity of each RRM of Npl3

Next, to understand the RNA-binding preferences and specificity for the individual RRM domains of Npl3, NMR-based binding experiments were carried out using ^{15}N labeled protein with a wide range of short oligonucleotide motifs and chemical shift perturbation (CSP) were

calculated. For that, single-stranded DNA oligonucleotides were selected as the first proxies for RNA binding to screen a range of diverse sequences, as shown in (**Figure 4.11 A-E; Figure 4.12. A**) (**Table 4.2**). It has been shown before that ssDNA ligands can well represent the RNA recognition by RRMs, especially in the case of SRSF1.

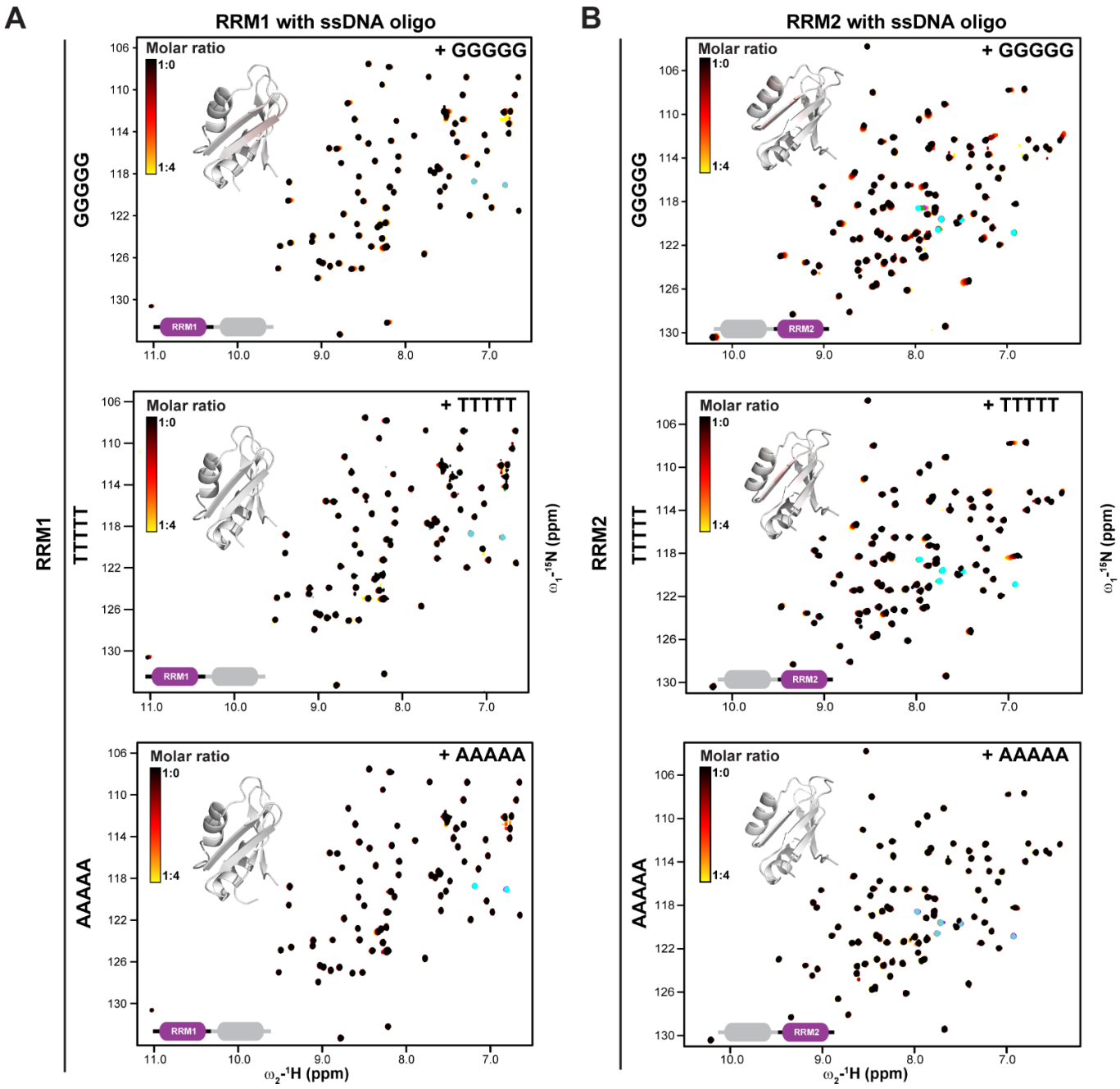
The NMR binding experiments show that cytosine-rich “CC” motif-containing ligands show significant chemical shift perturbation (CSP) at the canonical RNP1 and RNP2 RNA-binding surface of the RRM1 domain, while no considerable binding is observed for the RRM2 domain. In contrast, guanosine-rich “GG” motif-containing ligands strongly bind to RRM2 but not to RRM1 (**Figure 4.11 A; Figure 4.12. A**). The CSP with “GG” motif was observed at the helix α2 and strand β2 region of RRM2, a non-canonical RNA binding interface. This is consistent with the fact that Npl3 shares high sequence conservation with SRSF1 and other SR proteins, which exhibit similar features for RRM1 and RRM2 (**Figure 4.6 B**).

Overall, the analysis concludes that RRM1 has a strong preference for “CN”-type motifs, including “CCC”. At the same time, RRM2 prefers “GG”-type motifs, including “UGG” motif resembling the RNA-binding preference of the non-canonical RRM in the homologous SRSF1 protein (Cléry et al., 2013) (Cléry et al., 2021) (Deka et al., 2008).

Table 4.2. List of oligonucleotides used NMR binding studies

Ligand	domains	oligo	NMR CSP[b]	Binding[c]
DNA oligonucleotides (single stranded)[a]	RRM1	TTTTT	0.01	no
		GGGGG	0.05	no
		AAA AA	0.01	no
		CCCCC	0.44	yes
		AGCCCC	0.31	yes
		AGCACC	0.08	yes
	RRM2	TTTTT	0.11	yes
		CCCCC	0.00	no
		AAA AA	0.01	no
		GGGGG	0.08	yes
		GGGGAGA	0.05	no
		GTGGGGA	0.06	no
		GTGGAGA	0.41	yes
		GTAAAGA	0.11	yes
	RRM1,2	AGCACCGTGGAGA	0.34	yes
[a] DNA oligos were used as a proxy for RNA				

[b] Maximal NMR chemical shift perturbations observed at 4-fold excess of the oligonucleotides [c] Above a CSP threshold ≤ 0.08



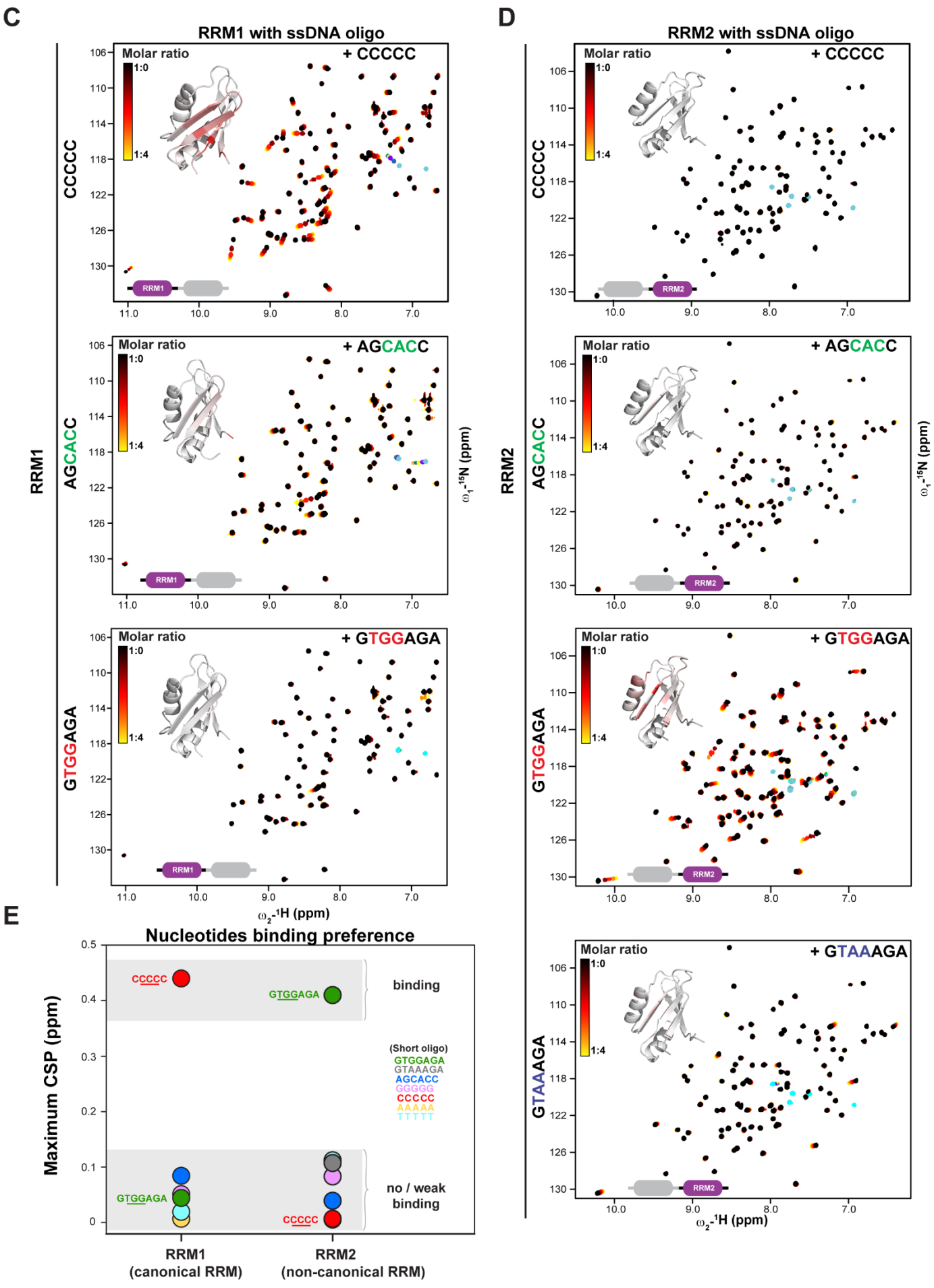

C
RRM1 with ssDNA oligo
RRM1
CCCCC
AGCACC
GTGGAGA
Molar ratio
1:0
1:4
+ CCCCC
+ AGCACC
+ GTGGAGA
RRM1
ω_2-^{1}H (ppm)
ω_1-^{15}N (ppm)
D
RRM2 with ssDNA oligo
RRM2
CCCCC
AGCACC
GTGGAGA
GTAAAGA
+ CCCCC
+ AGCACC
+ GTGGAGA
+ GTAAAGA
E
Nucleotides binding preference
Maximum CSP (ppm)
CCCCC
GTGGAGA
binding
(Short oligo)
GTGGAGA
GTAAAGA
AGCACC
GGGGG
CCCCC
AAAAA
TTTTT
no / weak binding
RRM1
(canonical RRM)
RRM2
(non-canonical RRM)

Figure 4.11. NMR titrations to assess the RNA-binding in preference of Npl3. Superposition of NMR ^{1}H-^{15}N correlation spectra of (A, C) RRM1 (left) and (B, D) RRM2 (right) titrated with various single-stranded DNA oligonucleotides. Spectra are colored black (free) and red to yellow (with increasing ligand concentration). Chemical shift perturbations (CSP) are mapped (red) onto the structure of the RRM domains. (E) Binding preferences for RRM1 (left) and RRM2 (right) based on maximum CSPs.

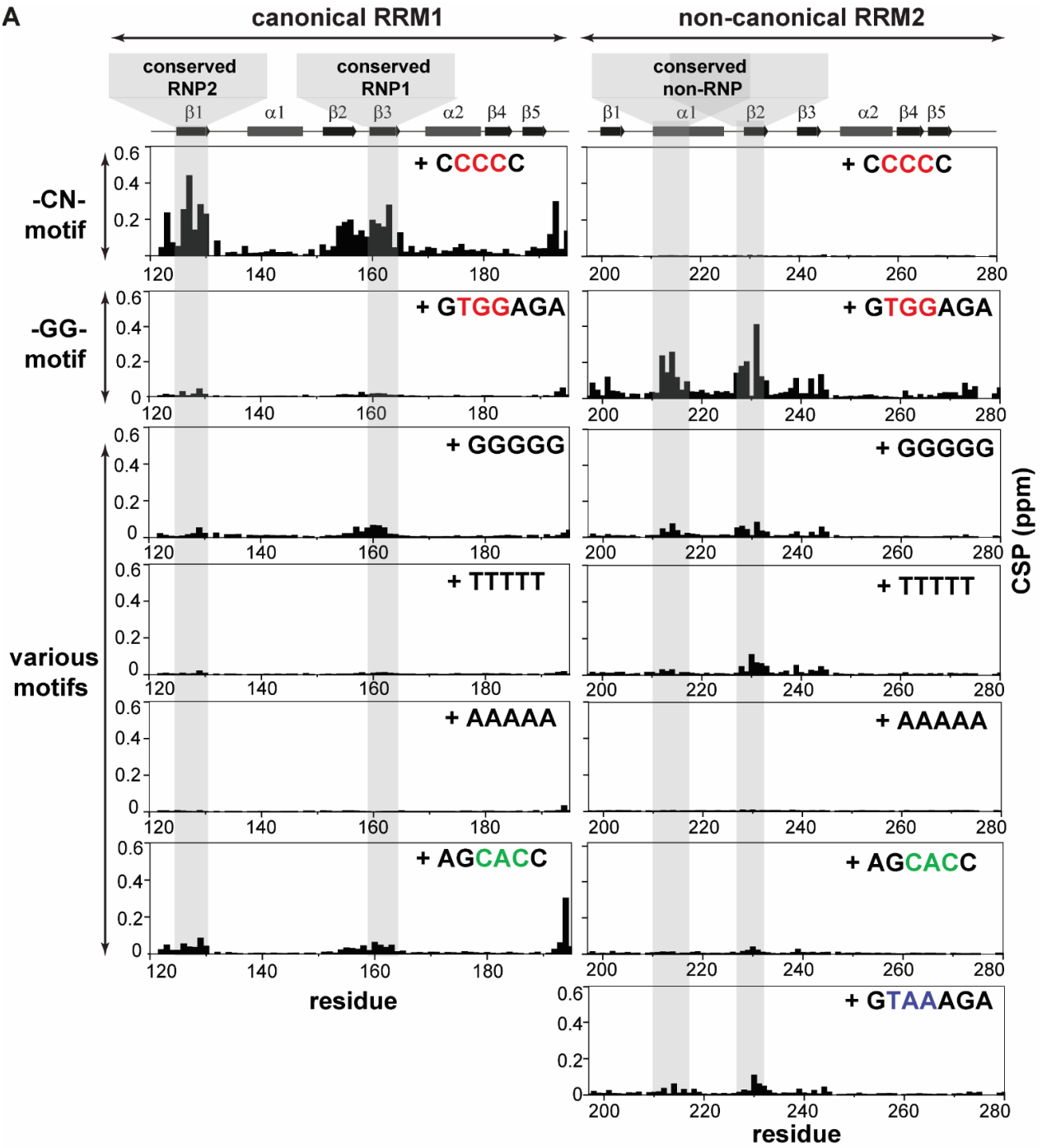


Figure 4.12. NMR CSP to assess the RNA-binding preference of Npl3 RRM1,2. (A) Summary of the CSPs. RNP1, RNP2 and non-canonical conserved sites are highlighted with gray boxes.

4.10 Characterization of RNA binding tandem RRMs of Npl3

To characterize the RNA-binding activity of the tandem RRM domain and the role of the linker, ITC, and NMR titration experiments were performed using RRM1,2 of Npl3$^{120\text{-}260}$ with two RNA sequences, which harbor the “CC” and “GG” nucleotide binding motifs for RRM1 and RRM2, respectively. Specifically, binding to CN--GG (5’-AGCACCGUGGAGA-3’) and a variant CN--AA (5’-AGCACCGUAAAGA-3’) were tested, where RNA binding by RRM2 is expected to be strongly reduced (**Figure 4.13 A**).

Titration suggests that the “CN--AA” oligo binds to both RRMs with modest affinity as NMR signals shift with increasing concentration of the RNA ligand, and saturation is obtained only at 4-fold molar excess (**Figure 4.13 B**). CSPs are observed for both RRMs and the linker region, indicating that this RNA interacts with both RRMs, and the canonical RNP motifs in RRM1 are most strongly affected (**Figure 4.13 C**). The positively charged residues K194, K198, and R199 in the linker between the RRMs also show significant perturbation, demonstrating that the linker contributes to RNA binding. However, the interaction of tandem RRMs (RRM1,2 of Npl3$^{120\text{-}280}$) with the CN--AA RNA was not detectable in ITC experiments indicating weak binding **(Figure 4.13 D**) (**Table 4.4**). This is confirmed by NMR titration experiments, which show an average dissociation constant (K_D) for the interaction of $K_D \approx 150$ μM (**Figure 4.13 E**).

In contrast, the binding to the CN--GG RNA is significantly stronger, consistent with NMR titrations showing binding kinetics in the intermediate to slow exchange regime (**Figure 4.13 B right, C lower panel**). Spectral changes map to the same binding surface seen for the CN--AA RNA (**Figure 4.13 B, C**). ITC shows high-affinity binding of the CN--GG RNA with Npl3$^{120\text{-}280}$ with $K_D = 0.66$ μM (**Figure 4.13 D**) (**Table 4.4**). Interestingly, the RNA-binding region maps to the β-sheets of the canonical RNP sites in RRM1, the non-canonical conserved regions in RRM2 and the positively charged surface in Npl3$^{120\text{-}280}$ (**Figure 4.10 B; Figure 4.6 B**).

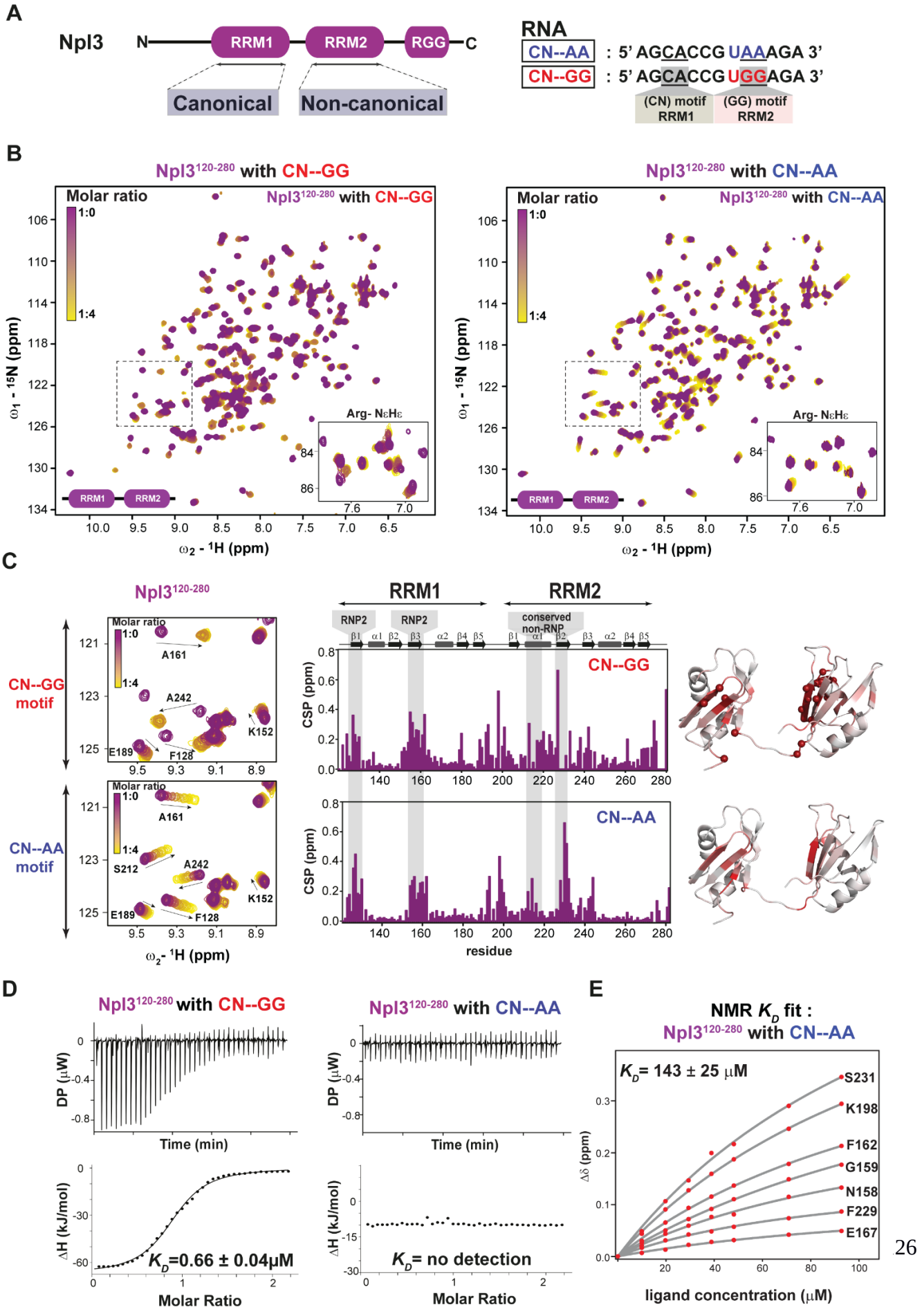

A
Npl3
N
RRM1
RRM2
RGG
C
Canonical
Non-canonical
RNA
CN--AA : 5' AGCACCG UAAAGA 3'
CN--GG : 5' AGCACCG UGGAGA 3'
(CN) motif RRM1
(GG) motif RRM2
B
Npl3120-280 with CN--GG
Npl3120-280 with CN--AA
Molar ratio
1:0
1:4
ω1 - 15N (ppm)
ω2 - 1H (ppm)
Arg- NεHε
RRM1
RRM2
C
Npl3120-280
CN--GG motif
CN--AA motif
A161
A242
K152
E189
F128
S212
RRM1
RRM2
RNP2
conserved non-RNP
CSP (ppm)
residue
D
Npl3120-280 with CN--GG
Npl3120-280 with CN--AA
DP (μW)
Time (min)
ΔH (kJ/mol)
KD=0.66 ± 0.04μM
KD= no detection
Molar Ratio
E
NMR KD fit :
Npl3120-280 with CN--AA
KD= 143 ± 25 μM
S231
K198
F162
G159
N158
F229
E167
Δδ (ppm)
ligand concentration (μM)

Figure 4.13. Npl3 RNA binding and structural analysis by NMR and ITC. (A) Domain organization of full-length Npl3 and the tandem RRM domains used for NMR and binding studies. The two 13-mer RNAs used for binding studies (CN–AA and CN–GG) are indicated below. The key nucleotides in the RNA-binding motifs of RRM1 and RRM2 are underlined. (B) A) ^{15}N labeled wild-type RRM1,2 (purple) titrated with “CN--AA” RNA (left) and the “CN--GG” RNA (right). Increasing chemical shift changes are colored from purple, green, dark or light blue (free form) to yellow (RNA-bound form). NMR signals of arginine NεHε side chains are shown as insets. (C) Zoom version of superimpose spectra (left), CSP plot (middle) and RNA-binding surface area are shown for wildtype RRM1,2 (purple) with “CN--AA” (upper) and “CN-_GG” (lower). Conserved sequence motifs (RNP1, RNP2 in RRM1 and the non-canonical site in RRM2) are highlighted with gray boxes. The inset shows a mapping of CSPs (red) onto the tandem RRM1,2 structure. (D) ITC binding curves for wt tandem RRM domains of Npl3 with “CN–GG” and “CC-AA” RNAs are shown. (E) K_D calculated from NMR titrations (right) upon the addition of ”CN--AA” RNA to Npl3 RRM1,2.

4.11 Structural model of RNA bound Npl3

To derive a structural model of the protein–RNA complex, PRE experiments were measured with spin labels attached to 185C (on RRM1) and 236C (on RRM2) positions in the presence of 3-fold excess of ‘CN–GG’ RNA. PRE data suggest that PRE position at 185C (on RRM1) shows intra-domain PRE, while no inter-domain PRE was observed. While potion 236C (on RRM2) shows intra- and inter-domain PRE **(Figure 4.14 A)**. Based on PRE, protein-RNA distance restraints were obtained from chemical shift perturbations seen in NMR titration and based on the homologous structure (PDB: 2M8D and 5DDR). CNS1.2 was used to generate a pool of 400 models, which were then scored against experimental SAXS data (**Figure 4.14 C, D**). To do this, theoretical SAXS curves were generated using CRYSOL from the ATSAS software package 3.0.0 and compared with the experimentally measured SAXS data. The final structural model of the protein–RNA complex shows a χ^2 of 1.9 with the experimental SAXS data (**Figure 4.14 D**, **E)**. The structural model indicates that the RNA-bound Npl3 stays extended where “CC” motif binds to the RRM1 and “GG” motif binds to non-canonical RRM2. Additionally, the PRE position at 185C shows inter-domain PRE effect for RNA free form of Npl3 while, RNA-bound form does not have any inter-domain PRE effect, indicating the RRM2 domain is further away from RRM1 in RNA-bound form compared to RNA free form. However, SAXS data indicated that the overall shape for Npl3 in RNA-free and RNA-bound Npl3 remains the same (**Figure 4.14 B**).

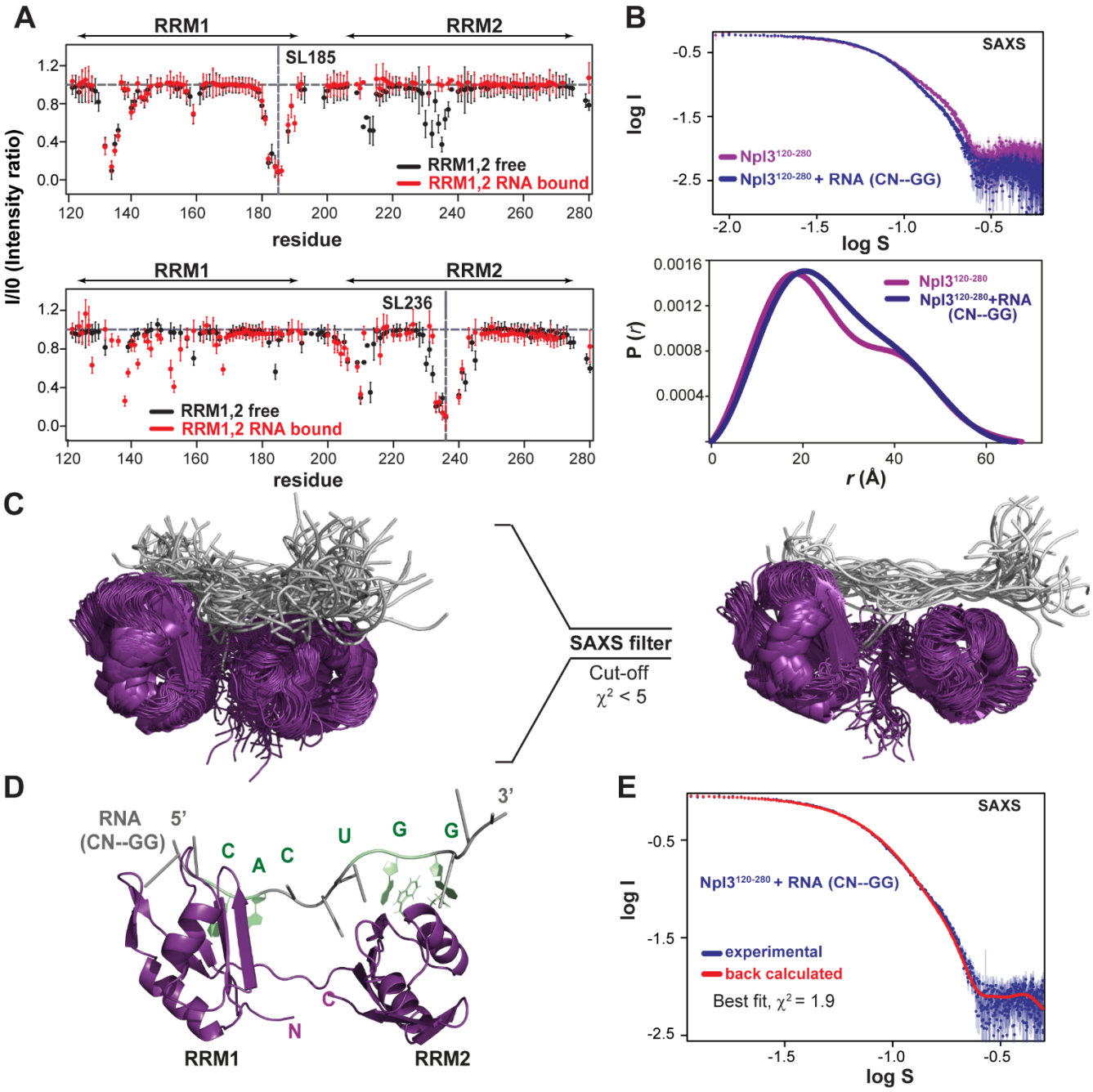


Figure 4.14. Structural model of RNA bound RRM1,2 domains of Npl3. (A) PRE data of spin-labeled Npl3 RRM1,2 bound to CN--GG RNA (red) compared to PRE data of the free protein (black) for spin labels at position 185C (RRM1) and 236C (RRM2). (B) Experimental SAXS data of free and "CN--GG" RNA-bound Npl3 RRM1,2. (C) Superimpose ensemble RNA-bound Npl3 models (left) and SAXS-filtered models are shown. (D) PRE and SAXS-derived representative structural model of the Npl3 tandem RRM domains with the "CN--GG" RNA. (E) Experimental and back-calculated SAXS data from the RNA-bound model on $Npl3^{120-280}$ are shown.

Table 4.3. Small-angle X-ray scattering (SAXS) of free and RNA-bound RRM1,2 of Npl3

Sample	R_g (Angstrom)	D_{max} (Angstrom)
Npl3$^{120-280}$	20.4 ± 0.12	67.7
Npl3$^{120-280}$ + “CN—GG”	20.4 ± 0.16	63.6
npl3$^{120-280}$-Linker	19.0 ± 0.10	66.5

Next, a structural model of the Npl3$^{120–280}$ complex with the high-affinity CN–GG RNA ligand was derived based on NMR and SAXS data (**Figure 4.15 A; Figure 4.14 C-D**), which rationalizes the RNA-binding features observed. The overall RRM domain arrangement is very similar to the one observed in the absence of RNA. As expected from the sequence conservation, the RNA-binding interfaces and interactions of the RRM domains strongly resemble the corresponding interactions in the human SRSF1 protein.

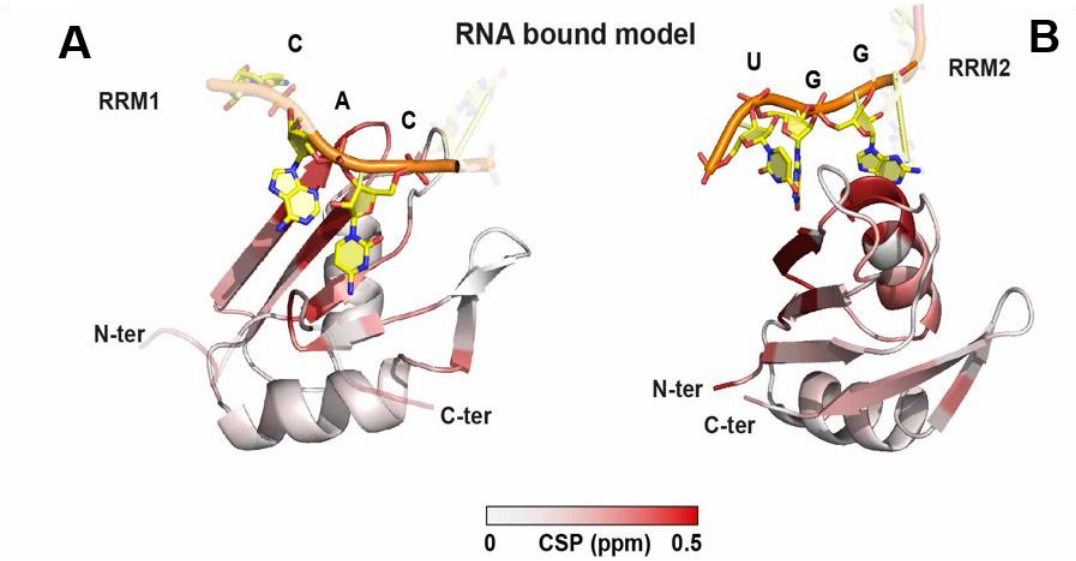


Figure 4.15. Structural analysis of individual RRM-RNA complex of Npl3. (A) NMR and SAXS-derived structural model of the complex of Npl3 with CN–GG RNA where RRM1 and RRM2 recognize a CAC motif (left) and UGG motif (right), respectively.

4.12 Structural analysis of Npl3 mutants

According to the results of UV-crosslinks and mass spectromentry analysis, RNA and amino acids crosslink in specific regions of Npl3, including the RRM domains (F162 of RRM1 and F229/S230 of RRM2), the linker region between two RRMs (P196 and A197), and the RGG motif. These regions appear to play a role in RNA binding. To investigate further, mutations

were created in Npl3 protein coding genes and their impact on yeast cell growth was assessed. The npl3-RRM1 (F162Y), npl3-Linker (P186D, A197D), and npl3-RRM2 (F245I) mutants showed growth defects, particularly at reduced and elevated temperatures, although less severe than a complete deletion of NPL3. Interestingly, combining the RRM1 and Linker mutations resulted in a stronger yeast growth defect (**Figure 4.16 A**). The combination of the RRM1 and RRM2, the Linker and RRM2 as well as all three mutations in one protein leads to lethality (**Figure 4.16 B**). Interestingly, strains with any of the three combinations of mutations grew worse than the NPL3 deletion strain, suggesting they may be dominant negative (**Figure 4.16 C**). These growth defects could be caused by a decrease in RNA-binding activity or by the combined defect in binding to different classes of RNAs by the different domains. Taken together, Npl3 mutants lead to a growth defect *in vivo* which displays the differential regulatory roles of each domain in cell.

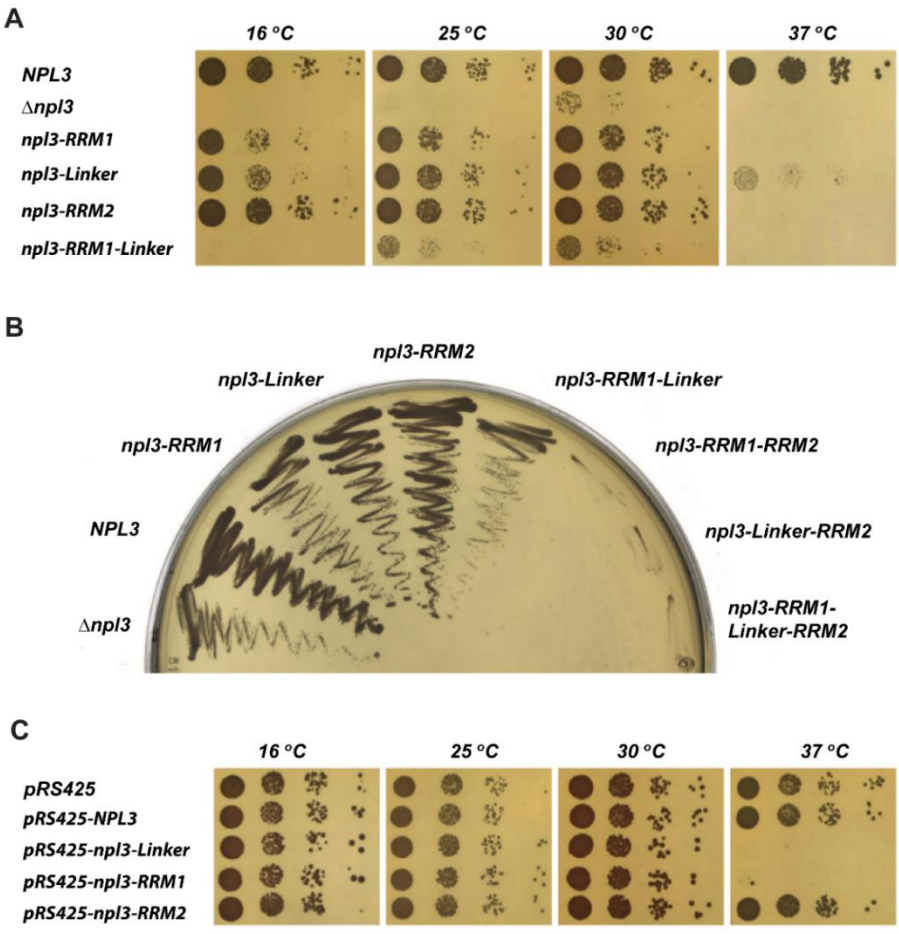


Figure 4.16. Dot plot assay of yeast cells carring wildetype and mutants of Npl3. (A)The mutants of Npl3 cause a growth defect *in vivo*. A serial dilution of wildtype and mutants of yeast cells were spotted on YPD plate at indicated temperature. (B) Combination of Npl3 mutants

cause synthetic lethality. A shuffle stain with deletion of genomic *Npl3* gene covered by a *URA3*-plasmid encoding wildtype or mutants of *Npl3* gene was transformed and streaked onto an FOA-plate to shuffle out the URA3-plasmid and incubated for 3 days at 30 °C. (C) Overexpression of npl3-RRM1, npl3-Linker, and npl3-RRM2 in yeast cell are shown as a dominant negative growth phenotype at the indicated temperature. A serial dilutions of wt cells transformed with the high-copy plasmid pRS425 encoding the indicated Npl3 mutatns were spotted onto SDC(-leu) plates and incubated for 3 days at the indicated temperature (experiments were performed by Philipp Keil).

These three Npl3 mutations (npl3-RRM1, npl3-linker and npl3-RRM2) that tested *in vivo* map to the positively surface charged region including the linker on the structure. To assess the impact of mutations on the structure of Npl3, the NMR spectra of the Npl3 (120-280 aa) fragment were measured *in vitro* and compared with the wild-type $Npl3^{120\text{-}280}$. The $npl3^{120\text{-}280}$-RRM1 mutation (F162Y) did not cause significant spectral changes, indicating that it does not affect the tandem RRM domains' fold (**Figure 4.17 A, B, left panel**). The $npl3^{120\text{-}280}$-Linker mutation (P196D, A197D) only caused notable spectral changes in the linker's proximity, but the overall structure remained largely undisturbed (**Figure 4.17 B, middle panel**), indicating that the overall structure is largely unperturbed. However, a comparison of the SAXS data for $Npl3^{120\text{-}280}$ and $npl3^{120\text{-}280}$-Linker (**Figure 4.17 D**) (**Table 4.3**) indicates that the RRM1,2 domain arrangement is slightly more compact in the linker mutant, perhaps as a consequence of replacing the more rigid and extended P196 residue. In contrast, the $npl3^{120\text{-}280}$-RRM2 mutation (F245Y) severely affected the RRM2 fold, impairing protein functions involving RRM2, while the NMR signals for the RRM1 are largely unaffected (**Figure 4.17 A, B, right panel**).

The overall analysis concludes that $npl3^{120\text{-}280}$-RRM1 and $npl3^{120\text{-}280}$-Linker mutants maintain the overall structure of the tandem RRMs, while $npl3^{120\text{-}280}$-RRM2 strongly affects the RRM2 fold and is thus expected to impair protein functions involving RRM2.

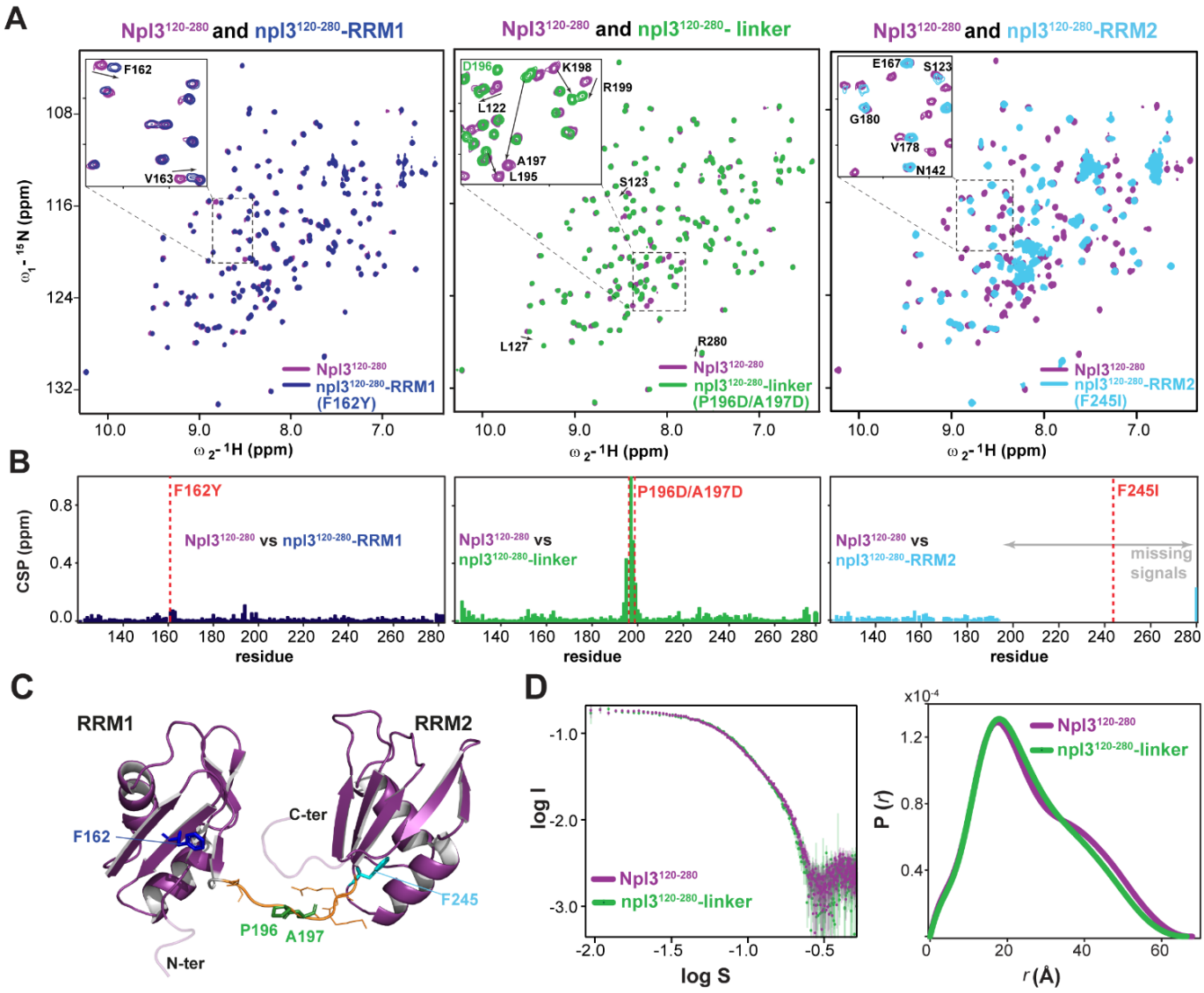


Figure 4.17. Structural effects of mutations on RRM1,2 of Npl3$^{120-280}$ by NMR and SAXS. (A) NMR ^{1}H-^{15}N correlation spectra (HSQC) of wildtype RRM1,2 (purple) superimposed with the RRM1 mutant (dark-blue), linker mutant (green) and RRM2 mutant (cyan), respectively. (B) Chemical shift perturbation (CSP) between spectra of wildtype Npl3$^{120-280}$ with npl3-RRM1, npl3$^{120-280}$-linker and npl3$^{120-280}$-RRM2 mutants, respectively. Red dashed lines indicate the mutated sites. (C) Mutation sites are highlighted on the Npl3 RRM1,2 structure. (D) SAXS data (left) and distance distribution function P(r) (right) for wild-type Npl3$^{120-280}$ and the linker mutant.

4.13 *In vitro* RNA binding analysis of Npl3 mutants

Npl3$^{120-280}$ complex with the high-affinity “CN--GG” RNA ligand shows that the overall RRM domain compactness remains the same as in RNA-free form (**Figure 4.14 B**). The in-vivo study showed that the combination of npl3-RRM1 (F162Y), npl3-Linker (P196D, A197D) and npl3-RRM2 (F245Y) mutants show severe growth defects in yeast at reduced and elevated

temperature. To study in detail, the effect of the npl3 mutations on RNA binding *in vitro* was investigated by NMR and ITC.

The linker mutation was found to have reduced RNA-binding affinity due to two aspartate residues in the mutant linker causing charge clashes. This resulted in a decrease in optimal contacts with the RNA and a small but significant change in the RRM domain distances (as seen by SAXS, **Figure 4.17 D**). Further studies were conducted by titrating RNA ligands to the npl3$^{120\text{-}280}$-Linker mutant fragment. It was observed that there were virtually no spectral changes with the "CN--AA" RNA, while the "CN--GG" RNA showed a strongly reduced interaction compared to wildtype Npl3$^{120\text{-}280}$ (**Figure 4.18; Figure 4.19 green panel**). These results are consistent with ITC experiments where no binding is detected (**Figure 4.19 D; Table 4.4**). Taken together, the data demonstrate that the linker mutation does not affect the structural integrity of RRMs, but it does strongly reduce RNA-binding affinity.

The npl3$^{120\text{-}280}$-RRM1 mutant shows reduced RNA binding with K_D =10 μM for the "CN—GG" RNA, thus 15-fold reduced compared to wt Npl3$^{120\text{-}280}$ (**Figure 4.18; Figure 4.19, dark blue panel, Table 4.4**). The npl3$^{120\text{-}280}$-RRM2 mutant spectrum shows the F245Y mutation destabilizes RRM2 fold (**Figure 4.17 A, B, light blue**). Also, npl3$^{120\text{-}280}$-RRM2 mutant shows 10-fold reduced RNA-binding affinity (K_D = 5 μM) for the "CN--GG" RNA by ITC, compared to wildtype Npl3$^{120\text{-}280}$ (**Figure 4.18 C; Figure 4.19 A, light blue panel**). In summary, the *in-vitro* finding shows that point mutations in the RRM1 (F162Y), linker (P196D, A197D), or RRM2 (F245Y) regions of Npl3 strongly reduce RNA-binding activity for both RNA sequences tested.

Table 4.4. ITC and NMR binding assays for the protein-RNA interactions

Protein	**RNA**	**K_D (μM)**	**N**	**ΔH (kJ/mol)**	**TΔS (kJ/mol)**	**ΔG (kJ/mol)**
Npl3$^{120\text{-}280}$	CN--GG	0.66 ± 0.04	0.9	-67.5 ± 0.8	-32.2	-35.3
npl3$^{120\text{-}280}$-Linker	CN--GG	Not detected	-	-	-	-
npl3$^{120\text{-}280}$-RRM1	CN--GG	10.4 ± 2	0.97	-127 ± 15.6	-98.2	-28.5
npl3$^{120\text{-}280}$-RRM2	CN--GG	4.8 ± 0.3	1.08	-118 ± 3.2	-87.8	-30.4
Npl3$^{120\text{-}280}$	CN--AA	Not detected	-	-	-	-
npl3$^{120\text{-}280}$-Linker	CN--AA	Not detected	-	-	-	-
Npl3$^{120\text{-}280}$	CN--AA	143 ± 25	(by NMR titration)			

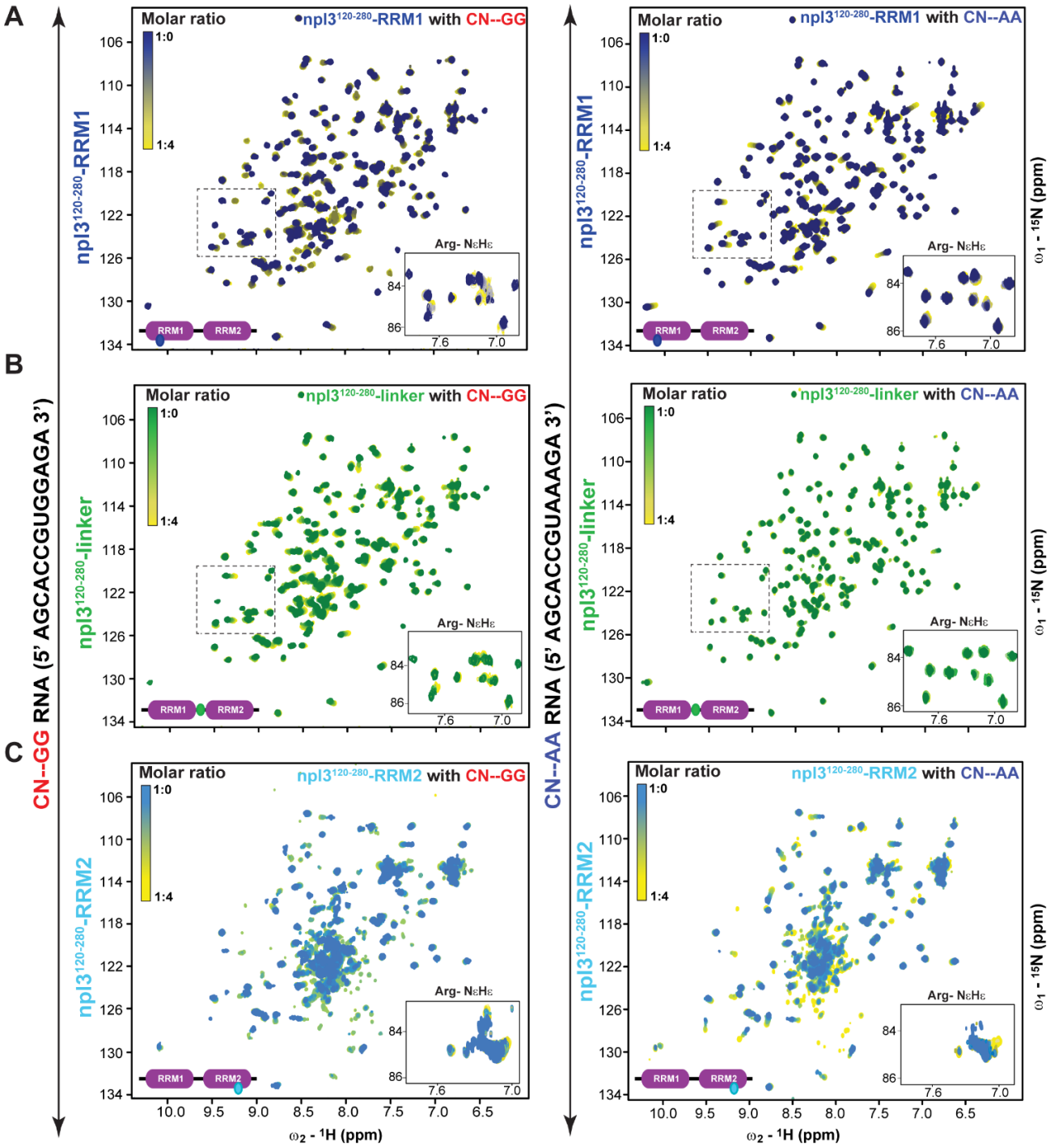


Figure 4.18. RNA binding with various mutants of Npl3$^{120\text{-}280}$ by NMR. Superposition of NMR ^{1}H-^{15}N correlation spectra of (A) npl3$^{120\text{-}280}$-RRM1 (dark blue), (B) npl3$^{120\text{-}280}$-Linker mutant (green) and (C) npl3$^{120\text{-}280}$-RRM2 (light blue) mutants titrated with "CN--AA" RNA (left) and the "CN--GG" RNA (right). Increasing chemical shift changes are colored from purple, green, dark or light blue (free form) to yellow (RNA-bound form). NMR signals of arginine NεHε side chains are shown as insets.

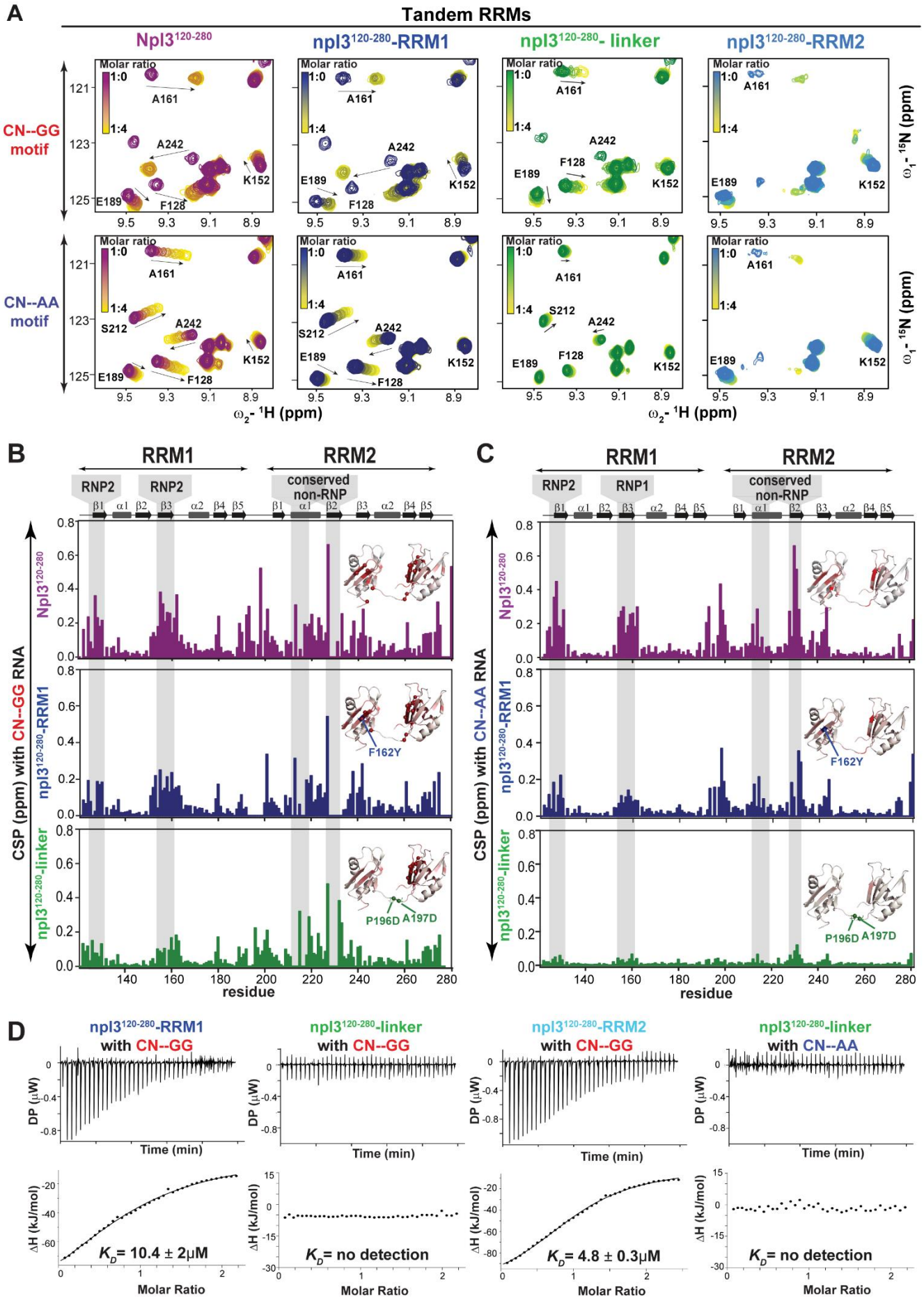

A
Tandem RRMs
Npl3120-280
npl3120-280-RRM1
npl3120-280- linker
npl3120-280-RRM2
CN--GG motif
CN--AA motif
Molar ratio
1:0
1:4
A161
A242
K152
E189
F128
S212
ω1- 15N (ppm)
ω2- 1H (ppm)
B
C
RRM1
RRM2
RNP2
RNP1
conserved non-RNP
CSP (ppm) with CN--GG RNA
CSP (ppm) with CN--AA RNA
F162Y
P196D A197D
residue
D
DP (μW)
Time (min)
ΔH (kJ/mol)
Molar Ratio
KD= 10.4 ± 2μM
KD= no detection
KD= 4.8 ± 0.3μM

Figure 4.19. RNA binding, structural analysis and effect of Npl3 mutations. (A) Overlay of zoom version of ^{1}H-^{15}N NMR correlation spectra of ^{15}N-labeled wt RRM1,2 (purple), npl3$^{120–280}$-RRM1 (dark blue), npl3$^{120–280}$-linker (green) and npl3$^{120–280}$-RRM2 (cyan) mutants titrated with the “CN--GG” (upper panel) and “CN--AA” (lower panel) RNAs, respectively. Increasing chemical shift changes are colored from purple, green, dark or cyan (free form) to yellow (RNA bound form), respectively. (B) CSP plot for wildtype (purple), npl3-RRM1 (dark blue) and npl3 linker (green) with “CN--AA” (left) and “CN--GG” (right) RNAs from spectra shown in Figure 4.3.12. Conserved sequence motifs (RNP1, RNP2 in RRM1 and the non-canonical site in RRM2) are highlighted with gray boxes. The inset shows mapping of CSPs (red) onto the tandem RRM1,2 structure.

4.14 *In vivo* RNA binding analysis of Npl3 mutants

To evaluate the functional impact of three mutations, the efficiency of nuclear mRNA export was tested. This involved performing RNA fluorescence in-situ hybridization against poly(A) tails to visualize the overall mRNA distribution. Despite reduced RNA-binding activity in all three mutants, they exhibited varying phenotypes (**Figure 4.20 A**). At 30°C, npl3-RRM1 cells did not display any mRNA export defect and only a minor one at 37°C. However, npl3-Linker cells showed a strong mRNA export defect even at 30°C, which worsened at 37°C. npl3-RRM2 cells exhibited an intermediate phenotype, with a mild defect at 30°C and a stronger one at 37°C (**Figure 4.16 A**). In summary, the strength of the mRNA export defect does not correlate with the severity of the growth defect, indicating that other processes than mRNA export are probably impaired in these mutants.

Next, to determine the composition of nuclear mRNPs in cells, they were purified using endogenously TAP-tagged Cbc2. Then Western blot assays were carried out to determine the amount of co-purifying Npl3, as well as six other nuclear mRNP components. These components included the THO/TREX subunits Hpr1, Sub2, and Yra1, the nuclear poly(A)-binding protein Nab2, Tho1, and the mRNA exporter subunit Mex67 (**Figure 4.20 B**). The analysis demonstrates that the *in vivo* RNA-binding activity of Cbc2 in npl3-RRM1 and npl3-Linker cells is similar to wildtype, and comparable amounts of nuclear mRNPs co-purify with Cbc2 in these two mutants. However, the amount of Npl3 co-purifying with nuclear mRNPs is decreased in the npl3-RRM1 and npl3-Linker mutants. The amount of Hpr1 is decreased only in the npl3-RRM1 mutant, while the abundance of Hpr1, Sub2, Tho1, Yra1, and Mex67 is decreased in nuclear mRNPs of npl3-Linker cells. This change in the composition of nuclear mRNPs is consistent with the mRNA export defect observed in npl3-Linker cells (**Figure 4.20 B**). The three npl3 mutants impact nuclear mRNA export and nuclear mRNP composition

differently, with the strongest effects observed for the npl3-Linker mutant. This suggests that they have different effects on gene expression processes, possibly due to divergent RNA sequence preferences of the three RNA-binding sites.

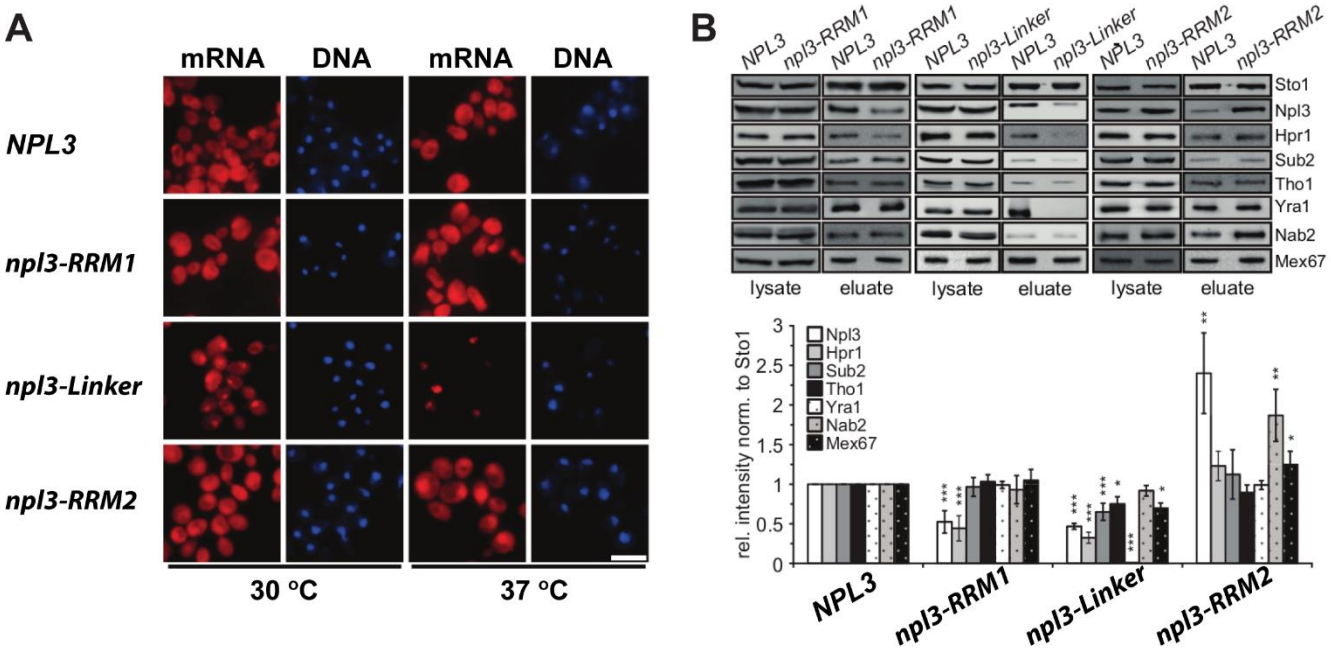


Figure 4.20. Mutations in Npl3 that decrease RNA-binding activity cause different effects on nuclear mRNA export and nuclear mRNP composition. (A) npl3 mutant cells show different degrees of mRNA export defect. The localization of bulk mRNA was visualized by in situ hybridization with fluorescently labeled oligo(dT) in wt and npl3 mutant cells grown at 30°C or shifted to 37°C for one hour. DNA was stained with DAPI. (B) Nuclear mRNP composition changes in the three different npl3 mutants. Western blots (upper panel) and quantification of three independent experiments (lower panel) of nuclear mRNPs purified via Cbc2-TAP purification from wt, npl3-RRM1, npl3-Linker and npl3-RRM2 cells. The amounts of Sto1, Npl3, Hpr1, Sub2, Tho1, Yra1, Nab2 and Mex67 were quantified and normalized to the amount of the CBC subunit Sto1. Values for wt cells were set to 1. (Experiments were performed by Philipp Keil).

Discussion

The RRMs and RGG domains of Npl3 are known to have RNA-binding activity. RRM1 adopts a canonical fold, while RRM2 lacks the conserved RNP1 and RNP2 motifs and features a conserved sequence motif 'SWQDLKD' in helix α1, characteristic of a non-canonical, so-called pseudo-RRM domain. The study presents a structural model for the tandem RRM domains that shows a large positively charged surface of Npl3 including the linker connecting the two RRM domains. The RNA-binding studies show distinct binding preferences for the two RRM domains. The canonical RRM1 recognizes a "CC" motif and the non-canonical RRM2 shows specificity for a "GG" motif. Accordingly, the binding affinity of the tandem RRM1,2 domains to "CN–AA" RNA (N indicates any nucleotide) is significantly lower (150-fold) compared to a "CN–GG" RNA ligand. Interestingly, the "AGCACCGUGGAGA" RNA binds to RRM1-RRM2 with 5' to 3' orientation, unlike to most other tandem RRM domains. The two RRM domains are partially prearranged for RNA binding, and, consistently, the overall domain arrangement of the free and RNA-bound tandem RRM domains is relatively similar.

The short linker shows some flexibility in the absence of RNA, potentially allowing fine-tuning of the domain arrangement to optimize RNA recognition. Npl3 is homologous to human SRSF1, and the analysis confirms that the RNA-binding preferences for RRM1 and RRM2 are comparable to human SRSF1. The data also show that the linker connecting RRM1 and RRM2 strongly contributes to RNA binding. Notably, the overall reduction in RNA-binding affinity upon linker mutation observed *in vitro* correlates well with effects observed *in vivo*. The increased flexibility by replacing the proline affects the domain orientation as indicated by SAXS data and together with the introduction of negatively charged aspartates rationalizes the reduced RNA-binding affinity.

UV-crosslinking and mass spectromentry analysis based derived mutants RRM1 (F162), linker (P196, A197) and RRM2 (F245Y) show the distinct growth phenotype *in vivo*. Hence, *in vitro* experiments were explored to characterize the mutants. From ITC analysis, RRM1 mutation reduces RNA binding (by 15-fold) compared to the wt protein, suggesting an important role for the tyrosine hydroxyl and altered stacking interactions. The mutation in RRM2 leads to domain unfolding, which rationalizes the significantly reduced binding affinity (~7 fold) to "CN–GG" RNA. Interestingly, no well-defined RNA-binding motif has been identified *in vivo*. This may reflect that differential contributions of the two tandem domains of Npl3 enable binding to

distinct substrates depending on the process.

In summary, NMR and biochemical data demonstrate the critical role of the linker connecting the two RRMs, the conserved sequence motifs in RRM1 and the non-canonical RRM2 for RNA binding. Although mRNA binding of all three mutant proteins is reduced in vitro and *in vivo*, the functional consequences differ. The yeast cell growth for the npl3 RNA-binding mutants is affected to varying extents, likely resulting from the sum of the different processes impaired in these mutants. For example, splicing efficiency, nuclear mRNP composition and mRNA export are differentially affected in the three mutants. Npl3 is the only SR-like protein that promotes splicing in *S. cerevisiae* through co-transcriptional recruitment of the U1 and U2 snRNPs to chromatin.

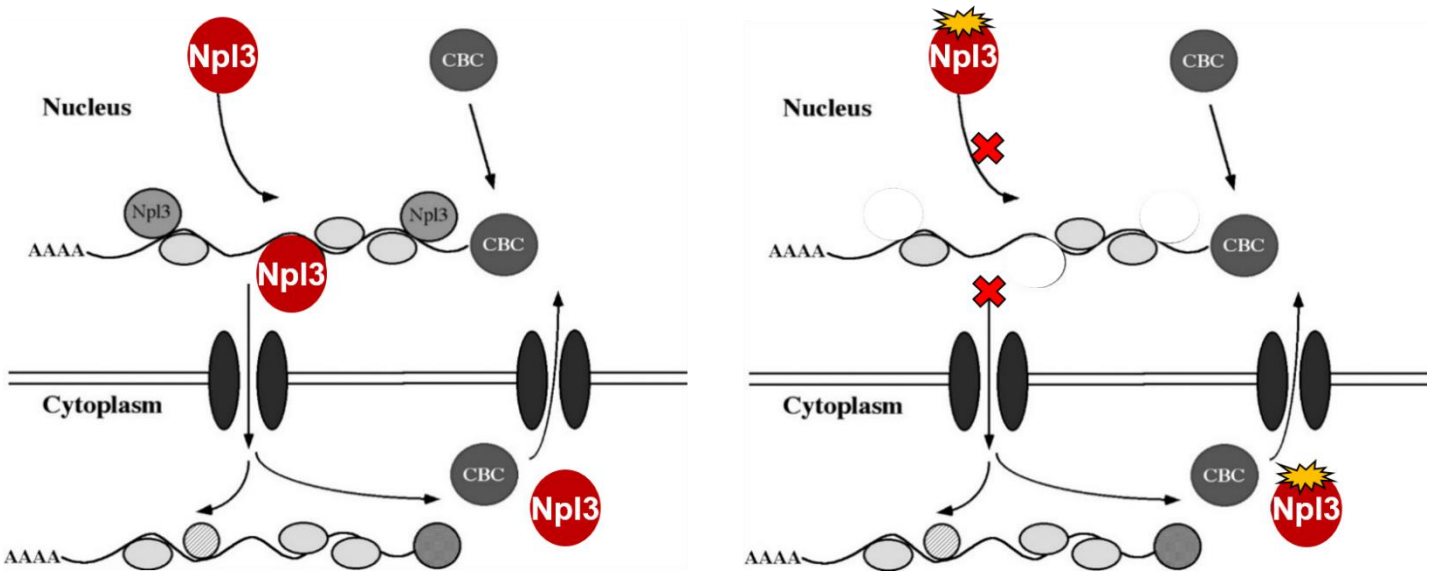


Figure 4.21. Role of Npl3 in nuclear mRNA recognition and mRNP assembly. Npl3 Wildtype and mutant are illustrated in the left and right panels, respectively.

Chapter 5 - Summary and Outlook

Summary and Outlook

Studying multi-domain RNA binding proteins and complexes such as the SF1-U2AF2 complex involved in 3’ splice site recognition or the nuclear mRNA assembly protein Npl3 requires solution-based biophysical methods which can provide a detailed understanding of RNA binding specificity, domain structure rearrangement, and dynamic regulations upon the RNA binding.

In Chapter 3, SF1's crucial role in recognizing canonical branch point sites through its KH and Qua2 domains is highlighted. The part of this work demonstrates how disease variants in branch sites can alter subsequent splicing machinery and affect the SF1 binding. However, it remains unclear how disease variants affect branch site recognition in the later stage of the splicing cycle when U2 hnRNP replaces SF1 in complex A of splicing cycle. As a future aspect, detailed structural studies with atomistic details could help in developing an understanding of the mechanistic role of splicing factors and associated proteins. U2AF2 iCLIP data showed that SF1 has a stabilizing effect that enhances the efficiency of 3’ splice site recognition. However, a more detailed analysis along with other splicing factors is required to understand the spliceosome assembly.

The individual domain structures for SF1 and U2AF2 have previously been identified. In the current study, the ensemble structure of SF1 having an N-terminal part of the domains, including ULM, HH, KH and Qua2 were characterized. In addition, the ensemble multiple-domain structures of U2AF2 having RRM1, RRM2, and connecting flexible linker to UHM were characterized in RNA-free and RNA with polypyrimidine tract sequences. SF1 and U2AF2 proteins exhibit inter-domain flexibility in RNA-free form, while the SF1-U2AF2 complex adapts characteristic open to compact conformations in the presence of RNA types that have variable splice sites, distances between them, or multiple BPS-like splice sites. The high-resolution structural models suggest that KHQua2 and RRM1,2 come closer in RNA-bound states to form compact structures. However, additional structural studies are needed to understand the splicing assembly fully. Therefore, the cryoEM structure of the entire complex E and complex A in the early stage of spliceosome assembly could provide more insights into how splice sites are recognized and which regulatory factors are involved. The complex system of complex E and A with disease variants of splice sites would provide additional details to our understanding.

Chapter 4 of this study explores the structure and functions of the yeast SR-like protein, Npl3, in mRNP assembly and mRNA export. The current study utilized UV-light crosslinking and mass spectrometry analysis to identify more than 100 RNA binding sites within 16 RNA binding proteins involved in mRNP formation, including Npl3, Nab2, Tho1, Mex67-Mtr2, and components of the TREX complex. The study also identified several RNA binding sites for Npl3 and conducted *in vivo* and *in vitro* experiments to investigate their functional study. The *in vitro* findings show that the RRM1 and pseudo-RRM2 recognize the "CC" and "GG" motifs, respectively. The study also looked into the tandem RRM structure and RNA binding surface area, although a more accurate crystal or NMR structure is needed to understand the structure and function better. Additionally, *in vivo* data revealed a novel role for Npl3 by mutating the linker residues, demonstrating Npl3's involvement in the recruitment and transfer of nuclear mRNP components to mRNA. Also, *in vitro* NMR and ITC data highlight reduced binding for linker mutants compared to wild-type Npl3 RRMs. While the current study focused on RNA recognition by RRMs and RRM mutants, the role of the N-terminal acidic linker and C-terminal RGG motif are not included. Thus, further analysis with the full-length protein would add more details to understanding the role of Npl3 for mRNP assembly. Npl3 also undergoes phosphorylation and methylation as post-translational modifications, and further experiments are needed to understand their *in vivo* and *in vitro* roles. Investigating the association of Npl3 with other RBPs in nuclear mRNPs would be interesting to understand how Npl3 contributes to their regulation.

This study demonstrates that combining NMR, SAXS and ITC is the most suitable and complementary solution methods to study the regulation of RBPs to study structure, dynamics, domain rearrangement, binding affinity and mapping of binding surface area.

Abbreviations

Short name	Full name
AA or aa	Amino acid
APS	Ammonium persulfate
ATP	Adenosine 5’-triphosphate
BMRB	Biological Magnetic Resonance Bank
BPS	Branch point site
CaCl2	Calcium Chloride
CCPN	Collaborative Computational Project for NMR
CN--AA RNA	RNA oligo “5’- AGCACCGUAAAGA -3’”
CN--GG RNA	RNA oligo “5’- AGCACCGUGGAGA -3’”
CSP	Chemical shift perturbation
CTP	Cytidine 5’-triphosphate
DTT	Dithiothreitol
E. coli	*Escherichia coli*
EDTA	Ethylene diamine tetra acetic acid
EOM	Ensemble Optimisation Method
ESI-MS	Electronspray Ionization Mass spectromentry
GTP	Guanosine-5'-triphosphate
HiSQC	$H_{\varepsilon}N_{\varepsilon}$-selective heteronuclear in-phase single quantum coherence
HSQC	Heteronuclear Single Quantum Coherence
IPSL	3-(2-Iodoacetamido)-PROXYL
IPTG	Isopropyl β- d-1-thiogalactopyranoside
ITC	Isothermal calorimetry titration
K_D	Dissociation constant
k_{ex}	Exchange rate constant
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIS kinase	Kinase interacting stathmin kinase
LB media	Luria Broth or Luria-Bertani medium
M9 media	M9 minimal medium

MD simulation	Molecular Dynamic simulation
$MgCl_2$	Magnesium Chloride
mRNA	Messenger RNA
MS	Mass spectrometry
MW	Molecular weight
NaPO buffer	Sodium phosphate buffer (Na_2HPO_4+NaH_2PO_4)
NMR	Nuclear Magnetic resonance
Npl3	Nucleolar protein 3
$Npl3^{120-280}$	Npl3 RRM1,2
$npl3^{120-280}$-linker	Npl3 RRM1,2 construct with linker mutation at P196D and A197D
$npl3^{120-280}$-RRM1	Npl3 RRM1,2 construct with RRM1 mutation at F162Y
npl3-RRM2	Npl3 RRM1,2 construct with RRM1 mutation at F245Y
OD	Optical density
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PPM	Parts per million
PPT	Polypyrimidine tract
PRE	Paramagnetic relaxation enhancement
Pre-mRNA	Precursor messenger RNA
pSF1	Phosphorylated splicing factor 1
Q-factor	Quality factor
R_1	Longitudinal relaxation rate
R_2	Transverse relaxation rate
RBD	RNA binding domain
RBP	RNA binding protein
RMSD	Root mean square deviation
RNA	Ribonucleic acid
RPM	Revolutions per minute
SAXS	Small angle X-ray scattering
SDS	Sodium dodecyl sulfate or sodium lauryl sulfate
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis

SEC-SAXS	Size-exclusion chromatography with small-angle X-ray scattering
SF1	Splicing factor 1
SIL	Segmental Isotope labeling
SL	Spin label
SR protein	Serine/arginine-rich protein
ssDNA	Single-stranded deoxyribonucleic acid
TauC (τ_C)	Total rotational correlation time
TEMED	N,N,N',N'-Tetramethylethylenediamine
TEV protease	Tobacco Etch Virus Protease
Tris-base	Tris (hydroxymethyl) aminomethan base
Tris-HCl	Tris hydrochloride
U2AF2	U2 auxiliary factor 2
UHM	U2AF homology motif
ULM	U2AF ligand motif
UTP	Uridine 5′-triphosphate
3’ ss	3’ splice site
5’ ss	5’ splice site

List of Tables

List of Figures

List of the papers published

1) Philipp Keil[#], Alexander Wulf[#], **Nitin Kachariya[#]**, Samira Reuscher, Kristin Hühn, Ivan Silbern, Janine Altmüller, Mario Keller, Ralf Stehle, Kathi Zarnack*, **Michael Sattler***, Henning Urlaub*, Katja Sträßer*. Npl3 functions in mRNP assembly by recruitment of mRNP components to the transcription site and their transfer onto the mRNA, *Nucleic Acids Research 2023*, 51(2): 831–851.

2) Sophie Vieweg, Katie Mulholland, Bastian Bräuning, **Nitin Kachariya**, Yu-Chiang Lai, Rachel Toth, Pawan Kishor Singh, Ilaria Volpi, **Michael Sattler**, Michael Groll, Aymelt Itzen, Miratul M. K. Muqit. PINK1-dependent phosphorylation of Serine111 within the SF3 motif of Rab GTPases impairs effector interactions and LRRK2-mediated phosphorylation at Threonine72. *Biochem J. 2020*; 477 (9): 1651–1668.

Acknowledgment

I am grateful to Prof. Michael Sattler for giving me an opportunity to work with him and engaging in detailed discussions on various project aspects. This experience has helped me to learn, build, and explore scientific skills during my Ph.D. journey. I want to thank TUM docGS graduate school for enabling my academic enrollment. Also, I extend my appreciation to TUM and Helmholtz Zentrum München for providing laboratory space, instruments, and IT facilities, and to the Bavarian NMR Center (BNMRZ) for offering state-of-the-art NMR facility services. I acknowledge SFB1035, SPP1935, GRP1721, and SPP2191 funding for financial and traveling support.

I am grateful to Niki Messini for her assistance with sample preparation and crystal setup. My thanks go to Dr. Sam Asami and Dr. Gerd Gemmecker for allocating NMR time and providing experimental support when necessary. I also appreciate Dr. Ralf Stehle and Dr. Matthias Brandl for their assistance with setting up SAXS measurements and analysis support. I am thankful to the staff members at DESY (Deutsches Elektronen-Synchrotron) and ESRF (European Synchrotron Radiation Facility) for providing their expertise and support with the SAXS measurements. A special thanks to Dr. Prince Prabhu Rajaiah for SAXS measuremetns and crystal setup.

I am also thankful to Dr. Saba Suladze, Zahra Harati Taji, Niki Messini, Christoph Muller-Hermes, Masood Aziz, and Olga Sieluzycka for having enjoyable discussions during lunch or tea time. I thank Dr. Mark Bostock, Dr. Peijian Zou, Dr. Arie Geerlof, Dr. Andre Mourao, Dr. Alisha Jones, Dr. Ana Messias, Dr. Filipe Menezes, Dr. Rajlaxmi Panigrahi, Dr. Mohanraj Gopalswami, and Dr. Sheeja Vasudevan for great scientific discussions and suggestions. I would like to thank Dr. Andreas Schlundt for his assistance during the initial phase of my Ph.D.

and his suggestions that always helped me whenever I encountered any trouble. I am also grateful to Dr. Winfried Meining for providing support for IT facilities and to Waltraud Wolfson, Karen Biniossek, and Gulden Yilmaz for taking care of administrative processes, which made my life a lot easier. I highly appreciate Dr. Riddhiman Sarkar, Amit Devra, and Dr. Tejaswini Pradhan for having general discussions and for providing a nice atmosphere. Last, I thank all other Sattler lab members, BNMRZ and HMGU members for sharing wonderful times during my Ph.D. journey.

References

Ackermann, K., Chapman, A., & Bode, B. E. (2021). A Comparison of Cysteine-Conjugated Nitroxide Spin Labels for Pulse Dipolar EPR Spectroscopy. *Molecules*, *26*(24). https://doi.org/10.3390/molecules26247534

Agrawal, A. A., Salsi, E., Chatrikhi, R., Henderson, S., Jenkins, J. L., Green, M. R., Ermolenko, D. N., & Kielkopf, C. L. (2016). An extended U2AF65–RNA-binding domain recognizes the 3′ splice site signal. *Nature Communications*, *7*(1), 10950. https://doi.org/10.1038/ncomms10950

Aguirre, C., Cala, O., & Krimm, I. (2015). Overview of Probing Protein-Ligand Interactions Using NMR. *Curr Protoc Protein Sci*, *81*, 17 18 11-17 18 24. https://doi.org/10.1002/0471140864.ps1718s81

Barnes, C. A., Starich, M. R., Tjandra, N., & Mishra, P. (2021). Simultaneous measurement of (1)H(C/N)-R(2)'s for rapid acquisition of backbone and sidechain paramagnetic relaxation enhancements (PREs) in proteins. *J Biomol NMR*, *75*(2-3), 109-118. https://doi.org/10.1007/s10858-021-00359-9

Becker, W., Bhattiprolu, K. C., Gubensak, N., & Zangger, K. (2018). Investigating Protein-Ligand Interactions by Solution Nuclear Magnetic Resonance Spectroscopy. *Chemphyschem*, *19*(8), 895-906. https://doi.org/10.1002/cphc.201701253

Berglund, J. A., Chua, K., Abovich, N., Reed, R., & Rosbash, M. (1997). The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell*, *89*(5), 781-787. https://doi.org/10.1016/s0092-8674(00)80261-5

Blech-Hermoni, Y., & Ladd, A. N. (2013). RNA binding proteins in the regulation of heart development. *Int J Biochem Cell Biol*, *45*(11), 2467-2478. https://doi.org/10.1016/j.biocel.2013.08.008

Brand, K., Dugi, K. A., Brunzell, J. D., Nevin, D. N., & Santamarina-Fojo, S. (1996). A novel A–>G mutation in intron I of the hepatic lipase gene leads to alternative splicing resulting in enzyme deficiency. *Journal of Lipid Research*, *37*(6), 1213-1223. https://doi.org/https://doi.org/10.1016/S0022-2275(20)39151-3

Brosey, C. A., & Tainer, J. A. (2019). Evolving SAXS versatility: solution X-ray scattering for macromolecular architecture, functional landscapes, and integrative structural biology. *Curr Opin Struct Biol*, *58*, 197-213. https://doi.org/10.1016/j.sbi.2019.04.004

Bryant, R. G. (1983). The NMR time scale. *Journal of Chemical Education*, *60*(11), 933. https://doi.org/10.1021/ed060p933

Burrows, N. P., Nicholls, A. C., Richards, A. J., Luccarini, C., Harrison, J. B., Yates, J. R., & Pope, F. M. (1998). A point mutation in an intronic branch site results in aberrant splicing of COL5A1 and in Ehlers-Danlos syndrome type II in two British families. *Am J Hum Genet*, *63*(2), 390-398. https://doi.org/10.1086/301948

Busch, A., & Hertel, K. J. (2012). Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdiscip Rev RNA*, *3*(1), 1-12. https://doi.org/10.1002/wrna.100

Cartegni, L., Chew, S. L., & Krainer, A. R. (2002). Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet*, *3*(4), 285-298. https://doi.org/10.1038/nrg775

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B. M., Strein, C., Davey, N. E., Humphreys, D. T., Preiss, T., Steinmetz, L. M., Krijgsveld, J., & Hentze, M. W. (2012). Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, *149*(6), 1393-1406. https://doi.org/10.1016/j.cell.2012.04.031

Chen, J., & Weiss, W. A. (2015). Alternative splicing in cancer: implications for biology and therapy. *Oncogene*, *34*(1), 1-14. https://doi.org/10.1038/onc.2013.570

Cheng, M. H., & Jansen, R. P. (2017). A jack of all trades: the RNA-binding protein vigilin. *Wiley Interdiscip Rev RNA*, *8*(6). https://doi.org/10.1002/wrna.1448

Cléry, A., Krepl, M., Nguyen, C. K. X., Moursy, A., Jorjani, H., Katsantoni, M., Okoniewski, M., Mittal, N., Zavolan, M., Sponer, J., & Allain, F. H. T. (2021). Structure of SRSF1 RRM1 bound to RNA reveals an unexpected bimodal mode of interaction and explains its involvement in SMN1 exon7 splicing. *Nature Communications*, *12*(1), 428. https://doi.org/10.1038/s41467-020-20481-w

Cléry, A., Sinha, R., Anczuków, O., Corrionero, A., Moursy, A., Daubner, G. M., Valcárcel, J., Krainer, A. R., & Allain, F. H.-T. (2013). Isolated pseudo–RNA-recognition motifs of SR proteins can regulate splicing using a noncanonical mode of RNA recognition. *Proceedings of the National Academy of Sciences*, *110*(30), E2802-E2811. https://doi.org/doi:10.1073/pnas.1303445110

Clore, G. M. (2011). Exploring sparsely populated states of macromolecules by diamagnetic and paramagnetic NMR relaxation. *Protein Sci*, *20*(2), 229-246. https://doi.org/10.1002/pro.576

Coppin, L., Leclerc, J., Vincent, A., Porchet, N., & Pigny, P. (2018). Messenger RNA Life-Cycle in Cancer Cells: Emerging Role of Conventional and Non-Conventional RNA-Binding Proteins? *Int J Mol Sci*, *19*(3). https://doi.org/10.3390/ijms19030650

Crisci, A., Raleff, F., Bagdiul, I., Raabe, M., Urlaub, H., Rain, J.-C., & Krämer, A. (2015). Mammalian splicing factor SF1 interacts with SURP domains of U2 snRNP-associated proteins. *Nucleic Acids Research*, *43*(21), 10456-10473. https://doi.org/10.1093/nar/gkv952

Crisci, A., Raleff, F., Bagdiul, I., Raabe, M., Urlaub, H., Rain, J. C., & Kramer, A. (2015). Mammalian splicing factor SF1 interacts with SURP domains of U2 snRNP-associated proteins. *Nucleic Acids Res*, *43*(21), 10456-10473. https://doi.org/10.1093/nar/gkv952

da Costa, P. J., Menezes, J., & Romao, L. (2017). The role of alternative splicing coupled to nonsense-mediated mRNA decay in human disease. *Int J Biochem Cell Biol*, *91*(Pt B), 168-175. https://doi.org/10.1016/j.biocel.2017.07.013

Da Vela, S., & Svergun, D. I. (2020). Methods, development and applications of small-angle X-ray scattering to characterize biological macromolecules in solution. *Current Research in Structural Biology*, *2*, 164-170. https://doi.org/https://doi.org/10.1016/j.crstbi.2020.08.004

De Conti, L., Baralle, M., & Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA*, *4*(1), 49-60. https://doi.org/10.1002/wrna.1140

Deka, P., Bucheli, M. E., Moore, C., Buratowski, S., & Varani, G. (2008). Structure of the yeast SR protein Npl3 and Interaction with mRNA 3'-end processing signals. *J Mol Biol*, *375*(1), 136-150. https://doi.org/10.1016/j.jmb.2007.09.029

Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., & Bax, A. (1995). NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *Journal of Biomolecular NMR*, *6*(3), 277-293. https://doi.org/10.1007/BF00197809

Dmitri, I. S., & Michel, H. J. K. (2003). Small-angle scattering studies of biological macromolecules in solution. *Reports on Progress in Physics*, *66*(10), 1735. https://doi.org/10.1088/0034-4885/66/10/R05

Feng, C., Kovrigin, E. L., & Post, C. B. (2019). NmrLineGuru: Standalone and User-Friendly GUIs for Fast 1D NMR Lineshape Simulation and Analysis of Multi-State Equilibrium Binding Models. *Sci Rep*, *9*(1), 16023. https://doi.org/10.1038/s41598-019-52451-8

Franke, D., Petoukhov, M. V., Konarev, P. V., Panjkovich, A., Tuukkanen, A., Mertens, H. D. T., Kikhney, A. G., Hajizadeh, N. R., Franklin, J. M., Jeffries, C. M., & Svergun, D. I. (2017). ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J Appl Crystallogr*, *50*(Pt 4), 1212-1225. https://doi.org/10.1107/S1600576717007786

Fullmer, M. N. a. L. (2015). In: Trends in Polyoxometalates Research. In L. R. a. D. Schaming (Ed.), *In: Trends in Polyoxometalates Research* (pp. 20). Nova Science Publishers, Inc.

Garcia-Maurino, S. M., Rivero-Rodriguez, F., Velazquez-Cruz, A., Hernandez-Vellisca, M., Diaz-Quintana, A., De la Rosa, M. A., & Diaz-Moreno, I. (2017). RNA Binding Protein Regulation and Cross-Talk in the Control of AU-rich mRNA Fate. *Front Mol Biosci*, *4*, 71. https://doi.org/10.3389/fmolb.2017.00071

Gerstberger, S., Hafner, M., & Tuschl, T. (2014). A census of human RNA-binding proteins. *Nat Rev Genet*, *15*(12), 829-845. https://doi.org/10.1038/nrg3813

Godavarthi, J. D., Polk, S., Nunez, L., Shivachar, A., Glenn Griesinger, N. L., & Matin, A. (2020). Deficiency of Splicing Factor 1 (SF1) Reduces Intestinal Polyp Incidence in Apc(Min/)(+) Mice. *Biology (Basel)*, *9*(11). https://doi.org/10.3390/biology9110398

Gupta, A., Jenkins, J. L., & Kielkopf, C. L. (2011). RNA induces conformational changes in the SF1/U2AF65 splicing factor complex. *J Mol Biol*, *405*(5), 1128-1138. https://doi.org/10.1016/j.jmb.2010.11.054

Gupta, P. A., Wallis, D. D., Chin, T. O., Northrup, H., Tran-Fadulu, V. T., Towbin, J. A., & Milewicz, D. M. (2004). FBN2 mutation associated with manifestations of Marfan syndrome and congenital contractural arachnodactyly. *J Med Genet*, *41*(5), e56. https://doi.org/10.1136/jmg.2003.012880

Guth, S., & Valcarcel, J. (2000). Kinetic role for mammalian SF1/BBP in spliceosome assembly and function after polypyrimidine tract recognition by U2AF. *J Biol Chem*, *275*(48), 38059-38066. https://doi.org/10.1074/jbc.M001483200

Hackmann, A., Gross, T., Baierlein, C., & Krebber, H. (2011). The mRNA export factor Npl3 mediates the nuclear export of large ribosomal subunits. *EMBO Rep*, *12*(10), 1024-1031. https://doi.org/10.1038/embor.2011.155

Heintz, C., Doktor, T. K., Lanjuin, A., Escoubas, C., Zhang, Y., Weir, H. J., Dutta, S., Silva-Garcia, C. G., Bruun, G. H., Morantte, I., Hoxhaj, G., Manning, B. D., Andresen, B. S., & Mair, W. B. (2017). Splicing factor 1 modulates dietary restriction and TORC1 pathway longevity in C. elegans. *Nature*, *541*(7635), 102-106. https://doi.org/10.1038/nature20789

Hennig, J., Warner, L. R., Simon, B., Geerlof, A., Mackereth, C. D., & Sattler, M. (2015). Structural Analysis of Protein-RNA Complexes in Solution Using NMR Paramagnetic Relaxation Enhancements. *Methods Enzymol*, *558*, 333-362. https://doi.org/10.1016/bs.mie.2015.02.006

Hentze, M. W., Castello, A., Schwarzl, T., & Preiss, T. (2018). A brave new world of RNA-binding proteins. *Nat Rev Mol Cell Biol*, *19*(5), 327-341. https://doi.org/10.1038/nrm.2017.130

Hiroaki, H., & Kohda, D. (2018). Protein–Ligand Interactions Studied by NMR. In J. The Nuclear Magnetic Resonance Society of (Ed.), *Experimental Approaches of NMR Spectroscopy: Methodology and Application to Life Science and Materials Science* (pp. 579-600). Springer Singapore. https://doi.org/10.1007/978-981-10-5966-7_21

Holmes, R. K., Tuck, A. C., Zhu, C., Dunn-Davies, H. R., Kudla, G., Clauder-Munster, S., Granneman, S., Steinmetz, L. M., Guthrie, C., & Tollervey, D. (2015). Loss of the Yeast SR Protein Npl3 Alters Gene Expression Due to Transcription Readthrough. *PLoS Genet*, *11*(12), e1005735. https://doi.org/10.1371/journal.pgen.1005735

Huang, J.-r., Warner, L. R., Sanchez, C., Gabel, F., Madl, T., Mackereth, C. D., Sattler, M., & Blackledge, M. (2014). Transient Electrostatic Interactions Dominate the Conformational Equilibrium Sampled by Multidomain Splicing Factor U2AF65: A Combined NMR and SAXS Study. *Journal of the American Chemical Society*, *136*(19), 7068-7076. https://doi.org/10.1021/ja502030n

Ikura, M., Kay, L. E., & Bax, A. (1990). A novel approach for sequential assignment of 1H, 13C, and 15N spectra of proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. *Biochemistry*, *29*(19), 4659-4667. https://doi.org/10.1021/bi00471a022

Irimia, M., & Roy, S. W. (2008). Evolutionary convergence on highly-conserved 3' intron structures in intron-poor eukaryotes and insights into the ancestral eukaryotic genome. *PLoS Genet*, *4*(8), e1000148. https://doi.org/10.1371/journal.pgen.1000148

Ishizuka, A., Hasegawa, Y., Ishida, K., Yanaka, K., & Nakagawa, S. (2014). Formation of nuclear bodies by the lncRNA Gomafu-associating proteins Celf3 and SF1. *Genes Cells*, *19*(9), 704-721. https://doi.org/10.1111/gtc.12169

Iwahara, J., Schwieters, C. D., & Clore, G. M. (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*(18), 5879-5896. https://doi.org/10.1021/ja031580d

Jaremko, L., Jaremko, M., Ejchart, A., & Nowakowski, M. (2018). Fast evaluation of protein dynamics from deficient (15)N relaxation data. *J Biomol NMR*, *70*(4), 219-228. https://doi.org/10.1007/s10858-018-0176-3

Jeong, S. (2017). SR Proteins: Binders, Regulators, and Connectors of RNA. *Mol Cells*, *40*(1), 1-9. https://doi.org/10.14348/molcells.2017.2319

Kang, H.-S., Sánchez-Rico, C., Ebersberger, S., Sutandy, F. X. R., Busch, A., Welte, T., Stehle, R., Hipp, C., Schulz, L., Buchbender, A., Zarnack, K., König, J., & Sattler, M. (2020). An autoinhibitory intramolecular interaction proof-reads RNA recognition by the essential splicing factor U2AF2. *Proceedings of the National Academy of Sciences*, *117*(13), 7140-7149. https://doi.org/doi:10.1073/pnas.1913483117

Keil, P., Wulf, A., Kachariya, N., Reuscher, S., Huhn, K., Silbern, I., Altmuller, J., Keller, M., Stehle, R., Zarnack, K., Sattler, M., Urlaub, H., & Strasser, K. (2023). Npl3 functions in mRNP assembly by recruitment of mRNP components to the transcription site and their transfer onto the mRNA. *Nucleic Acids Res*, *51*(2), 831-851. https://doi.org/10.1093/nar/gkac1206

Kelaini, S., Chan, C., Cornelius, V. A., & Margariti, A. (2021). RNA-Binding Proteins Hold Key Roles in Function, Dysfunction, and Disease. *Biology (Basel)*, *10*(5). https://doi.org/10.3390/biology10050366

Kikhney, A. G., & Svergun, D. I. (2015). A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett*, *589*(19 Pt A), 2570-2577. https://doi.org/10.1016/j.febslet.2015.08.027

Klare, J. P., & Steinhoff, H.-J. (2009). Spin labeling EPR. *Photosynthesis Research*, *102*(2), 377-390. https://doi.org/10.1007/s11120-009-9490-7

Kramer, A. (1992). Purification of splicing factor SF1, a heat-stable protein that functions in the assembly of a presplicing complex. *Mol Cell Biol*, *12*(10), 4545-4552. https://doi.org/10.1128/mcb.12.10.4545-4552.1992

Kramer, S. (2021). Nuclear mRNA maturation and mRNA export control: from trypanosomes to opisthokonts. *Parasitology*, *148*(10), 1196-1218. https://doi.org/10.1017/S0031182021000068

Kress, T. L., Krogan, N. J., & Guthrie, C. (2008). A single SR-like protein, Npl3, promotes pre-mRNA splicing in budding yeast. *Mol Cell*, *32*(5), 727-734. https://doi.org/10.1016/j.molcel.2008.11.013

Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R., & Thornton, J. M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR*, *8*(4), 477-486. https://doi.org/10.1007/BF00228148

Lee, K. C., Chung, K. S., Lee, H. T., Park, J. H., Lee, J. H., & Kim, J. K. (2020). Role of Arabidopsis Splicing factor SF1 in Temperature-Responsive Alternative Splicing of FLM pre-mRNA. *Front Plant Sci*, *11*, 596354. https://doi.org/10.3389/fpls.2020.596354

Li, C., & Tarn, W.-Y. (2006). Splicing. In *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine* (pp. 1788-1792). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-29623-9_2880

Li, X., Liu, S., Zhang, L., Issaian, A., Hill, R. C., Espinosa, S., Shi, S., Cui, Y., Kappel, K., Das, R., Hansen, K. C., Zhou, Z. H., & Zhao, R. (2019). A unified mechanism for intron and exon definition and back-splicing. *Nature*, *573*(7774), 375-380. https://doi.org/10.1038/s41586-019-1523-6

Linder, B., Fischer, U., & Gehring, N. H. (2015). mRNA metabolism and neuronal disease. *FEBS Lett*, *589*(14), 1598-1606. https://doi.org/10.1016/j.febslet.2015.04.052

Lipp, J. J., Marvin, M. C., Shokat, K. M., & Guthrie, C. (2015). SR protein kinases promote splicing of nonconsensus introns. *Nature Structural & Molecular Biology*, *22*(8), 611-617.

https://doi.org/10.1038/nsmb.3057
Liu, Z., Luyten, I., Bottomley, M. J., Messias, A. C., Houngninou-Molango, S., Sprangers, R., Zanier, K., Kramer, A., & Sattler, M. (2001). Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science*, *294*(5544), 1098-1102. https://doi.org/10.1126/science.1064719
Loerch, S., & Kielkopf, C. L. (2016). Unmasking the U2AF homology motif family: a bona fide protein-protein interaction motif in disguise. *RNA*, *22*(12), 1795-1807. https://doi.org/10.1261/rna.057950.116
Long, J. C., & Caceres, J. F. (2009). The SR protein family of splicing factors: master regulators of gene expression. *Biochem J*, *417*(1), 15-27. https://doi.org/10.1042/BJ20081501
Long, Y., Sou, W. H., Yung, K. W. Y., Liu, H., Wan, S. W. C., Li, Q., Zeng, C., Law, C. O. K., Chan, G. H. C., Lau, T. C. K., & Ngo, J. C. K. (2019). Distinct mechanisms govern the phosphorylation of different SR protein splicing factors. *J Biol Chem*, *294*(4), 1312-1327. https://doi.org/10.1074/jbc.RA118.003392
Lukong, K. E., & Richard, S. (2004). Arginine methylation signals mRNA export. *Nat Struct Mol Biol*, *11*(10), 914-915. https://doi.org/10.1038/nsmb1004-914
Lunde, B. M., Moore, C., & Varani, G. (2007). RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol*, *8*(6), 479-490. https://doi.org/10.1038/nrm2178
Mackereth, C. D., Madl, T., Bonnal, S., Simon, B., Zanier, K., Gasch, A., Rybin, V., Valcárcel, J., & Sattler, M. (2011). Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. *Nature*, *475*(7356), 408-411. https://doi.org/10.1038/nature10171
Maguire, M. L., Guler-Gane, G., Nietlispach, D., Raine, A. R. C., Zorn, A. M., Standart, N., & Broadhurst, R. W. (2005). Solution Structure and Backbone Dynamics of the KH-QUA2 Region of the Xenopus STAR/GSG Quaking Protein. *Journal of Molecular Biology*, *348*(2), 265-279. https://doi.org/https://doi.org/10.1016/j.jmb.2005.02.058
Manalastas-Cantos, K., Konarev, P. V., Hajizadeh, N. R., Kikhney, A. G., Petoukhov, M. V., Molodenskiy, D. S., Panjkovich, A., Mertens, H. D. T., Gruzinov, A., Borges, C., Jeffries, C. M., Svergun, D. I., & Franke, D. (2021). ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. *J Appl Crystallogr*, *54*(Pt 1), 343-355. https://doi.org/10.1107/S1600576720013412
Manceau, V., Swenson, M., Le Caer, J.-P., Sobel, A., Kielkopf, C. L., & Maucuer, A. (2006). Major phosphorylation of SF1 on adjacent Ser-Pro motifs enhances interaction with U2AF65. *The FEBS Journal*, *273*(3), 577-587. https://doi.org/https://doi.org/10.1111/j.1742-4658.2005.05091.x
Masunaga, T., Niizeki, H., Yasuda, F., Yoshida, K., Amagai, M., & Ishiko, A. (2015). Splicing abnormality of integrin beta4 gene (ITGB4) due to nucleotide substitutions far from splice site underlies pyloric atresia-junctional epidermolysis bullosa syndrome. *J Dermatol Sci*, *78*(1), 61-66. https://doi.org/10.1016/j.jdermsci.2015.01.016
Mazroui, R., Puoti, A., & Kramer, A. (1999). Splicing factor SF1 from Drosophila and Caenorhabditis: presence of an N-terminal RS domain and requirement for viability. *RNA*, *5*(12), 1615-1631. https://doi.org/10.1017/s1355838299991872
McBride, A. E., Cook, J. T., Stemmler, E. A., Rutledge, K. L., McGrath, K. A., & Rubens, J. A. (2005). Arginine methylation of yeast mRNA-binding protein Npl3 directly affects its function, nuclear export, and intranuclear protein interactions. *J Biol Chem*, *280*(35), 30888-30898. https://doi.org/10.1074/jbc.M505831200
McKee, A. E., & Silver, P. A. (2007). Systems perspectives on mRNA processing. *Cell Res*, *17*(7), 581-590. https://doi.org/10.1038/cr.2007.54
Mercer, T. R., Clark, M. B., Andersen, S. B., Brunck, M. E., Haerty, W., Crawford, J., Taft, R. J., Nielsen, L. K., Dinger, M. E., & Mattick, J. S. (2015). Genome-wide discovery of human splicing branchpoints. *Genome Res*, *25*(2), 290-303. https://doi.org/10.1101/gr.182899.114

Mikheeva, S., Murray, H. L., Zhou, H., Turczyk, B. M., & Jarrell, K. A. (2000). Deletion of a conserved dinucleotide inhibits the second step of group II intron splicing. *RNA*, *6*(11), 1509-1515. https://doi.org/10.1017/s1355838200000972

Moehle, E. A., Ryan, C. J., Krogan, N. J., Kress, T. L., & Guthrie, C. (2012). The yeast SR-like protein Npl3 links chromatin modification to mRNA processing. *PLoS Genet*, *8*(11), e1003101. https://doi.org/10.1371/journal.pgen.1003101

Myung, J. K., & Sadar, M. D. (2012). Large scale phosphoproteome analysis of LNCaP human prostate cancer cells. *Mol Biosyst*, *8*(8), 2174-2182. https://doi.org/10.1039/c2mb25151e

Nameki, N., Takizawa, M., Suzuki, T., Tani, S., Kobayashi, N., Sakamoto, T., Muto, Y., & Kuwasako, K. (2022). Structural basis for the interaction between the first SURP domain of the SF3A1 subunit in U2 snRNP and the human splicing factor SF1. *Protein Sci*, *31*(10), e4437. https://doi.org/10.1002/pro.4437

Ni, J. Z., Grate, L., Donohue, J. P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T. A., Blume, J. E., & Ares, M., Jr. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev*, *21*(6), 708-718. https://doi.org/10.1101/gad.1525507

Olivieri, C., Subrahmanian, M. V., Xia, Y., Kim, J., Porcelli, F., & Veglia, G. (2018). Simultaneous detection of intra- and inter-molecular paramagnetic relaxation enhancements in protein complexes. *J Biomol NMR*, *70*(3), 133-140. https://doi.org/10.1007/s10858-018-0165-6

Ottoz, D. S. M., & Berchowitz, L. E. (2020). The role of disorder in RNA binding affinity and specificity. *Open Biol*, *10*(12), 200328. https://doi.org/10.1098/rsob.200328

Padgett, R. A. (2012). New connections between splicing and human disease. *Trends Genet*, *28*(4), 147-154. https://doi.org/10.1016/j.tig.2012.01.001

Palmer, A. G., 3rd. (2004). NMR characterization of the dynamics of biomacromolecules. *Chem Rev*, *104*(8), 3623-3640. https://doi.org/10.1021/cr030413t

Pastuszak, A. W., Joachimiak, M. P., Blanchette, M., Rio, D. C., Brenner, S. E., & Frankel, A. D. (2011). An SF1 affinity model to identify branch point sequences in human introns. *Nucleic Acids Res*, *39*(6), 2344-2356. https://doi.org/10.1093/nar/gkq1046

Pereira, B., Billaud, M., & Almeida, R. (2017). RNA-Binding Proteins in Cancer: Old Players and New Actors. *Trends Cancer*, *3*(7), 506-528. https://doi.org/10.1016/j.trecan.2017.05.003

Plass, M., Agirre, E., Reyes, D., Camara, F., & Eyras, E. (2008). Co-evolution of the branch site and SR proteins in eukaryotes. *Trends Genet*, *24*(12), 590-594. https://doi.org/10.1016/j.tig.2008.10.004

Qin, H., Ni, H., Liu, Y., Yuan, Y., Xi, T., Li, X., & Zheng, L. (2020). RNA-binding proteins in tumor progression. *J Hematol Oncol*, *13*(1), 90. https://doi.org/10.1186/s13045-020-00927-w

Query, C. C., Strobel, S. A., & Sharp, P. A. (1996). Three recognition events at the branch-site adenine. *The EMBO Journal*, *15*(6), 1392-1402. https://doi.org/https://doi.org/10.1002/j.1460-2075.1996.tb00481.x

Reddy, T., & Rainey, J. K. (2010). Interpretation of biomolecular NMR spin relaxation parameters. *Biochem Cell Biol*, *88*(2), 131-142. https://doi.org/10.1139/o09-152

Ren, P., Lu, L., Cai, S., Chen, J., Lin, W., & Han, F. (2021). Alternative Splicing: A New Cause and Potential Therapeutic Target in Autoimmune Disease. *Front Immunol*, *12*, 713540. https://doi.org/10.3389/fimmu.2021.713540

Rino, J., Desterro, J. M., Pacheco, T. R., Gadella, T. W., Jr., & Carmo-Fonseca, M. (2008). Splicing factors SF1 and U2AF associate in extraspliceosomal complexes. *Mol Cell Biol*, *28*(9), 3045-3057. https://doi.org/10.1128/MCB.02015-07

Romero-Barrios, N., Legascue, M. F., Benhamed, M., Ariel, F., & Crespi, M. (2018). Splicing regulation by long noncoding RNAs. *Nucleic Acids Res*, *46*(5), 2169-2184. https://doi.org/10.1093/nar/gky095

Rutz, B., & Seraphin, B. (2000). A dual role for BBP/ScSF1 in nuclear pre-mRNA retention and splicing.

EMBO J, *19*(8), 1873-1886. https://doi.org/10.1093/emboj/19.8.1873
Sagar, A., Jeffries, C. M., Petoukhov, M. V., Svergun, D. I., & Bernado, P. (2021). Comment on the Optimal Parameters to Derive Intrinsically Disordered Protein Conformational Ensembles from Small-Angle X-ray Scattering Data Using the Ensemble Optimization Method. *J Chem Theory Comput*, *17*(4), 2014-2021. https://doi.org/10.1021/acs.jctc.1c00014
Sandhu, R., Sinha, A., & Montpetit, B. (2021). The SR-protein Npl3 is an essential component of the meiotic splicing regulatory network in Saccharomyces cerevisiae. *Nucleic Acids Res*, *49*(5), 2552-2568. https://doi.org/10.1093/nar/gkab071
Sandhu, R., Sinha, A., & Montpetit, B. (2021). The SR-protein Npl3 is an essential component of the meiotic splicing regulatory network in Saccharomyces cerevisiae. *Nucleic Acids Research*, *49*(5), 2552-2568. https://doi.org/10.1093/nar/gkab071
Sattler, M. (1990). Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Progress in Nuclear Magnetic Resonance Spectroscop*, *34*, 66. https://doi.org/10.1016/S0079-6565(98)00025-9
Scott, D. D., Aguilar, L. C., Kramar, M., & Oeffinger, M. (2019). It's Not the Destination, It's the Journey: Heterogeneity in mRNA Export Mechanisms. In M. Oeffinger & D. Zenklusen (Eds.), *The Biology of mRNA: Structure and Function* (pp. 33-81). Springer International Publishing. https://doi.org/10.1007/978-3-030-31434-7_2
Selenko, P., Gregorovic, G., Sprangers, R., Stier, G., Rhani, Z., Kramer, A., & Sattler, M. (2003). Structural basis for the molecular recognition between human splicing factors U2AF65 and SF1/mBBP. *Mol Cell*, *11*(4), 965-976. https://doi.org/10.1016/s1097-2765(03)00115-1
Shen, Y., & Bax, A. (2013). Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *Journal of Biomolecular NMR*, *56*(3), 227-241. https://doi.org/10.1007/s10858-013-9741-y
Simon, B., Madl, T., Mackereth, C. D., Nilges, M., & Sattler, M. (2010). An efficient protocol for NMR-spectroscopy-based structure determination of protein complexes in solution. *Angew Chem Int Ed Engl*, *49*(11), 1967-1970. https://doi.org/10.1002/anie.200906147
Sjodt, M., & Clubb, R. T. (2017). Nitroxide Labeling of Proteins and the Determination of Paramagnetic Relaxation Derived Distance Restraints for NMR Studies. *Bio Protoc*, *7*(7). https://doi.org/10.21769/BioProtoc.2207
Skrisovska, L., & Allain, F. H. (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*(1), 151-164. https://doi.org/10.1016/j.jmb.2007.09.030
Sodium phosphate. (2006). *Cold Spring Harbor Protocols*, *2006*(1), pdb.rec8303. https://doi.org/10.1101/pdb.rec8303
Softley, C. A., Bostock, M. J., Popowicz, G. M., & Sattler, M. (2020). Paramagnetic NMR in drug discovery. *J Biomol NMR*, *74*(6-7), 287-309. https://doi.org/10.1007/s10858-020-00322-0
Stanley, R. F., & Abdel-Wahab, O. (2022). Dysregulation and therapeutic targeting of RNA splicing in cancer. *Nat Cancer*, *3*(5), 536-546. https://doi.org/10.1038/s43018-022-00384-z
Tanackovic, G., & Kramer, A. (2005). Human splicing factor SF3a, but not SF1, is essential for pre-mRNA splicing in vivo. *Mol Biol Cell*, *16*(3), 1366-1377. https://doi.org/10.1091/mbc.e04-11-1034
Tsuiji, H., Yoshimoto, R., Hasegawa, Y., Furuno, M., Yoshida, M., & Nakagawa, S. (2011). Competition between a noncoding exon and introns: Gomafu contains tandem UACUAAC repeats and associates with splicing factor-1. *Genes Cells*, *16*(5), 479-490. https://doi.org/10.1111/j.1365-2443.2011.01502.x
Urbanski, L. M., Leclair, N., & Anczukow, O. (2018). Alternative-splicing defects in cancer: Splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics. *Wiley Interdiscip Rev RNA*, *9*(4), e1476. https://doi.org/10.1002/wrna.1476

Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, M., Ulrich, E. L., Markley, J. L., Ionides, J., & Laue, E. D. (2005). The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins*, *59*(4), 687-696. https://doi.org/10.1002/prot.20449

Wagner, R. E., & Frye, M. (2021). Noncanonical functions of the serine-arginine-rich splicing factor (SR) family of proteins in development and disease. *Bioessays*, *43*(4), e2000242. https://doi.org/10.1002/bies.202000242

Wan, L., Deng, M., & Zhang, H. (2022). SR Splicing Factors Promote Cancer via Multiple Regulatory Mechanisms. *Genes (Basel)*, *13*(9). https://doi.org/10.3390/genes13091659

Wan, R., Bai, R., & Shi, Y. (2019). Molecular choreography of pre-mRNA splicing by the spliceosome. *Curr Opin Struct Biol*, *59*, 124-133. https://doi.org/10.1016/j.sbi.2019.07.010

Wang, W., Maucuer, A., Gupta, A., Manceau, V., Thickman, Karen R., Bauer, William J., Kennedy, Scott D., Wedekind, Joseph E., Green, Michael R., & Kielkopf, Clara L. (2013). Structure of Phosphorylated SF1 Bound to U2AF65 in an Essential Splicing Factor Complex. *Structure*, *21*(2), 197-208. https://doi.org/https://doi.org/10.1016/j.str.2012.10.020

Wang, X., Bruderer, S., Rafi, Z., Xue, J., Milburn, P. J., Krämer, A., & Robinson, P. J. (1999). Phosphorylation of splicing factor SF1 on Ser20 by cGMP-dependent protein kinase regulates spliceosome assembly. *The EMBO Journal*, *18*(16), 4549-4559. https://doi.org/https://doi.org/10.1093/emboj/18.16.4549

Waudby, C. A., Ramos, A., Cabrita, L. D., & Christodoulou, J. (2016). Two-Dimensional NMR Lineshape Analysis. *Sci Rep*, *6*, 24826. https://doi.org/10.1038/srep24826

Wilkinson, M. E., Charenton, C., & Nagai, K. (2020). RNA Splicing by the Spliceosome. *Annual Review of Biochemistry*, *89*(1), 359-388. https://doi.org/10.1146/annurev-biochem-091719-064225

Williamson, M. P. (2013). Using chemical shift perturbation to characterise ligand binding. *Prog Nucl Magn Reson Spectrosc*, *73*, 1-16. https://doi.org/10.1016/j.pnmrs.2013.02.001

Wright, C. J., Smith, C. W. J., & Jiggins, C. D. (2022). Alternative splicing as a source of phenotypic diversity. *Nat Rev Genet*, *23*(11), 697-710. https://doi.org/10.1038/s41576-022-00514-4

Xu, B., Meng, Y., & Jin, Y. (2021). RNA structures in alternative splicing and back-splicing. *Wiley Interdiscip Rev RNA*, *12*(1), e1626. https://doi.org/10.1002/wrna.1626

Zhang, P., Philippot, Q., Ren, W., Lei, W.-T., Li, J., Stenson, P. D., Palacín, P. S., Colobran, R., Boisson, B., Zhang, S.-Y., Puel, A., Pan-Hammarström, Q., Zhang, Q., Cooper, D. N., Abel, L., & Casanova, J.-L. (2022). Genome-wide detection of human variants that disrupt intronic branchpoints. *Proceedings of the National Academy of Sciences*, *119*(44), e2211194119. https://doi.org/doi:10.1073/pnas.2211194119

Zhang, Y., Dai, Y., Huang, Y., Wang, K., Lu, P., Xu, H., Xu, J. R., & Liu, H. (2020). The SR-protein FgSrp2 regulates vegetative growth, sexual reproduction and pre-mRNA processing by interacting with FgSrp1 in Fusarium graminearum. *Curr Genet*, *66*(3), 607-619. https://doi.org/10.1007/s00294-020-01054-2

Zhang, Y., Madl, T., Bagdiul, I., Kern, T., Kang, H.-S., Zou, P., Mäusbacher, N., Sieber, S. A., Krämer, A., & Sattler, M. (2012). Structure, phosphorylation and U2AF65 binding of the N-terminal domain of splicing factor 1 during 3′-splice site recognition. *Nucleic Acids Research*, *41*(2), 1343-1354. https://doi.org/10.1093/nar/gks1097