Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University Heidelberg

Presented by

M.Sc. Andrea Lomoschitz

born in: Sopron (Hungary)

Oral examination: 27.04.2023

Structural and RNA binding studies of Hrp48 – a regulator of translation in female *Drosophila* dosage compensation

Referees: Dr. Julia Mahamid

Prof. Dr. Rebecca Wade

Abstract

Post-transcriptional gene control is essential in gene expression, and one major step is the regulation of translation, controlling functions in a plethora of biological processes. A well-studied example of translation regulation is the repression of *msl-2* mRNA translation, a crucial step in the regulation of X-chromosome dosage compensation in females of *Drosophila melanogaster*. The repression is coordinated by the protein Sex-lethal (Sxl), which binds to both untranslated regions (UTRs) of the *msl-2* mRNA. At the 3' UTR Sxl recruits further RNA binding proteins, Unr and Hrp48 to form a ribonucleoprotein (RNP) complex which targets an early translation initiation step. Hrp48 directly interacts with the eIF3d subunit of the 43S preinitiation complex, but the detailed molecular role of Hrp48 during translational repression and its interaction with *msl-2* is not well understood.

Hrp48 consists of two N-terminal RNA recognition motifs (RRM) and in this work I solved the crystal structure of RRM1 at 1.2 Å resolution and validated the structure prediction model of RRM2 by NMR spectroscopy. In order to identify the RNA interaction site and binding affinities of Hrp48, I utilized NMR spectroscopy titrations as the differences in affinities were not resolved by isothermal titration calorimetry. The two RRM domains of Hrp48 bind the RNA simultaneously and synergistically forming a 1:1 complex in solution. Based on NMR relaxation and RDC experiments, the complex behaves very dynamically and the two RRMs remain flexible with respect to each other upon RNA-binding, suggesting a binding mode unusual for tandem-RRMs. The identified RNA-binding sites were corroborated by cellular assays performed by our collaborator. Studies directed to the understanding of the complex formation between Hrp48, Unr, Sxl and *msl-2* suggest no interaction of the proteins in the absence of RNA, however the three proteins bind *msl-2* simultaneously. I established a protocol to reproducibly form the quaternary complex of Sxl, Unr, Hrp48 and *msl-2*. It has been shown previously, that Sxl and Unr synergistically bind *msl-2*. My data shows that the incorporation of Hrp48 to the RNP complex occurs in a non-cooperative fashion.

Zusammenfassung

Post-transkriptionelle Genregulation spielt eine essenzielle Rolle in der Genexpression. Ein wichtiger Schritt in diesem Vorgang ist die Translationregulation, die eine Vielzahl biologischer Prozesse kontrolliert. Ein gut untersuchtes Beispiel für Translationregulation ist die Unterdrückung der Translation der *msl-2* mRNA, ein wichtiger Schritt in der Regulation der X-Chromosom-Dosiskompensation in weiblichen *Drosophila melanogaster*. Diese Regulation wird koordiniert durch das Protein Sex-lethal (Sxl), dass die beiden untranslatierten Regionen (UTR) der *msl-2* mRNA bindet. Sxl rekrutiert die RNA-bindenden Proteine Unr und Hrp48 an der 3‘-UTR und formt mit diesen einen Ribonucleoproteinkomplex (RNP), der einen frühen Schritt in der Initiation der Translation reguliert. Hrp48 interagiert direkt mit der elF3d Untereinheit des 43S Präinititationskomplex, aber seine Rolle in der Unterdrückung der Translation und die der Interaktion mit *msl-2* ist kaum verstanden.

Hrp48 besitzt zwei strukturierte RNA-Erkennungsmotive (engl. RNA recognition motif, RRM). In dieser Arbeit habe ich die Kristallstruktur von RRM1 mit 1.2 Å Auflösung gelöst und die vorhergesagte Struktur von RRM2 mit Hilfe von Kernspinresonanzspektroskopie (engl. nuclear magnetic resonance spectroscopy, NMR) validiert. Die Bindungsstelle und Affinität für RNA wurde durch NMR-Titrationen bestimmt, da isothermale Titrationskalometrie (engl. Isothermal titration calorimetry, ITC) keine Unterschiede in der Affinität auflösen konnte. Die beiden RRM-Domänen in HRP48 binden gleichzeitig und synergistisch an die RNA. Dies führt zur Bildung eines 1:1-Komplexes in Lösung. NMR-Relaxationsmessungen und dipolare Restkopplungen (engl. residual dipolar coupling, RDC) deuten darauf hin, dass der Komplex in Lösung dynamisch ist und die beiden RRMs in Lösung auch nach der Bindung an RNA flexible Orientierungen relativ zueinander zeigen. Die so bestimmten RNA-Bindestellen wurden mit unseren Kollaborationspartnern in Zellassays untersucht. Experimente, um den Mechanismus der Komplexformation zu untersuchen, zeigen keine Interaktion der Proteine in Abwesenheit der RNA, aber alle drei Proteine binden die RNA gleichzeitig. Ein Protokoll für die reproduzierbare Rekonstitution des quartären Komplexes aus Hrp48, Unr, Sxl und *msl*-2 wurde entwickelt. Frühere Arbeiten haben gezeigt, dass Sxl und Unr *msl-2* synergistisch binden, während meine Arbeit zeigt, dass die Integration von HRP48 in den RNP-Komplex nicht-kooperativ ist.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ATP	Adenosine triphosphate
CITE	Cap-independent translation enhancer
Cryo-EM	Cryo-electron microscopy
CSD	Cold-shock domain
CSP	Chemical shift perturbation
DCC	Dosage compensation complex
dsRNA	Double-stranded RNA
eIF	Eukaryotic initiation factor
EMSA	Electrophoretic mobility shift assay
grk	Gurken
GSC	Germline stem cell
HDE	Hsp83 degradation element
hnRNP	Heterogeneous nuclear ribonucleoprotein
How	Held-out-wings
Hrp48	(also: Hrb27C) Heterogeneous nuclear ribonucleoprotein 27C
HSQC	Heteronuclear single quantum coherence
IRES	Internal ribosome entry site
ITC	Isothermal titration calorimetry
lncRNA	Long non-coding RNA
Mle	Maleless
Mof	Male absent on first
MR	Molecular replacement
mRNA	Messenger RNA
Msl	Male-specific lethal
ncCSD	Non-canonical cold-shock domain
NMR	Nuclear magnetic resonance spectroscopy
NOE	Nuclear Overhauser effect
osk	Oskar
pAbP	Poly(A)-binding protein
PIC	Preinitiation complex
ppm	Parts per million
RBD	RNA binding domain
RBP	RNA binding protein
RDC	Residual dipolar coupling
RMSD	Root mean square deviation
RNP	Ribonucleoprotein
roX	RNA on X
RRM	RNA recognition motif
SAXS	Small angle X-ray scattering
SEC-MALLS	Size exclusion chromatography-multi-angle laser light scattering
snRNP	Small nuclear ribonucleoprotein
smFRET	Single-molecule fluorescence resonance energy transfer

ssRNA	Single stranded RNA
Sxl	Sex-lethal
tra	Transformer (gene)
Ubx	Ultrabithorax
Unr	Upstream of N-Ras
uORF	Upstream open reading frame
UTR	Untranslated region
WT	Wild type

1 Introduction

The formulation of the central dogma of molecular biology by Francis Crick in 1958 [1] laid the foundation for novel biological fields to emerge. Comprehending the various aspects of gene expression includes the research of transcriptional and translational regulation. These are essential steps in expressing a gene product and they are studied immensely. Yet, our atomic-level mechanistic understanding on these processes remains to be broadened.

An excellent model system for studying gene expression regulation is the dosage compensation of *Drosophila melanogaster.* This system allows to study both transcription control and epigenetics on the male side [2] and post-transcriptional regulation mediated by RNA-binding proteins on the female side [3].

1.1 Dosage compensation

Expanding genetic diversity and adaptive advantages lead to the evolution of different mechanisms of genetic recombination, such as sexual reproduction found in most animals and plants [4]. This is often genetically determined and involves differentiated sex chromosomes [5], while other species rely on such environmental ques, as temperature [6]. Sex-determination and chromosome systems vary heavily even across vertebrates [7]. For example, most mammals are male XY heterogametic, which means they have different sex chromosomes, while females are XX homogametic. On the contrary, most reptiles and birds are female heterogametic with ZW females and ZZ males. Both Y and W chromosomes are smaller in size and carry only a few active genes, yet despite the apparent similarity, these systems evolved independently multiple times and from different autosomes [8]. However, evolution of sex chromosomes poses an aneuploidy to the heterogametic sex, which is balanced by dosage compensation. There are two aspects of this imbalance: one is the ratio of X or Z-linked gene expression to the autosomal gene expression, and the other one is the X or Z-linked expression of genes across the sexes.

In my thesis work I focus on the latter one: the inequality in the numbers of X (or Z) chromosomes across the biological sexes, and the dosage compensation of this. These mechanisms operate on the transcriptional and translational levels to reach equal expression levels of X-linked genes. Deficiency or lack of this process leads to unequal expression levels of proteins coded by the X chromosome across sexes, which reduces fitness and leads to lethality. Different species evolved diverse

mechanisms for dosage compensation, however there are three ways in common eukaryotic model organisms, which are well-studied (Figure 1).

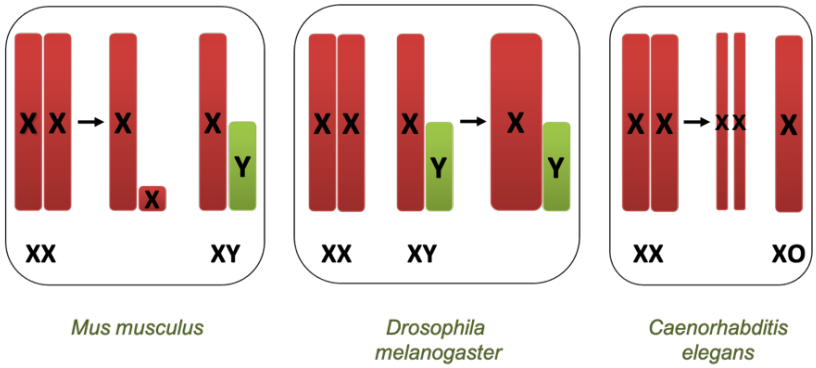


Figure 1 The three most studied examples of dosage compensation.

In each cell of placental mammalian females, one of the X chromosomes gets randomly inactivated (XCI, X chromosome inactivation, Figure 1 left panel). In other monotreme mammals, birds, and reptiles, the dosage compensation is restricted and partial [8]. Middle panel, in *D. melanogaster* males, the only X chromosome is hyper-transcribed about twofold [9]. Right panel, in hermaphroditic *C. elegans,* the expression from both X chromosomes is repressed by half [3].

1.2 *Drosophila* dosage compensation

In *D. melanogaster,* sex-chromosome dosage compensation on the transcriptional level takes place in males, where the only X chromosome is about twofold upregulated by the control of the male-specific dosage-compensation complex. In females, a counterbalancing mechanism occurs, which is regulated post-transcriptionally and mediated by the master sex-determination protein, sex-lethal (Sxl).

1.2.1 Drosophila genes and chromosomes

The use of the *D. melanogaster* species of Arthropoda as a model organism in biology originated from Charles W. Woodworth's laboratory at the very beginning of the 20th century [10] and has been used since in various fields such as genetics, developmental biology, and physiology. There are several reasons for its popularity as a model system, including a short life cycle; requiring little space, care and equipment; together with high fecundity, morphological aspects and numerous genetic reasons [11]. *D. melanogaster* cells are diploid, comprising of four pairs of chromosomes.

The chromosome pair number 1 are the sex chromosomes, females have two X chromosomes and males have an X and a Y chromosome. There are three pairs of autosomes: the 2 and 3 are the two

larger, and 4 is the smallest of the chromosomes (Figure 2A) [12]. The genome of *D. melanogaster* was first sequenced in 2000 [13-14] and today the known genome consists of around 143 million base pairs standing for 13968 coding and 4044 non-coding genes [15].

1.2.2 Sex determination system and sex-lethal

Sex-determination in *Drosophila melanogaster* takes place based on the X chromosome number. The number of autosomes sets the ratio (X:A), unlike in humans, where the presence or absence of the Y chromosome is the sex-determining switch [16]. The *Drosophila* Y chromosome is important for male fertility, but not for sex determination. The XY and XO chromosomes determine sex for males, and XO is infertile, while both XX and XXY determine for fertile females [17].

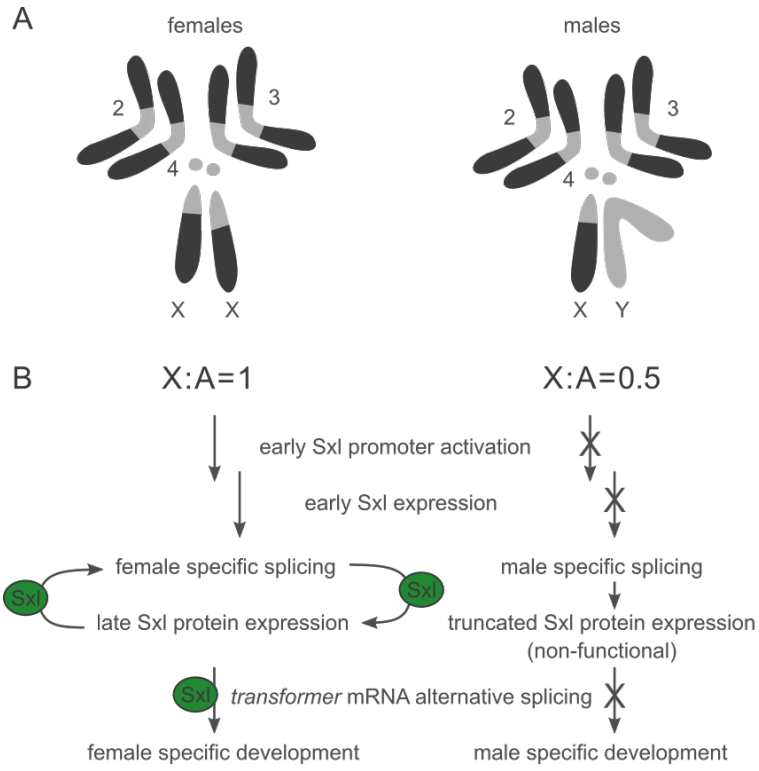


Figure 2 Karyotypes and sex-determination in *D. melanogaster*

A: *Drosophila* diploid chromosome systems consist of three pairs of autosomes and two sex chromosomes: XX in females and XY in males. **B:** Sex-determination and sexual differentiation is regulated based on the X:A signal. The early expression of Sxl in females results in the establishment of a feedback-loop for Sxl self-splicing and downstream regulation of female development. In the absence of Sxl, the genes responsible for female development remain silent and male development takes place.

The expression of the master sex-determination gene, *sex-lethal* (*Sxl*) becomes activated in early embryogenesis in females of *Drosophila*, because the X:A ratio is 1 (2X and 2 sets of autosomes,

since it is diploid), and not in males, where the ratio is 0.5 [18]. The molecular 'counting' is encoded in X-linked (numerator in X:A) and A-linked (denominator) transcription factors. During the early developmental blastoderm stage, products of the numerator genes activate the early/establishment promoter of *Sxl*, while denominator transcription factors compete with the numerators [19]. As in males the effect of the denominator transcription factors is predominant, no or significantly less Sxl is expressed via the early promoter compared to females. After the blastoderm stage, a late promoter becomes activated and consequently *Sxl* begins to be transcribed in both sexes. The early Sxl present from the blastoderm stage contributes to the female specific alternative splicing of the *Sxl* pre-mRNA, which also sets up an autoregulatory loop for female specific splicing and the maintenance of its own protein levels [20-21]. In males, a different splicing pattern takes place and the transcripts are spliced to a non-functional isoform since they lack early Sxl [22-23].

The functionally spliced Sxl in females then regulates a plethora of post-transcriptional events [24], some of which were already mentioned in the previous chapter. Sxl binds tightly [25] and controls the female specific alternative splicing of *transformer* (*tra*) pre-mRNA [18]. Tra proteins then control downstream female specific splicing of transcription factors accounting for somatic sexual differentiation and morphological development (Figure 2) [26].

1.2.3 The translation of the one X chromosome of males is upregulated

In *Drosophila melanogaster*, the single male X chromosome is twofold upregulated transcriptionally, in order to match the expression levels of the two X chromosomes of females [3]. This hyper-transcription is mediated by the chromatin-modifying multi-protein dosage compensation complex (DCC), also termed the male-specific lethal (Msl) complex [2]. This complex consists of two long noncoding RNAs (lncRNAs), *roX1* and *roX2* (RNA on X) encoded by the X chromosome, and five proteins (Msl-1, Msl-2, Msl-3, Mof; 'males-absent on the first' and Mle; 'maleless') of which Msl-2 is a limiting key component for the assembly, stabilization, and function of the DCC [2, 27]. Msl-1 acts as a scaffold for the complex, by interacting with Msl-2, Msl-3 and Mof [28]. The *roX* lncRNAs are necessary for selective X chromosomal targeting, and the double-stranded RNA helicase Mle remodels the *roX* lncRNA in an ATP dependent fashion, and this remodeling enhances the association of Msl-2 to the complex [29-30].The Msl proteins and the *roX* lncRNAs are recruited specifically to the X chromosome, where the DCC complex binds to specific regions, the high affinity binding sites on the X chromosome, and spreads to low affinity sites to cover the whole chromosome [31-32]. There are more hypotheses explaining the X-chromosome upregulation [33-34], which are complementary to each other. The canonical hypothesis is linked to the histone acetyltransferase Mof, which acetylates lysine 16 on histone tail 4 (H4K16ac) [35-36], a distinct feature of the X-chromosome compared to autosomes.

This modification is thought to promote transcription and to loosen the highly packed structure of chromatin, facilitating the attachment of chromatin-binding factors [37-38]. A direct link between enhanced recruitment of Pol II to the X-chromosome and the Msl proteins and further studies revealed increased elongation by Pol II [2]. An additional mechanism suggests a role for Msl-1 in Pol II Ser5 phosphorylation promoting transcription initiation [2, 39].

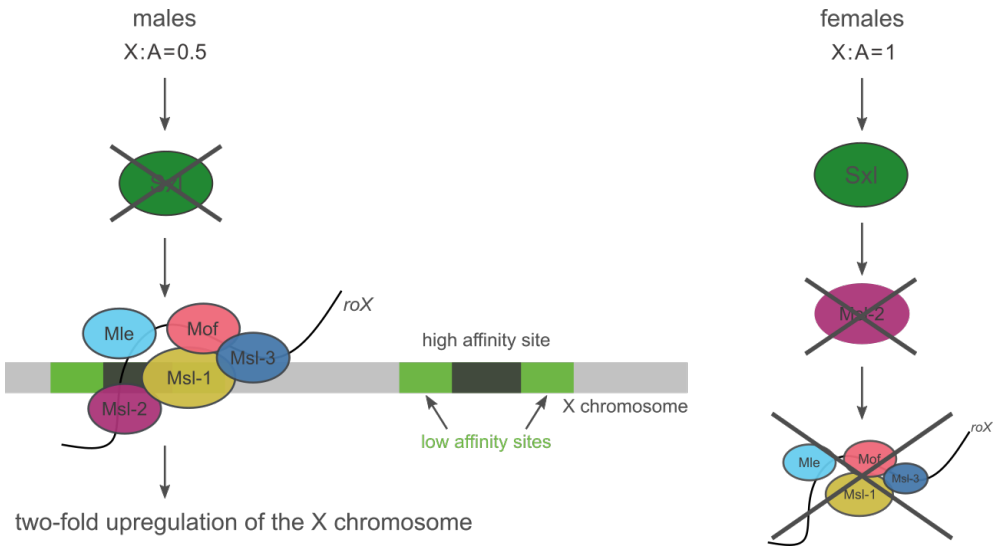


Figure 3 Dosage compensation by the DCC complex

In males (left), the dosage compensation is achieved by the chromatin-remodeling DCC complex accomplishing a twofold hyper-transcription of the X chromosome. In females (right), the DCC does not assemble and function, as the limiting component Msl-2 is not expressed.

1.2.4 Dosage compensation in females is repressed

In females of *D. melanogaster,* the assembly and the action of the DCC needs to be inhibited by a set of post-transcriptional events [3, 40] to ensure survival [27] as dosage compensation in females would lead to ectopic upregulation of X-linked genes and abnormal protein levels (Figure 3). In female Msl-2, the limiting component of the DCC complex is not expressed [27, 41-42], *roX1* is expressed until the first 2 hours after egg laying and then the transcripts gradually fade during embryogenesis [43], and *roX2* is undetectable [44-45]. Interestingly, in females, the other four Msl proteins (Msl-1, Msl-3, Mle and Mof) are still expressed, however, Msl-1 exists at lower levels compared to in males [46]. Induced expression of Msl-2 in females initiates higher expression levels of Msl-1, as well as the ectopic assembly of the Msl proteins on the female X chromosomes and loosening of the chromatin structure [27]. These transgenic female flies showed a delay in female-specific development, compromised

fertility, and significantly decreased viability. Curiously, females still produced the *msl-2* transcript, although about 60% of the quantity in males [27, 47]. Therefore, a repression mechanism evolved also on the post-transcriptional level.

Repression of *msl-2* translation is orchestrated by female-specific isoforms of Sex-lethal, by multiple mechanisms on different gene expression steps, and its literature has been reviewed extensively (Figure 4) [3, 24].

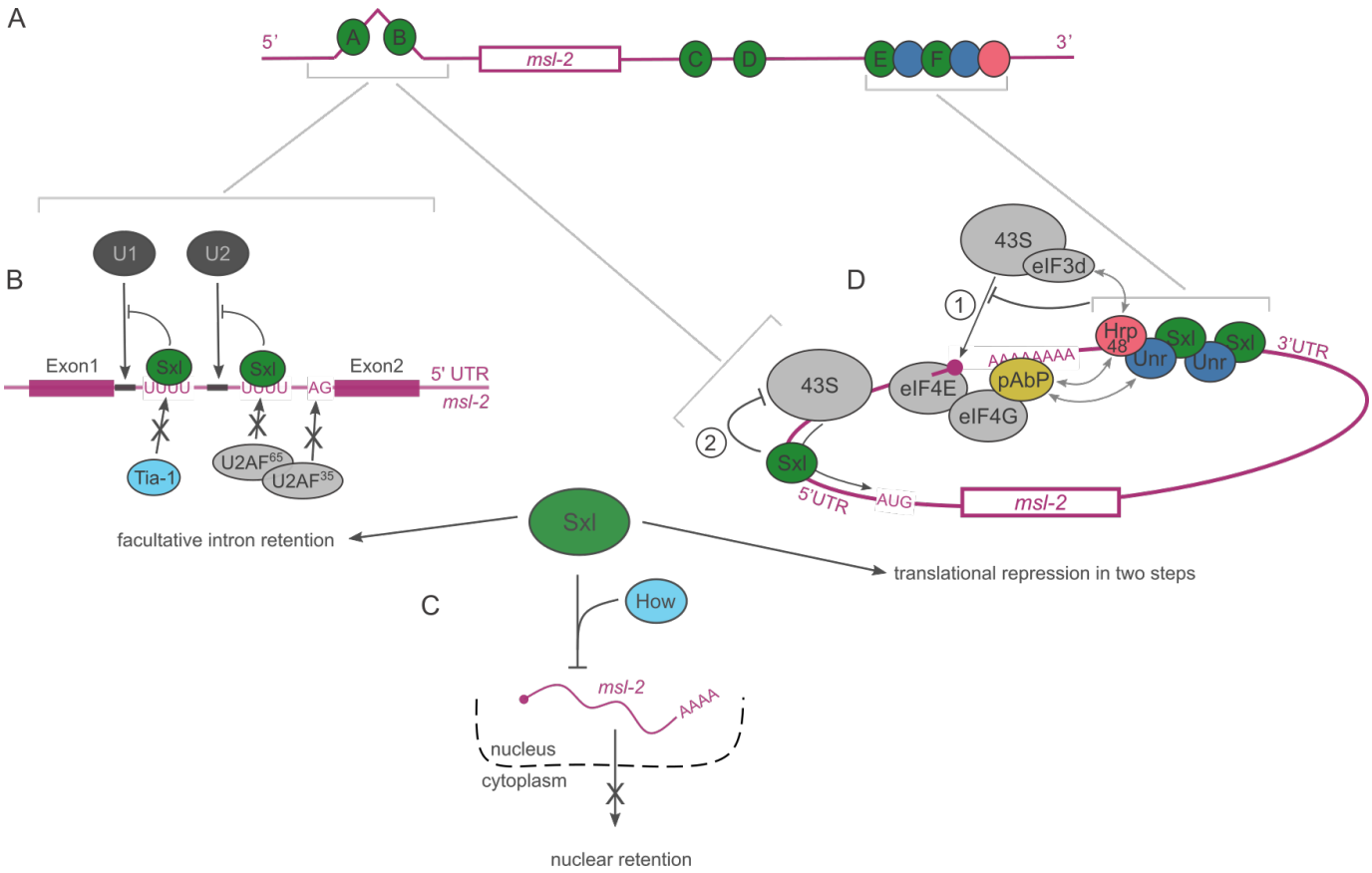


Figure 4 Sxl is the master regulator of Msl-2 expression inhibition

Sxl mediates three post-transcriptional events to prevent the expression of Msl-2 in females. **A**: schematic representation of the *msl-2* mRNA with the Sxl binding sites A – F (green), Unr binding sites (blue) and Hrp48 binding site (pink). The respective sites for the post-transcriptional control events are connected on the figure. **B**: Sxl regulates the female-specific splicing of *msl-2* by retention of a 5' UTR facultative intron, thereby retaining its own binding sites for downstream regulatory steps. **C**: Sxl along with How promotes the nuclear retention of the *msl-2* transcripts. **D**: Sxl regulates the translational repression of *msl-2* in two distinct steps. At the 3' UTR along with other cofactors it inhibits the recruitment of the 43S ribosomal subunit to the mRNA ①. Those 43S units that escaped from this control are inhibited at the AUG scanning step by Sxl bound to the 5' UTR ②.

In a first step, Sxl regulates the alternative splicing of a 5' UTR female-specific facultative intron of *msl*-2-pre-mRNA at both 3' and 5' splice sites (ss) (Figure 4A) [41-42, 48]. At the 3' ss, Sxl interferes

with the recruitment of the U2AF (Figure 4B) [49]. U2AF, the U2 auxiliary factor, is not part of the U2 snRNP, but it is required for supporting U2 snRNP binding to the splice site and it consists of the $U2AF^{35}$ 35 kDa and the $U2AF^{65}$ 65 kDa subunits. At the 3' ss, the minimum sequence necessary for the retention of the facultative intron is the combination of the polypyrimidine poly(Y)-tract (U_{16}) together with the long downstream sequence until the AG signal at the 3' end of the intron (Figure 4B) [48]. The distance between the poly(Y) and AG is essential for the splicing inhibition [49]. The $U2AF^{35}$, which is in complex with the $U2AF^{65}$ subunit, recognizes and binds the AG motif with high affinity. Subsequently, $U2AF^{65}$ recognizes the poly(Y)-tract, but if the distance between the AG and poly(Y) is large, as in the case of *msl-2*, the binding of $U2AF^{65}$ can be impaired by Sxl competing for the same poly(Y) sequence (Figure 4B) [49]. Sxl also binds at a high affinity binding site close to the 5' ss of the facultative intron and inhibits the recruitment of U1 snRNP to the 5' ss [50]. Tia-1 (Cytotoxic granule associated RNA binding protein or T-cell-restricted intracellular antigen-1), an important factor for splicing and U1 snRNP recruitment, competes with Sxl at the 5' splice site, which can compromise splicing of the facultative intron (Figure 4B) [49]. The female-specific intron retention is crucial for downstream translation control steps.

Second, Sxl also supports the nuclear retention of *msl-2* transcripts (Figure 4C), thereby reducing the available *msl-2* mRNA level in the cytosol [40]. Along with the RNA-binding cofactor Held-Out-Wings (How), Sxl forms a complex at the 5'UTR of *msl-2*. How interacts with the mRNA at two binding sites, about 30 nucleotides upstream from the Sxl binding site within the female-specific intron. RNA binding of How alone does not account for the nuclear retention. The presence of Sxl and both of the How binding sites are required for optimal nuclear retention [40].

Third, Sxl controls the translational repression of the *msl-2* mRNA in two distinct steps targeting two stages in early translation initiation [47-48, 51-52]. For this reason, before going into details on the *msl-2* translational repression, I introduce a short summary on translational initiation.

1.2.4.1 Translation initiation in eukaryotes

Post-transcriptional gene control is essential in gene expression. One major step is the regulation of translation that controls functions in a plethora of biological processes, possibly in all eukaryotic cell types. These can be divided in two main categories: global control, which acts on the bulk of the transcriptome and is mediated by alteration of translation-initiation factor activities, and mRNA-specific control, which selectively targets individual or a subset of mRNAs. RNA binding proteins and microRNAs play a prominent role in the latter [53-54]. RNA secondary structure features, specific sequence motifs, and nucleotide modifications, act as cis-regulatory elements on the transcripts and guide RNA binding proteins and initiation factors to their interaction sites. Due to their high target

specificity, RBPs can fine-tune protein expression levels and achieve tight control of translation [54]. There are three main steps of translation: initiation, elongation, and termination. Translation initiation can happen in both cap-dependent and cap-independent fashion.

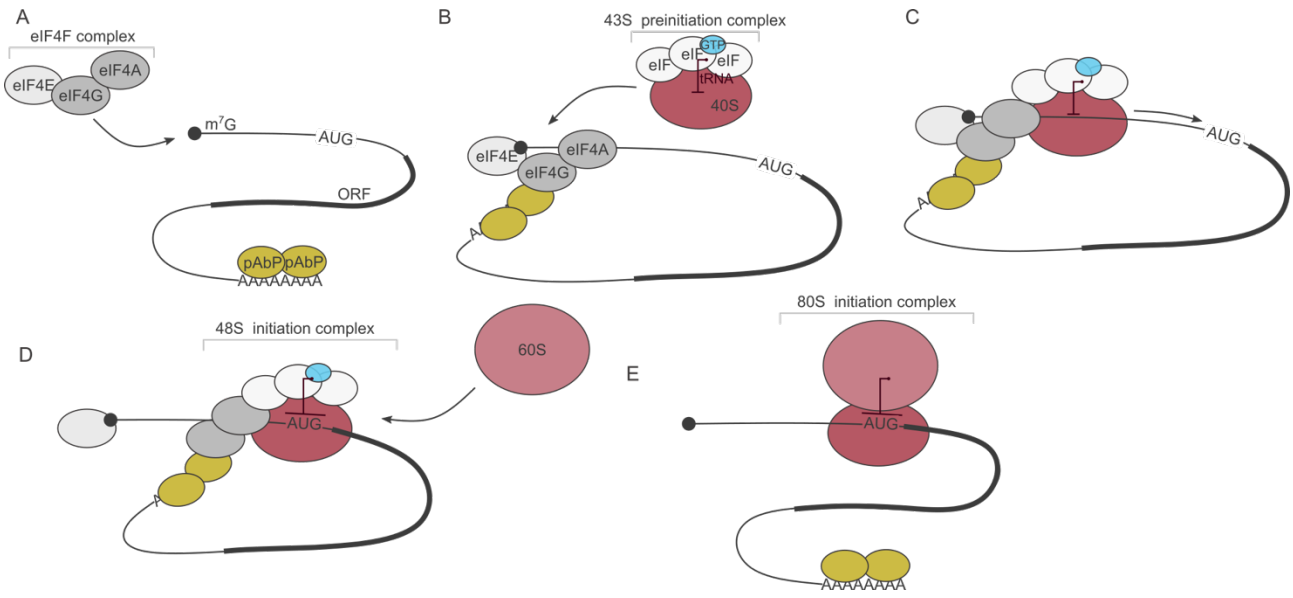


Figure 5 Main steps of cap-dependent eukaryotic translation initiation

A: The eIF4F complex is recruited to the 5' m^7G cap structure of the mRNA. At the 3' end, pAbP covers the poly-A tail. **B:** Interaction between the eIF4G subunit of the eIF4F complex and pAbP promotes the closed loop formation of the mRNA. The 43S preinitiation complex recognizes and binds at the cap-structure in an eIF4E-mediated mechanism. **C:** The PIC scans for the AUG codon in 5' to 3' direction. **D:** At the initiation codon, the 60S subunit merges with the initiation complex. **E:** the formation of the 80S is the last step before the elongation starts.

Cap-dependent translational initiation generally starts with the binding of the eukaryotic initiation factor 4F (eIF4F) complex to the m^7GpppN (7-methylguanylate) cap-structure at the 5' end of the mRNA in an ATP-dependent mechanism (Figure 5A). The eIF4F complex consists of three proteins: the cap-binding protein eIF4E; the ATP-dependent DEAD-box helicase eIF4A; and the scaffold protein eIF4G that connects with eIF4E and eIF4A [53, 55]. eIF4G also directly interacts with the poly(A)-binding protein (pAbP-1), and pAbP-1 is also bound to the poly-A tail at the 3' UTR [56], thereby forming a loop of the mRNA (Figure 5B). Keeping the two UTR regions in close proximity supports the recycling of the terminated ribosomes for a new translation cycle [57]. The eIF4F complex recruits the pre-assembled 43S preinitiation complex (PIC), that includes the 40S small ribosomal subunit, other eIF eukaryotic initiation factors, GTP, and the initiator tRNA (Figure 5B). Subsequently, the PIC scans the mRNA in the 5' to 3' direction for the downstream initiation codon (Figure 5C). After the start codon recognition, the formation of the 48S initiation complex occurs

(Figure 5D), which then merges with the 60S large ribosomal subunit to form the 80S ribosome, while some of the eIFs get released (Figure 5E), which then starts the peptide synthesis in the elongation phase [54].

A cap-independent mechanism has also evolved for translation initiation, in circumstances when the cap-dependent pathway is not available, such as during mitosis, under stress conditions, or oftentimes in viruses [58-59]. A well-studied, but not the only mechanism, is mediated by internal ribosome entry sites (IRES) at the 5'UTR, which host the direct recruitment of the 43S PIC. IRES are diverse in sequence, but they have similar secondary structure elements. In some cases, they require IRES trans-acting elements or initiation factors for their function as well [59]. Different to IRES are the cap-independent translation enhancers (CITEs), secondary structure elements residing at both 3' and 5' UTRs where they bind initiation factors or ribosomal subunits, mainly in viral translation systems [58, 60].

1.2.5 Translational repression of the *msl-2* mRNA

Sxl controls the translational repression of the *msl-2* mRNA by interacting with both UTRs of *msl-2* and targeting to steps of translation initiation (Figure 4D) [47-48, 51-52]. Initiation is frequently the rate-limiting step of translation, and many translational regulators focus on this to control protein expression [53].

Sxl recognizes and binds uridine-rich motifs [61] at both 3' and 5' UTRs of *msl-2* mRNA [27, 47, 62-63]. There are two Sxl binding sites (A and B) at the 5' UTR region and four at the 3' UTR (C-F) (Figure 4 A) [47]. The RNA binding and the translational repression activity of Sxl are coupled [63]. The binding sites at the 5' UTR overlap with the retained facultative intron, supposing the function of splicing inhibition is the maintenance of the Sxl-binding sites (Figure 4A and B). Out of the six binding sites, B, E, and F are necessary for optimal translational repression [64]; consequently, Sxl interaction is required on both untranslated regions [47, 64] (Figure 4A).

Bound to the 3' UTR, Sxl mediates the inhibition of the translational initiation step of 43S preinitiation complex (PIC) recruitment (Figure 5 B) [51, 65]. The mechanism involves a multiprotein complex formation at the E and F Sxl binding sites. Sxl bound to these sites recruits Unr (Upstream-of-N-ras) as a co-factor (Figure 4 D) [63]. The RNA-binding protein Unr binds sequences directly downstream of Sxl E and F binding motifs (Figure 4 A) [66-67]. Sxl and Unr do not interact with each other in absence of RNA, but specific protein-protein contacts are established upon RNA binding [68]. The binding of Sxl and Unr to *msl-2* is highly cooperative and enhances the affinities of both proteins [68]; thereby, Sxl grants sex specificity to Unr in *msl-2* recognition [66-67].

Sxl or Unr alone are not sufficient for translational repression at the 3' UTR, and both of them are necessary for this process [66]. Unr bound at the 3' UTR also directly binds and forms a complex with pAbP (poly(A)-binding protein) [69] and this interaction further promotes translational repression activity with a yet unidentified mechanism [70]. Interestingly, the Unr – pAbP interaction does not have an inhibitory effect on the closed loop formation of the mRNA through pAbP and the eukaryotic initiation factors eiF4E and eiF4G [70]; however, the closed loop formation (Figure 5 B) has been shown to primarily promote translation [57, 71].

Besides Sxl and Unr, Hrp48 was also found to directly interact with the downstream Region 5 of *msl-2* on the 3' UTR (Figure 4 D); and its binding is also necessary for optimal translation repression activity [72]. Differently to the synergistic binding of Sxl and Unr, Hrp48 does not require either of these factors to bind *msl-2*, yet it interacts with them in the absence and presence of RNA based on immunoprecipitation experimental data [72]. In addition, Hrp48 interacts with How, which binds at the 5' UTR and promotes nuclear retention of the *msl-2* transcript [40]. Mass-spectrometry and immunoprecipitation studies uncovered further factors interacting with Hrp48 in the context of *msl-2*, and the d subunit of the eIF3 initiation factor was identified as a direct interaction partner of Hrp48 independent of RNA, Sxl, and Unr [72]. eIF3 organizes different steps in translation initiation [73-74]. In the case of *msl-2*, the eIF3d binds to the 5' UTR and it is required for both *msl-2* translation and 3'UTR mediated translational repression. The binding can occur in a 5' Cap dependent or independent mechanism, but for both cases eIF3d can be targeted by the 3'UTR repressor complex (Figure 4 D) [72]. Recent results of Daniel Leopoldus, former Master student in our lab, suggest an additional interaction: the direct binding between Hrp48 and pAbP based on NMR spectroscopy data [75]. These interactions hint the formation of an hnRNP complex involving the association on the 3' UTR of *msl-2*, which might be in parts a dynamic assembly, where not all interactions occur simultaneously.

Those ribosomal subunits that escaped from the 3' UTR mediated control are inhibited by Sxl when bound at the 5' UTR intron [48, 51]. Here, Sxl regulates a later step of translation initiation by blocking the 43S PIC scanning for the AUG translational initiation codon (Figure 4D and Figure 5 C) [51]. The 5' UTR also encloses a short upstream open reading frame (uORFs) in between the two Sxl binding sites A and B [76]. uORFs are abundant on mRNAs of eukaryotes, and they are often responsible for reduced protein expression. The *msl-2* uORF is in a weak initiation context, which means the ribosomal scanning is leaky. Still, in the absence of Sxl, it provides an about 2-fold repression of translation of the downstream ORF (*msl-2*) compared to no uORF sequence present. However, in the presence of both the uORF and Sxl, translation gets about 14-fold repressed. Interestingly, and different to previously reported uORF mediated translational repression examples, the regulatory

activity does not depend on the coding sequence of the uORF or the presence of a termination codon [76]. In males, that do not express the Sxl isoform required for translation repression, hindrance of alternative splicing of the 5' intron does not affect Msl-2 expression [47], but Sxl transfection rescued the translation repression activity, indicating a prominent role of Sxl in translational repression [48].

Both translational repression mechanisms are necessary for female fly viability as they act synergistically [47, 62]; however, the mechanistic description of this cooperativity remains elusive. My thesis work targets questions on the structural and biophysical level of the translation repression mechanism.

Msl-2 is regulated by a plethora of RNA-binding proteins that form multi-subunit RNP complexes at the 3' UTRs and in the next section I describe these.

1.2.5.1 RNA binding proteins

RNAs are typically accompanied by proteins throughout their lifetime. RNA-binding proteins (RBPs) form ribonucleoprotein (RNP) complexes on RNA as soon as they are transcribed and control different aspects of RNA-metabolism, production, processing, nucleocytoplasmic export, localization, translation, and degradation of the RNA [77-78].

The conventional RBPs interact with RNA through distinct and well-conserved RNA-binding domains (RBDs) [79]. Some of the most prominent and well-studied RBDs are the RNA-recognition motif (RRM, details in The RNA recognition motif section) [80-81], the hnRNP K-homology domain (KH) binding both single-stranded (ss) RNA and DNA [82], the double-stranded RNA-binding domain (dsRBD) recognizing RNA helical structures rather than specific sequences [83], the cold-shock domain (CSD) that binds ssRNA purine-rich motifs, and the zinc-finger domain (ZF) that have been described as DNA binding proteins but also interact with RNA. RNA-binding proteins function in a myriad of biological processes, which could imply a similarly high number of structures for RNA recognition. On that note, RNA binding also can occur in the absence of these domains enormously extending the RNA interactome [84].

Yet, RBPs are constructed from a rather limited pool of modular domains, and the combination of these domains in various arrangement results in a diversity of function of the RBPs [85]. These domains individually bind often to short stretches of RNA with moderate affinity and specificity, but jointly, the binding interface can be extended to bring about elevated binding affinity and specificity. The domains can also deviate from the canonical topology, i.e., featuring additional secondary-structure elements or specific mutations modifying the RNA-binding activity. In addition, not just the order

and variation of the modules, but also the presence or absence of linker regions between the domains is important for defining target specificity [86]. Shorter linkers usually account for limited conformational freedom and smaller number of potential RNA targets but tighter binding, while longer and flexible linkers can allow the domains to bind less spatially restricted sequences, further apart on the RNA strand or on completely different RNAs. More restricted conformations can also drive structure specificity, or restrict the plastic RNA to function- specific conformations [85]. Limited conformational space can also occur upon interaction of the domains to one another fixing them into specific relative arrangements.

1.2.5.2 The RNA recognition motif

Particularly interesting to my thesis work is the RRM, a highly abundant single stranded RNA-binding protein domain, also called as ribonucleoprotein domain (RNP-1) or RNA-binding domain (RBD).

RRM domains were first reported in 1988 in the context of heterogeneous nuclear ribonucleoprotein (hnRNP) particles, which are complexes of nuclear proteins and newly transcribed RNA in the nucleus [87]. Through sequencing the cDNAs of different RNP proteins (first yeast pAbP [88-89] and human hnRNP A1 [90]) and aligning the sequences, they discovered an about 90 amino acid long sequence repeated four times in pAbP and two times in hnRNP A1. Further cDNA analysis resulted in discovery of the two RRM consensus sequences: RNP-1 [88], a highly conserved 8 amino acid long segment (R/K-G-**F/Y**-G/A-**F/Y**-I/L/V-X-**F/Y**) approximately in the middle of the domain, then RNP-2, a shorter and less well conserved sequence (I/L/V-**F/Y**-I/L/V-X-N-L) closer to the N-terminus (Figure 6 A and B) [80]. These RNP-motifs are conserved across different proteins and species and usually play a central role in RNA binding [81, 87]. The canonical RRM domain is arranged into a $\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$ topology, where the four β-strands fold into a sheet in a $\beta_4\beta_1\beta_3\beta_2$ order facing the two helices (Figure 6 B and C). The loops connecting the strands and helices are named as the following: loop 1 between β_1 and α_1 loop 2 between β_2 and β_3 and so on (Figure 6 A and C). RNP-1 is located in the β_3 strand and RNP-2 in the β_1 and along with the other strands, forming an RNA-binding platform in the middle of the β -sheet. In the usual binding mode, two aromatic residues in RNP-1 (Nr. 3 and 5 in Figure 6 C) and in RNP-2 (Nr. 2) interact with two nucleotides specifically. Nr. 2 from RNP-2 and 3 from RNP-1 then form stacking interactions with the nucleotides, while Nr. 5 from RNP-1 might interact with the RNA sugar-phosphate backbone [81]. A single RRM domain usually binds 3 – 4 nucleotides but can accommodate up to eight nucleotides when additional structural features extend the RNA-binding interface [91]. Another way to increase specificity and/or affinity is

to employ additional RRM domains to extend the surface for interaction, as in the case of pAbP [92] or Sex-lethal [25, 68].

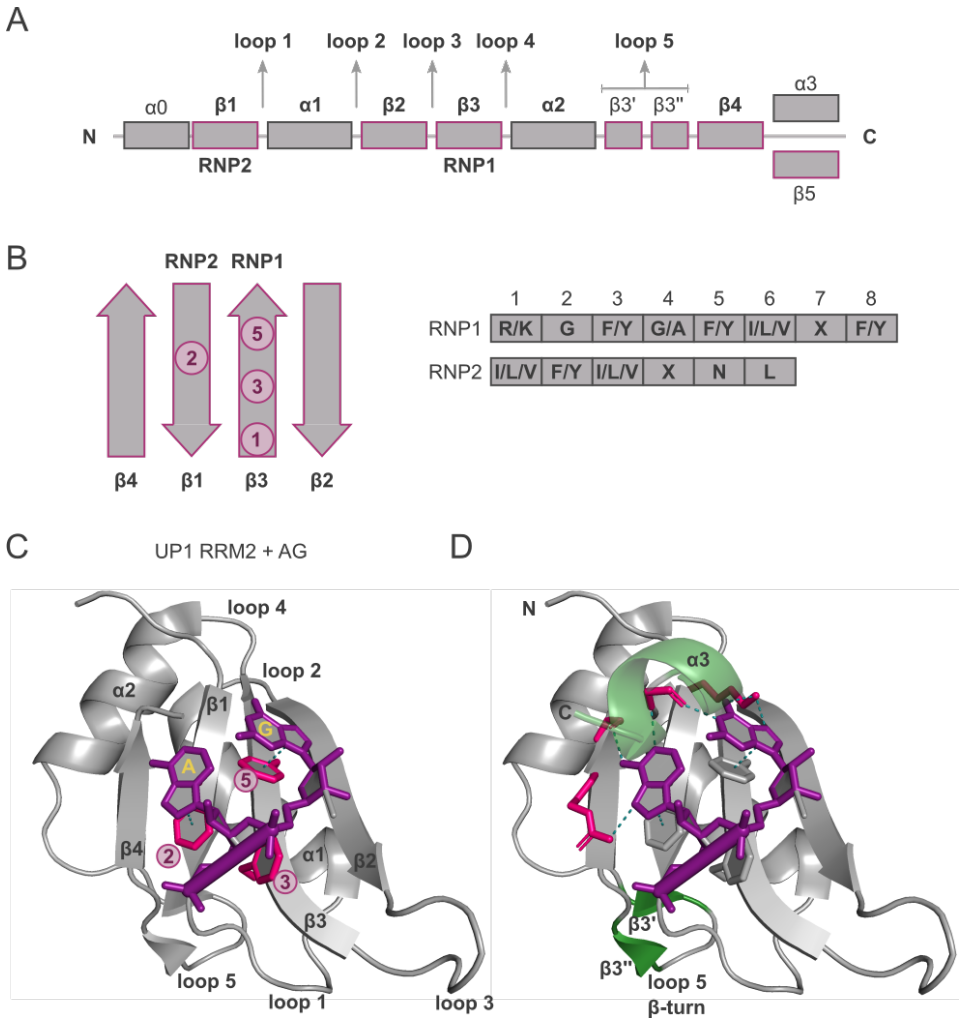


Figure 6 The RNA recognition motif

A: Schematic representation of the secondary structure elements in a canonical (bold fonts) and non-canonical (regular fonts) structural elements, with the placement of the RNP-motifs. **B**: Consensus sequences and arrangement of the conserved RNP-amino acids on the β-sheet. **C**: Model of a canonical RRM-RNA interaction exemplified by UP1 RRM2 domain (grey, interacting residues pink) bound to pri-mir-18a RNA (purple) [93]. For illustrative purposes only two nucleotides, AG from the whole RNA are shown. (Adapted from Afroz et al. [94]) **D**: Interaction of the pri-mir-18a with residues outside of the RNP-motifs and non-canonical secondary structural elements (green). In panel C, the helix 3 is not shown for better visibility.

There are numerous examples in RRM structure variation, accounting for an expansion in function of this otherwise small domain [80]. Regions additional to the β-sheet can contribute to RNA recognition to increase binding affinity and specificity, such as loops, α-helices, non-canonical helices and strands, N- and C-terminal extensions and linkers. Loop regions, most commonly 1, 3, or 5 can also be involved in RNA-binding either alone, or more loops simultaneously [81, 95]. For example, loop

1 in 25% of human RRMs contains an aromatic residue [81], and in case of Fox-1 F126 [96] or Hrp-1 W168 [97], structural data supports their role in RNA binding via direct H-bonding with the RNA. Fox-1 Loop 5 is also important for the RNA binding, and similarly to other cases, loop 5 tends to fold into a β-turn (Figure 6 D). Additional non-canonical β-strands can extend the available surface for RNA binding and consequently accommodating more nucleotides, which in addition can result in tighter binding. The alternative splicing factor PTB consists of four tandem RRM domains, of which two contains an additional β_5 strand [98]; RRM2 can bind one and RRM3 two additional nucleotides [99]. The N-terminal extension of the TIA-1 RRM3 domain forms and additional helix situated in the linker region between RRM2 and RRM3 [100]. This non-canonical helix strongly contributes to the RNA recognition based on NMR chemical shift perturbation data in a tandem construct [100].

Apart from RNA recognition, RRM domains have been shown to interact with DNA (hnRNP A1 and telomeric DNA) [101] and proteins. RRMs can interact intramolecularly with adjacent RRM domains (hnRNP A1, Sex-lethal) [25, 102], or intermolecularly with other proteins such as the Mago – Y14 complex, where the Y14 RNA-binding interface is completely buried upon interaction with Mago [103]. However, the protein recognition of RRM domains cannot be generalized as the RNA recognition via the β-sheet. Non-canonical structural elements can also contribute for protein-protein interactions, such as loop 5 β-turns in eIF3b – eIF3j peptide interaction [104], additional, elongated α-helices in $U2AF^{35}$ – $U2AF^{65}$ binding interface or other non-canonical β-strands. RRM domains can also undergo self-association and form dimers or higher-order oligomers. The RRM3 domain of HuR promotes the dimerization of the protein, thereby fine-tuning its RNA-binding affinity [105]. In the case of CPEB4, both the RRM1 domain and the tandem RRM12 construct can dimerize, but this does not have an effect the RNA-binding [106].

Proteins enclosing RRM domains have diverse biological functions, among others they are involved in all post-transcriptional regulation events, for example pre-mRNA processing, splicing and alternative splicing, transport, localization, and translation regulation [80]. This functional versatility is a result of their adaptability in their interaction properties.

1.2.6 RNA-binding proteins of the 3' UTR mediated msl-2 translational repression

1.2.6.1 Sex-lethal

Sxl is alternatively spliced in females and males, and the isoform needed for translational repression is female-specific [22]. It consists of an N-terminal GN-rich region and two canonical RRM domains that participate in RNA binding synergistically [107-109] and are connected by a short flexible linker (Figure 7 B). While the two RRM domains (residues 123 – 294, dRBD3, *Drosophila* RNA binding

domain 3) are sufficient for the recognition and binding of *msl-2*, a longer sequence (aa 122 – 301, dRBD4) is necessary for full translation repression activity (Figure 7 B) [63]. Interestingly, an even longer construct (aa 94 – 322, RBD) appeared to be less efficient in reporter assays, than dRBD4, although it was not directly compared, but to the full-length Sxl construct [62]. Sxl specifically and with high affinity interacts with uridine-rich motifs, not just in context of *msl-2* translation repression, but also in the case of alternative splicing of its own and the *transformer* pre-mRNA [25, 61, 108-109]. There are two structures published of dRBD3 in complex with RNA, one in complex with *tra* mRNA (UGUUUUUUU) and one in complex with *msl-2* 18-mer and Unr (Figure 8 A) [25, 68]. Interestingly, the two structures overlap very well (backbone coordinate RMSD = 0.52Å) and recognize RNA in a similar way. This suggests a canonical binding mechanism for Sxl-dRBD3.

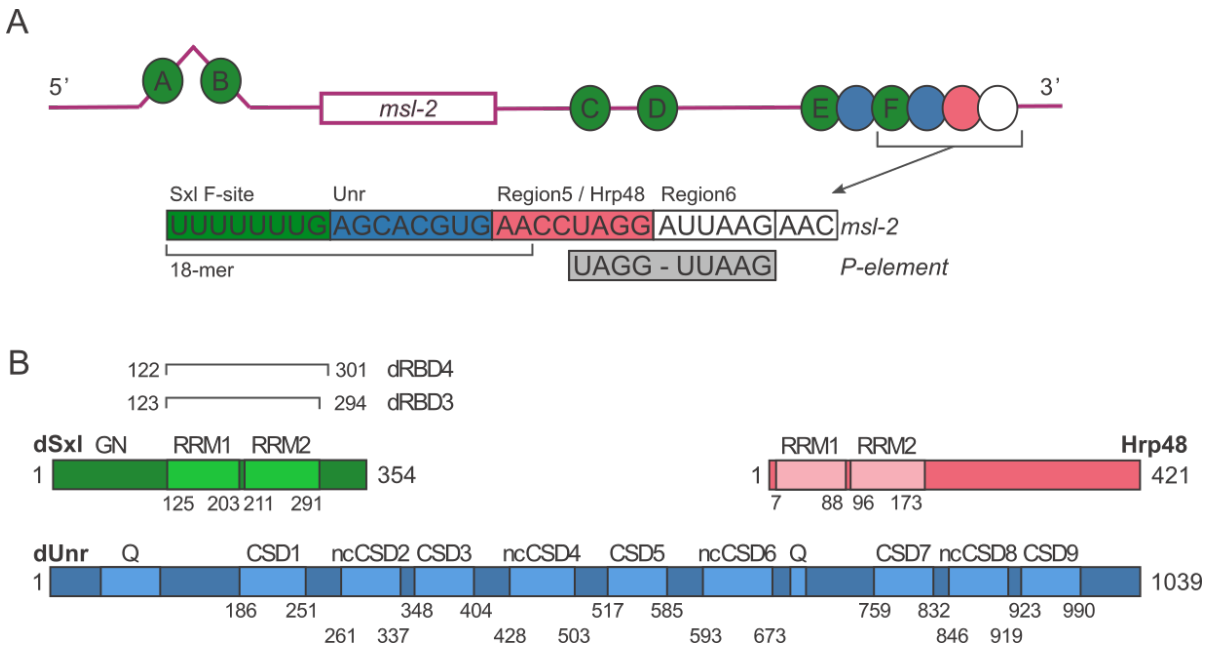


Figure 7 Elements of the *msl-2* translation repression

A: Schematic representation of the *msl-2* mRNA with the binding sites for the RNA binding factors (Sxl green, Unr blue, Hrp48 pink, downstream region white). The highlighted section depicts the region important for my thesis work. The grey sequence depicts the *P element* UTR binding site of Hrp48, that is highly similar to the *msl-2* sequence. **B**: Domain arrangement of Sxl, Hrp48 and Unr.

1.2.6.2 Upstream-of-N-ras

Unr is an abundant RBP found at comparable levels in male and female cells of *Drosophila*, and has a sex-specific role in the dosage compensation in both sexes [110-111]. In males, it possibly promotes the assembly of the DCC complex [110]; in females, it is part of the translational repression complex, and binds *msl-2* mRNA only jointly with Sxl in females [68]. Curiously, Unr is known to interact as translational activator; thus, Sxl conveys a rather unexpected function to this essential cofactor Unr

in a sex-specific fashion [66-67]. Unr proteins consist of five RNA-binding cold-shock domains (CSD) and two glutamine-rich domains (Q) [112-113]. Cold-shock domains are about 70 amino acid long single-stranded RNA binding motifs. Their topology follows a β-barrel in which five anti-parallel β-strands back against each other (Figure 8 B). Four aromatic residues protruding from the β-barrel are mainly responsible for the RNA-binding activity (**F**-G-**F** and **F**/**Y**-F-**H** motifs, Figure 8 B). Surprisingly, it was shown recently that *D. melanogaster* Unr features four additional non-canonical cold-shock domains (ncCSD) in-between the five predicted canonical ones [114] (Figure 7 B). These domains follow a highly similar fold to canonical CSDs, but lack the aromatic residues. Isolated ncCSD domains of Unr do not bind RNA per se, but in constructs together with canonical CSDs RNA – ncCSD contacts can form. For example in the Unr-CSD789 – poly-A complex, ncCSD8 directly binds the RNA [69]. In addition, ncCSD domains can contribute to the overall conformation of Unr, which helps to define RNA tertiary structure specificity [114].

The CSD1 (aa 182 – 252) is sufficient for joint high-affinity binding with *msl-2* and Sxl [68, 112], however for translational repression the first 397 amino acids are needed, a construct enclosing the N-terminal Q-rich domain and the first three CSD domains (CSD1, ncCSD2, CSD3). Interestingly, the Q-rich domain in connection with the CSD domains significantly augmented the repression activity. Altogether, the data suggested that residues in the Q-rich domain and between CSD1 and CSD3 (ncCSD2 and linkers, specific for *D. melanogaster*) are substantial for repression [112]. The structures of all CSD and ncCSD domains, except of CSD3 has been solved experimentally [114] and also the CSD789 in complex with an RRM domain of pAbP (Figure 8 C-F) [69].

Sxl and Unr form a complex on the *msl-2* mRNA EF binding sites and downstream sequences (Figure 8H and I) of which dRBD3 and CSD1 binding has been structurally characterized by X-ray crystallography and small-angle X-ray scattering [68]. A highly specific and synergistic recognition between the three parts of the complex was discovered. In this ternary interaction, both Unr-CSD1 and Sxl-dRBD3 bind *mls-2* with low nanomolar affinity. These affinities are 100-fold stronger for dRBD3 and 1000-fold stronger for CSD1, than the binding to the same RNA alone. The ternary interaction is stabilized through unique, interlaced, triple-zipper-like interactions between aromatic entities of all three components (Figure 8 I) [68].

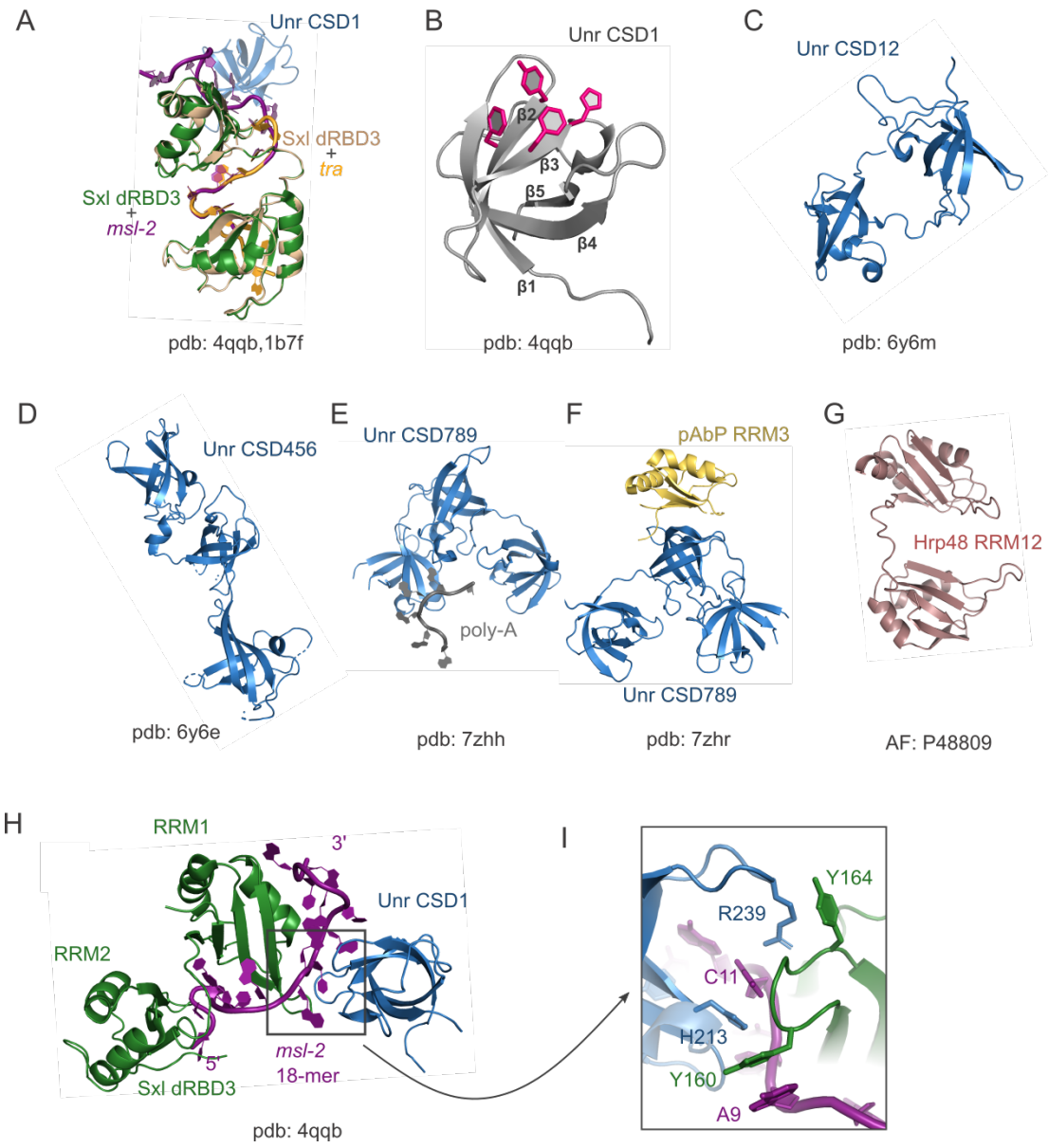


Figure 8 Structural knowledge on the proteins of the 3' UTR control of *msl-2* translational repression.

Existing structures of different subdomains of Sxl and Unr and their complexes with proteins and RNAs and the AlphaFold2 prediction of the Hrp48 RRM domains. **A:** Structures of Sxl-dRBD3 bound to *tra* and *msl-2* mRNA are overlaid (pdbs: 1b7f [25] and 4qqb respectively). **B:** Canonical CSD-fold on the example of Unr-CSD1 of the 4qqb crystal structure [68]. Canonical aromatic RNA-binding residues highlighted in pink. **C-F:** Structures of various Unr CSD and ncCSD-domains [69, 114]. C: Unr-CSD12, pdb: 6y6m D: Unr-CSD456, pdb: 6y6e E: Unr-CSD789 bound to poly-A RNA, pdb:7zhh F: Unr-CSD789 bound to pAbP-RRM3, pdb:7zhr **G:** AlphaFold2 prediction of Hrp48-RRM12. **H:** Crystal structure of the ternary complex of *msl-2* – Sxl-dRBD3 – Unr-CSD1, pdb: 4qqb [68] **I:** Specific triple-zipper-like interactions involving Unr-CSD1, Sxl-dRBD3 and *mls-2* in 4qqb.

1.2.6.3 Hrp48

Hrp48 or Hrb27C (Heterogeneous nuclear ribonucleoprotein 27C) is a member of the heterogeneous nuclear ribonucleoprotein A/B (hnRNP A/B) family of RNA-binding proteins. The hnRNPs are

proteins that associate with hnRNAs – bulks of unprocessed transcripts – to form hnRNP complexes that host mRNA processing steps. Hrp48 was first isolated in 1992 from *D. melanogaster* embryos by ssDNA affinity chromatography [115], and characterized along with other major nuclear proteins of hnRNP complexes [116]. Here, they termed the different proteins based on their apparent molecular weight; for Hrp48, it was measured to be around 48kDa. The name of the gene, Hrb27C comes from cytological localization studies in the same article, as the Hrp48 hybridization signal came from the 27C position of the left arm of the polytene chromosome [116].

Hrp48 is a highly abundant and essential protein in *D. melanogaster.* Its expression starts as early as the first stages of oogenesis, being present in germline and follicle cells [117]. Hrp48 mutations are lethal already at the larval stage [118], and studies of different Hrp48 mutant alleles showed failed development at early stages of oogenesis [117]. Localization of Hrp48 in the cells is dependent on the developmental stage. In early oogenesis, it already localizes with distinct patterns in the germarium [119]; in the follicle and nurse cells, it is mainly cytoplasmic and situated around the nucleus. In the oocyte, it localizes at the posterior half, and gradually concentrates at later stages to the posterior pole of the oocyte [117]. Hrp48 is also expressed in the somatic cells of the ovary and was also detected in Kc cells, the pupa and embryos, thus present in germline and somatic cells and in most cases primarily in the cytoplasm [120].

Sequencing data evaluation shows that Hrp48 consist of two N-terminal RRMs connected by a short (7aa long) linker and a low complexity C-terminal glycine-serine-rich region (Figure 7 B) [116]. The latter is responsible for liquid-to solid phase separation and RNP granule formation [121], and could account for protein-protein and protein-RNA interactions too. This is consistent with the two RRM domains in the AlphaFold predicted structure of Hrp48 (Figure 8 G) [122]. The putative binding site of Hrp48 on the *msl-2*, Region 5 (AACCUAGG, Figure 7 A) has been determined via reporter assays and immunoprecipitation experiments [72]. However, in the case of the P-element transposase mRNA [123], the binding sequence of Hrp48 (UAGGUUAAG, Figure 7 A) is almost identical to the second half of Region 5 and the downstream Region 6 of *msl-2* (Figure 7 A), and similar binding sequences to the *P element transposase* UTR has been determined in the case of the Hrp48 *oskar* mRNA recognition, though the actual sequences were not published [117]. Based on this, the *msl-2* binding site could be located four nucleotides downstream of Region 5. As well, the fact that Hrp48 was still pulled down with Region 5 could be explained by tight binding of Hrp48 even to only a part of the optimal RNA motif combined with protein-protein interactions with Sxl or Unr that were also present in the immunoprecipitation experiments. In addition, in earlier reporter assay studies with wild-type *msl-2* UTRs, the significance of Region 6 in translational repression was recognised [65], indicating

that Region 6, which overlaps in sequence with Hrp48 *P element* binding site, is highly involved in translational repression.

As previously mentioned, it has been shown that Hrp48 binds the 3'UTR of *msl-2* at Region 5, downstream of the Sxl and Unr binding sites. How Hrp48 separately and together with Sxl and Unr interacts with the *msl-2* is unknown and it is the central question of my thesis.

1.2.7 Hrp48 and its diverse roles in post-transcriptional regulation

Apart from its role in the 3' UTR- mediated translational repression, Hrp48 has been reported to be involved in numerous post-transcriptional gene control events. In most of these cases Hrp48 functions in concert with other RNA binding factors interacting with sequences at the UTRs of (pre-)mRNAs. On Figure 9 I summarized the various roles I describe below.

1.2.7.1 Splicing and alternative splicing

Hrp48-mediates the splicing of the pre-mRNA of the P transposable element, a 2.9 kb mobile DNA element in animals. Its transcript undergoes alternative splicing based on the tissue type. In germ cells, all three introns are spliced, resulting in the expression of the 87 kDa transposase enzyme, while in somatic and germ cells, the third intron (IVS3) remains unspliced and translates a 66 kDa repressor protein, that represses the transposition [124]. Additional alternative splicing can also occur in the P element. The U1 snRNP (small nuclear ribonucleoprotein particle) interacts with the 5' splice site of IVS3 as an initial step of spliceosome assembly. A multi-protein complex binding to an inhibitory element in the exon upstream of IVS3 inhibits the U1 snRNP binding to the 5' splice site and instead associates to a pseudo-5' splice site, deterring splicing of IVS3 [118, 123]. This inhibitory complex contains soma-specific PSI (P-element somatic inhibitor) [120] and Hrp48, which recognizes a sequence (F2 element, UAGGUUAAG) in the inhibitory element [123, 125]. In a later genome-wide study focused on method development, Hrp48 along with PSI, dASF/SF2, and B52/RSp55 was tested in microarray experiments as splicing regulators [126]. They tested the whole set of annotated alternatively spliced transcripts in *D. melanogaster* with knockdown cells of the splicing regulators. They found that Hrp48 was up- or down-regulated in 90 distinct splicing or alternative splicing events. Hrp48 was also found to be up- or down-regulated significantly and consistently together with PSI in seven splicing events, suggesting their cooperativity in further examples than the P element regulation. In addition, PSI and Hrp48 seemed to regulate many common splicing events, of which

the PSI targets were almost all influenced in the same direction as Hrp48 knockdowns, implying PSI entails Hrp48 for splicing. Hrp48 however has targets not similarly affected by PSI, proposing Hrp48 can regulate splicing independently from PSI [126].

Alternative splicing in higher eukaryotes enables the expression of different proteins with diverse functionality derived from the same locus, often based on the cell/tissue type or the developmental stage [127]. There are six alternative mRNAs produced of the *Ultrabithorax* (Ubx) homeotic gene of *D. melanogaster*, differentially spliced based on developmental stage or tissue/cell types. The complexity of isoform expression patterns is especially high in the central nervous system, where even individual neurons can have different expression ratios of the Ubx isoforms [128]. The *Ubx* transcription unit consists of 3 splice sites that connect 3 regions, which can be alternatively spliced: the B element and microexons I and II. Hrp48 has been found to exhibit a crucial role in the inclusion of the microexons in both larvae and adults [129].

1.2.7.2 RNA localization

Spatial and temporal control of gene expression of cytoplasmic mRNA is achieved through mRNA localization [130]. Cells can localize RNA for different reasons, for example promoting co-translational assembly of complexes and supramolecular structures, or achieving efficient localizing and concentration of the coded protein. During *D. melanogaster* oogenesis, establishment of antero-posterior and dorsal-ventral axes through pattering of protein expression is accomplished by asymmetric localization of the maternal mRNA [131]. Dorsal-ventral patterning requires *gurken (grk)* mRNA and protein localization, the anterior-posterior patterning requires *oskar (osk)* and *nanos* localization and translational control. Hrp48 is involved in both *osk* and *grk* localization.

Oskar activity is restricted to the posterior pole of the oocyte in developing embryos. The mRNA is translationally repressed during transport orchestrated by a repressor element in the 3'UTR and translation becomes de-repressed once localized to the posterior pole by a mechanism involving both UTRs [132]. Hrp48 binds to both 3' and 5' UTR regions of *osk* and co-localizes with other repressors such as Bruno, forming an mRNA complex and together repressing *osk* translation during localization [117, 133]. The proteins along with *osk* undergo liquid-solid phase separation which is needed for its localization. The C-terminal low complexity region of Hrp48 and N-terminal region of Bruno was found to be responsible for this phase separation [121]. In the study by *Yano et. al.,* the Hrp48 binding sites on *osk* were identified bearing homology to the F2 Hrp48 binding site (UAGGUUAAG) on the P element transposase [117, 132]. In the localization of *grk*, specific isoforms of the RNA-binding protein

Squid (Sqd) functions together with Hrp48, Cup, pAbp [134] and other factors to enable ordinary dorsal-ventral concentration of *grk* to the anterior pole [135]. Hrp48 interacts with Sqd, Cup, and pAbp in an RNA-dependent manner; and they physically bind the *grk* mRNA at the 3'UTR [134-136]. Additional factors mediating *grk* localization and directly interacting with Hrp48 in an RNA-independent manner have also been identified, such as Ovarian tumor (Otu) [135] and Glorund (Glo) [137]. Numerous proteins involved in *grk* spatial organization also function in *oskar* localization implying similarities in the organization of their control [138].

1.2.7.3 Translation efficiency

Modulating the efficacy of translation contributes to countless cellular processes and is often orchestrated by trans-acting factors binding to autoregulatory mRNA sequences at both 5' and 3'UTRs. Hrp48 acts as a bifunctional modulator. Examples show its involvement in both translational enhancement and translational repression processes.

One example for the translational upregulation is the translation of the heat shock protein 83 coding mRNA (*Hsp83*). The Hsp83 degradation element (HDE) at the 3' UTR of *Hsp83* is bound by a multiprotein complex, of which three proteins have been identified: Hrp48, the *Drosophila* dodeca-satellite protein 1 (DDP1), and pAbp [139]. These proteins form a single RNP complex on *Hsp83*, and each of them directly and simultaneously interact with the HDE and enhance *Hsp83* expression [139].

Another example, which is studied in this thesis, is the translational downregulation of *msl-2* mRNA mediated by several RNA binding proteins. Here, Sxl, Unr, pAbp, and Hrp48 act together as translation repressors (already discussed in the *Drosophila* dosage compensation chapter) [72].

Hrp48 mediates translational downregulation of Sxl. Studies directed to understand the joint effect of Hrp48 and Sxl on the Notch pathway revealed that Hrp48 represses Sxl expression [140]. In this work, they analysed on which level Hrp48 acts on the Sxl transcript, as Hrp48 has various functions in post-transcriptional gene control. They excluded alternative splicing, as the sex-specific transcript levels of Sxl remained the same, and Sxl regulates its own alternative splicing in a sex-specific manner. RNA stability and modification at the poly-A tail were also eliminated, leaving translational control as the possible mechanism for Sxl downregulation [140]. Inhibition of Sxl expression by Hrp48 was also observed in the study of Szostak et al. [72].

These experimental data suggests that the function of Hrp48 as an activator or repressor relies on the modulating effect of other trans-acting factors co-recruited to the mRNA with Hrp48.

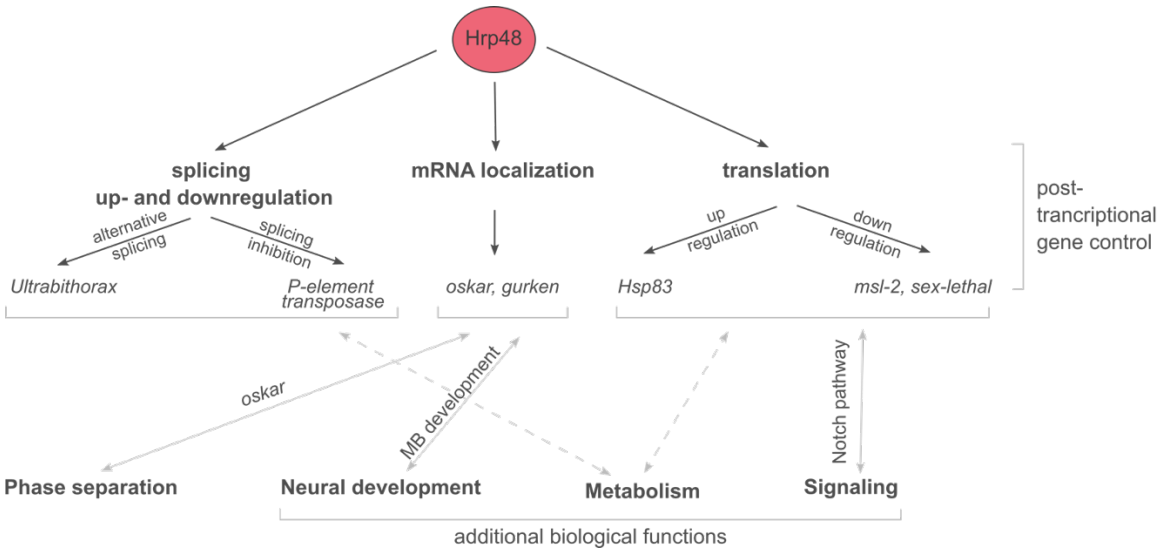


Figure 9 Various roles of the RNA-binding protein Hrp48

Hrp48 is involved in diverse post-transcriptional gene control events through its RNA-binding capabilities (upper part of the figure). Additional roles of Hrp48 in various biological events and phase-separation are also described, but most of them derive from RNA binding and can be connected to post-transcriptional mechanisms (lower part of the figure).

1.2.7.4 Additional roles of Hrp48

Hrp48 has been identified as regulatory element in functions different from post-transcriptional control, but in most of these cases, its RNA-binding properties are exploited, and the final function is mediated by translational repression or alternative splicing activity of Hrp48 (Figure 9).

Signaling pathways play a crucial role in metazoan development, enabling the cell to communicate with its environment or itself and to coordinate processes such as cell division, cell fate determination, stem cell maintenance, proliferation, differentiation, and maintaining homeostasis.

The Notch (N) signaling pathway is an evolutionarily conserved signal transduction pathway involved in embryonal development [140]. The female-specific *D. melanogaster* protein Sex-lethal has been found to downregulate the N pathway resulting in sex-specific patterning likely by translational repression of Notch [141] and Hrp48 counteracts the effect of Sxl by suppression of Sxl translation and thereby rescuing Notch expression [140]. In female tissues, the overexpression of Hrp48 results in a decrease of Sxl and consequently, an increase of Notch protein levels. The knockdown of Hrp48 leads to elevated Sxl and reduced Notch protein levels, and in both cases, there is no effect, or only a minor one in males depending on the tissue type. The minor effect on males indicates that Hrp48 might also have an Sxl-independent, hence not sex-specific effect on the N-pathway. In later studies they

observed that Hrp48 is involved in the Deltex (Dx) mediated regulation of the Notch pathway by binding to Dx [142], and affecting the transport of Notch from the membrane to the cytoplasm in an Sxl-independent fashion. Dx directly interacts with Notch and Hrp48 via the same domain, and Dx together with Hrp48 reduces the Notch protein levels. These two proteins have also been found to induce caspase-mediated apoptosis in eye tissue, through activating the JNK (c-Jun N-terminal kinase) pathway [143].

Germline stem cells (GSCs) are fundamental for reproductive ability, and they are regulated by networks of signaling mechanisms. In the *D. melanogaster* germarium, loss-of-function analyses suggest intrinsic cell-specific roles of Hrp48 along with other hnRNPs, including Sqd in GSC maintenance; and specifically, Hrp48 supports GSC proliferation [119, 144]. Further results also imply that these hnRNPs function by binding to distinct sets of nascent transcripts.

Based on findings of genome-wide studies on fat **metabolism** and storage, where genes primarily known for regulating mRNA splicing were identified as hits, *Bhogal et.al.* investigated various hnRNPs for metabolic functions by RNAi in the *D. melanogaster* fat body [145]. The fat body is a tissue functioning as liver and fat storage. Decreasing levels of Hrp48 resulted in increase in stored lipid quantity, by elevated fat body cell numbers and deposited fat per cell. Parallel to this, changes in splicing or expression of genes responsible for lipid-breakdown were also observed. This suggests that the metabolic functions of Hrp48 (and possibly other hnRNPs) is mediated through its post-transcriptional regulatory effect on targets involved in metabolism.

Hrp48, similarly to its human homologue hnRNPA1, mediates **phase separation**, as it has been shown in the context of *oskar* mRNA localization [121, 146].

hnRNP A2/B1, a member of the hnRNP A/B family of RNA-binding proteins was observed to play a role in **development of the nervous system** of humans by their ability to control mRNA localization [147-148]. *In vivo* experiments have shown that the homologous *D. melanogaster* Hrp48 has multiple functions in the morphogenesis of the mushroom body (MB) of fruit fly brains. Hrp48 regulates axonal branching and prevents ectopic axonal growth, which is the overextension of the MB in the dorsal direction [149].

2 Aims of thesis

The aim of my thesis is to understand the structural and RNA-binding properties of the *Drosophila* protein Hrp48 and its interaction with the *msl-2* mRNA. Hrp48, along with Unr and Sxl form a complex at the 3' UTR of *msl-2* mRNA, mediating the repression of *msl-2* translation. Getting structural insight into the assembly of these factors on the *msl-2* mRNA ensures a better understanding of the function of this protein-RNA complex, its role in translational repression, and also extends our sparse knowledge of multi-subunit RNP complexes and translational control. A crystal structure of the ternary Sxl–Unr–*msl-2* complex has already been solved [68] utilizing subdomains of these proteins (Figure 8 G), and the next step is to understand how Hrp48 is involved in this mRNP assembly.

First, the structural information on Hrp48 has to be explored and extended. To date, there is neither structural data on how Hrp48 is integrated into the repressor complex, nor on the structure of Hrp48 itself or its domains, apart from the AlphaFold prediction.

Next, in order to link the structural data to the biological function, the RNA-binding properties, sequence specificity, affinities, and binding mode of Hrp48 has to be explored. For this, the putative binding site of Hrp48 will be dissected into smaller motifs to obtain the optimal binding sequences for both RRM domains and the tandem RRM12 construct. The refined RNA-binding sites will be utilized in biophysical characterization of complex formation of Hrp48 with RNA. Based on this foundation, the quaternary complex formation including Sxl, Unr, Hrp48, and *msl-2* will be optimized and studied.

The refined RNA-binding site will be validated by mutational analysis using *in-vitro* translation assays. This would also give insights into the biological relevance of the previous results.

Achieving these goals requires a multidisciplinary approach, integrating structural and biophysical/biochemical techniques, computer modeling, and cell-based studies. For the structural studies, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy supported by MD simulations and rigid body modeling will be employed. NMR will give further insight into the biophysical properties and dynamics of RNA binding. Further biophysical/biochemical approaches complement these, including isothermal titration calorimetry (ITC), size-exclusion chromatography coupled with multiple-angle laser light scattering (SEC-MALLS), and electrophoretic mobility shift assay (EMSA).

3 Structural biology methodology

Structural biology studies the three-dimensional arrangement biomolecules at nearly atomic resolution in order to understand their functions. There are several methods that can be used to determine the structure of a protein or nucleic acid molecule experimentally. The three main high-resolution techniques are cryo-electron microscopy, NMR-spectroscopy and X-ray crystallography. The first atomic resolution structure solved of a biological molecule was cholesterol, solved by Dorothy Hodgkin in 1937 [150], and more than 10 years later, the first protein structure, myoglobin, was published by John Kendrew [151]. Since then, over 200.000 structures of proteins, nucleic acids and their complexes have been deposited in the Protein Data Bank [152], the majority of which were solved by X-ray crystallography (86%), and the rest by nuclear magnetic resonance spectroscopy (NMR) (7%), or by electron microscopy (7%). This impressive community effort of 200.000 structures was the foundation for the success of the artificial intelligence-driven protein structure prediction AlphaFold2, which since 2021 contributed to an additional almost 1.000.000 structure predictions [122, 152].

This great advancement also works for protein-protein complex structure perditions [153], but protein-RNA complexes and transient interactions cannot be predicted yet. Therefore, integrative structural biology methods are needed. For this, the high resolution methods can be combined and complemented with additional lower resolution techniques, such as small-angle scattering (SAS), electron paramagnetic resonance (EPR), mass spectrometry approaches (cross-linking mass-spectrometry and hydrogen-deuterium exchange mass spectrometry), and single-molecule microscopy techniques aided by computer-based structural modeling [154]. This approach is especially useful in structure determination of macromolecular complexes or multi-domain proteins with flexibility in their structural arrangement.

However, proteins and biomolecular complexes are not rigid entities – they are naturally flexible at biologically relevant temperatures and undergo complex conformational changes on different time scales as they are in a dynamic equilibrium. Dynamics refers to any time-dependent change of an atomic coordinate, for example conformational changes, such as loop motions, domain motions, rotation of side-chains; diffusion and molecular tumbling in solution, folding or catalytic turnover rate (Figure 12) [155]. Information on the rate of these processes and population of the respective states complements the knowledge on the structure, enabling a better understanding of the function of biomolecules and their complexes. Depending on the rate of the dynamic motion, different techniques need to be utilized as there is no single method capable of probing all types of dynamic motion. Various NMR-spectroscopy techniques can cover the entire range of possible time scales [156]. Other

methods can also be implemented to study molecular motions, such as single-molecule fluorescence resonance energy transfer (sm-FRET), time-resolved X-ray crystallography, time-resolved small-angle X-ray scattering and time-resolved mass-spectrometry, along with molecular dynamics simulations [157]. A great advantage of single-molecule methods compared to ensemble methods such as NMR spectroscopy, is that they can provide information on the heterogeneity of different states rather than on the average bulk description of the system.

3.1 NMR spectroscopy in structural biology

Nuclear magnetic resonance spectroscopy is a powerful and versatile physical method for elucidating the structure, interaction, and dynamic properties of small and large macromolecules. The basic background, theory, methodology, and spectral interpretation of NMR can be found in several monographs and books [158-161]. For this reason, only the basic phenomenon and principles are summarized here, along with those spectral parameters and techniques that were relevant to my thesis.

3.1.1 The principles of NMR spectroscopy

Only atomic nuclei in which either the number of protons or neutrons or both are odd, because only these nuclei have a non-zero spin quantum number, $I \neq 0$, and therefore a non-zero $\mathbf{I} \neq 0$ spin angular momentum (spinning of the nuclei about an axis crossing the center of mass), can be investigated by NMR. As the nuclei are charged, this causes a magnetic moment $\boldsymbol{\mu}$ described by the relationship $\boldsymbol{\mu} = \gamma\mathbf{I}$, where γ is the gyromagnetic ratio [158].

Without an external magnetic field, the magnetic moment vector orientation is random. In the bore of the superconducting magnet in the NMR spectrometer, there is a strong and homogeneous permanent magnetic field $\mathbf{B}_0$. Using a Cartesian coordinate system to describe the NMR phenomenon, $\mathbf{B}_0$ is defined so as to point in the z direction. When a **s**pin is placed in this field, it can only take up some specific orientations along cones: the z component of the magnetic moment vector can only have a magnitude of $\mu_z = m\gamma\hbar$, where $m = I, I-1, \ldots, -I$. The length of the magnetic moment vector has the value $|\boldsymbol{\mu}| = \gamma\hbar\sqrt{I(I+1)}$. These two conditions imply that $\boldsymbol{\mu}$ may point onto surface points of $2(I+1)$ cones. Similarly as observed in a compass, $\boldsymbol{\mu}$ tries to align itself towards one pole of the z axis, but it is only allowed to adopt these limited number of orientations. Since $\boldsymbol{\mu}$ cannot be aligned parallel to the $\mathbf{B}_0$ axis, classical physics states that $\boldsymbol{\mu}$ will experience a torque $\mathbf{T} = \boldsymbol{\mu} \times \mathbf{B_0}$ that forces the $\boldsymbol{\mu}$ vector to rotate on the surface of the circular cone at a constant rate. This phenomenon is called the Larmor precession, and its frequency can be described by

$$\nu_{\mathrm{L}} = \frac{\gamma B_0}{2\pi}$$

where ν_{L} is the Larmor frequency, the frequency of the circular motion about the $\mathbf{B}_0$ axis. ν_{L} is directly proportional to the gyromagnetic ratio, which is an intrinsic atomic property dependent on the isotope and the external field, which is dependent on the strength of the NMR magnet. For high-resolution liquid-state NMR spectroscopy of biomolecules, experiments on the ^{1}H, ^{13}C, ^{15}N and ^{31}P nuclei are the most relevant, for which nuclei $I = 1/2$, denoting for 2 possible cones for Larmor precession [158].

It is vital to make a distinction between the quantum mechanical description and the classical physical vector model of NMR. For the purposes of this discussion, the following description is restricted to protons in the framework of an often used but superficial and limited vector model of NMR spectroscopy [158].

Considering an ensemble of a large number of $I = 1/2$ nuclei that are present in the NMR sample, the magnetic moments are aligned to two oppositely orientated cone surfaces. As the two possible alignments do not have the same energy, their population must differ according to the Boltzman distribution. This means there is a very slight difference between the numbers of magnetic moments aligned parallel (slightly preferred) and antiparallel to the *z* axis. For this reason, the vector sum of the individual $\boldsymbol{\mu}$ magnetic moments gives the macroscopic magnetization vector $\mathbf{M} = \sum \boldsymbol{\mu}$. In equilibrium, this bulk magnetization is not zero (Boltzmann-equilibrium state) and lies parallel with the z axis (if $\gamma > 0$, if $\gamma < 0$ anti-parallel), therefore it is equal to its z component $\mathbf{M}_{\mathrm{z}} = \mathbf{M}$ and no component is present in the (*x*,*y* transversal) plane, $\mathbf{M}_{\mathrm{xy}} = 0$. For acquiring an NMR spectrum, a specific non-equilibrium state of the bulk magnetization is needed, because the $\mathbf{M}_{\mathrm{xy}}$ component is what we can measure. In order to tilt **M** from its equilibrium position, an additional $\mathbf{B_1}$ magnetic field is needed that rotates with the Larmor frequency about $\mathbf{B_0}$. This causes resonance and rotates **M** from the z axis to the x,y plane, so $\mathbf{M}_{\mathrm{xy}} \neq 0$. The $\mathbf{B_1}$ field is transmitted by sinusoidal alternating radio frequency pulses via the coil of the NMR spectrometer probe head [158].

After the pulse, the nuclear magnetic moments return to the equilibrium state as a result of relaxation mechanisms. The $\mathbf{M}_{\mathrm{xy}}$ component disappears gradually (called transversal / spin-spin relaxation), while $\mathbf{M}_{\mathrm{z}}$ returns to its initial state (longitudinal / spin-lattice relaxation) [162]. During this process $\mathbf{M}_{\mathrm{xy}}$ rotates with the Larmor frequency in the x,y plane, inducing a sinusoidal alternating current in the receiver coil of the NMR spectrometer probe head.

The FID (free induction decay) is the superposition of temporally sinusoid/harmonic currents induced by the different decaying precessions of different spins detected by the coil. In order to differentiate

between the constituent frequencies, the FID (signal intensity – time function) has to be Fourier transformed to an NMR spectrum (signal intensity – frequency function).

The appearance of signals in the one-dimensional ^{1}H NMR spectrum is determined by four main spectral parameters, all of them supplying useful information about the molecular structure. The **chemical shift** reflects the chemical or electronic environment of each nucleus. Its basic principle lies in the relation between the nuclei and the secondary field (described by the shielding factor σ) induced by the surrounding electrons. As a result of this shielding, a given *i* nucleus in the molecule is in a slightly different magnetic field strength B_i in comparison with an isolated, bare nucleus, that would sense the B_0 field [158]:

$$B_\mathrm{i} = B_0(1 - \sigma_\mathrm{i})$$

For this reason, each nucleus with a specific chemical environment has its own Larmor frequency, and a specific signal in the spectrum, making it possible to distinguish between the signals of the nuclei:

$$\nu_\mathrm{i} = \frac{\gamma B_\mathrm{i}}{2\pi}$$

Since Larmor frequencies cannot be measured on an absolute scale by NMR, a reference material is needed (with the resonance frequency of ν_ref) to define the chemical shift (δ) which is a dimensionless quantity:

$$\delta(\mathrm{ppm}) = \frac{(\nu - \nu_\mathrm{ref})}{\nu_\mathrm{ref}} 10^6$$

Thus, chemical shift values are not dependent on the different B_0 of different spectrometer field strengths. In biomolecular NMR the primary reference material is H_2O. In case of multinuclear experiments, when simultaneously more nuclei are magnetized, the referencing of only one nucleus is required, while the rest can be calculated from this via indirect referencing, a common practice in biomolecular NMR [163]. **Spin-spin coupling** or scalar coupling is a magnetic interaction between nuclei mediated by electrons. It can be homonuclear or heteronuclear according to the types of interacting nuclei (same isotopes or different ones). Among other things, spin-spin coupling provides information on the number of the neighboring nuclei, but very importantly can be exploited for magnetization transfer through covalent bonds in multinuclear experiments [164-165]. The **integral** is the area of the NMR peaks and provides information about the number of nuclei giving rise to the signal. The **NMR linewidth** is the half peak width of the peaks. It is determined mainly by the transversal T_2 relaxation, the rate at which transversal magnetization decays and the macroscopic magnetization returns to equilibrium, but exchange processes can also influence it.

3.1.2 NMR spectroscopy of proteins

In NMR spectroscopy of proteins, the one-dimensional ^{1}H NMR spectrum (Figure 10) contains lots of overlaps because all the hydrogens in a protein give rise to a signal with similar chemical shift values. The overlaps make it impossible to identify the signals separately, and there is spectral information that cannot be extracted from this type of spectrum. To gain more information for the structure elucidation or dynamical characterization of the molecules, it is the common practice to use extra dimensions such as the resonance frequency of different nuclei, coupling constants, diffusion coefficient and so on, which can reduce the complexity of the spectrum by separating the overlapping peaks and reveal the additional spectral information.

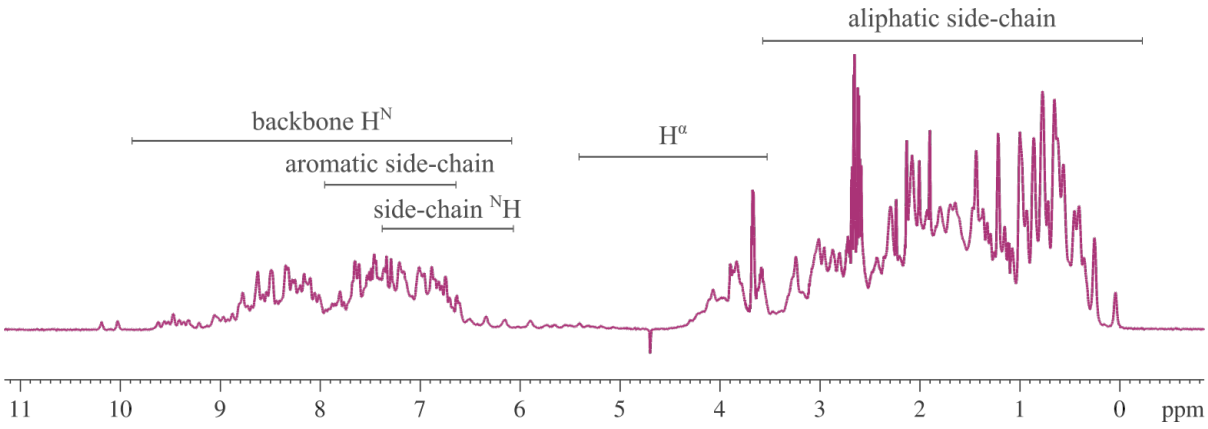


Figure 10 One-dimensional proton spectrum of Hrp48 RRM1

The figure indicates the characteristic chemical shift ranges corresponding to different types of hydrogen nuclei in a protein. The spectrum indicates that the protein is folded, because the peaks appear well dispersed. The one-dimensional spectrum is useful for quality check of protein samples, but multi-dimensional experiments are needed to extensively analyze proteins by NMR spectroscopy.

A powerful experiment for protein NMR spectroscopy is the ^{1}H,^{15}N-HSQC (Heteronuclear Single Quantum Coherence) spectrum which is a heteronuclear experiment that correlates the chemical shifts of 1H and 15N and thus reports on the backbone conformation. In a protein, each amino acid has one bond like this in the amide group, except of prolines. This makes the ^{1}H, ^{15}N-HSQC a fingerprint spectrum of proteins. Additional N-H peaks can arise from side-chains of asparagine, arginine, glutamine, lysine or tryptophan. In order assign each peak in the ^{1}H,^{15}N-HSQC to its corresponding amino acid in the sequence of the observed protein, additional experiments are needed. Based on the size of the protein, different assignment strategies are possible, and here I describe the heteronuclear assignment based on doubly labelled (^{13}C and ^{15}N) proteins that I used for my thesis work. As the natural abundance of NMR-active nuclei ^{13}C and ^{15}N is low, they have to be enriched during

expression of the protein. This is achieved with special growth media for cell cultures in which the carbon and nitrogen source consists of ^{13}C and ^{15}N isotopes (e.g., D-Glucose-$^{13}C_6$ and $^{15}NH_4Cl$ in Minimal medium for bacterial cultures). For the heteronuclear assignment, various ^{1}H,^{13}C,^{15}N triple-resonance experiments are acquired, such as the HNCA, HNCACB and CBCA(CO)NH. These experiments can correlate the amide NH resonances of each amino acid with the C_α and C_β side-chain chemical shifts of the same and the predecessor spin systems [166-167]. Based on this sequential information, the spin systems can be linked to each other. Due to the known characteristic C_α and C_β chemical shifts of each type of amino acid, the linked spin systems can be linked to the protein sequence. Thereby, the complete the assignment of the protein backbone can be achieved.

3.1.3 NMR interaction studies and chemical exchange

Chemical exchange is a process in which a spin exchanges between two or more environments. This can happen due to dynamic chemical or structure-rearrangement equilibria, a phenomenon which results in a periodic fluctuation of the local magnetic field. Thus, the nuclei are exposed to different magnetic environments as time passes, for example when a ligand binds or when protonation – deprotonation events occur. This can have an effect on the peak width (relaxation) and integral values, the chemical shift, or the scalar coupling [158]. According to the frequency of this change, three different scenarios are possible (Figure 11). In the following I describe a bimolecular equilibrium to exemplify e.g. ligand binding of a protein:

$$P + L \underset{k_{diss}}{\overset{k_{ass}}{\rightleftharpoons}} PL$$

Where P is the protein and L is the ligand, PL denotes for the complex, k_{ass} and k_{diss} are the rates of association and dissociation. The dissociation constant for this reaction is defined as:

$$K_D = \frac{[P][L]}{[PL]} = \frac{k_{diss}}{k_{ass}}$$

In diffusion-controlled binding, k_{ass} is typically around 10^{-9} $M^{-1}s^{-1}$ [168] and k_{diss} can be approximated as:

$$k_{diss} \approx 10^{-9} K_D$$

For this reaction, the rate of exchange (k_{ex}) which is the rate of changing between P and PL states can be described as:

$$k_{ex} = [L]k_{ass} + k_{diss} \approx k_{diss} \approx 10^{-9} K_D$$

This means there is a direct correlation between the k_{ex} exchange rate and the K_D dissociation constant of the binding.

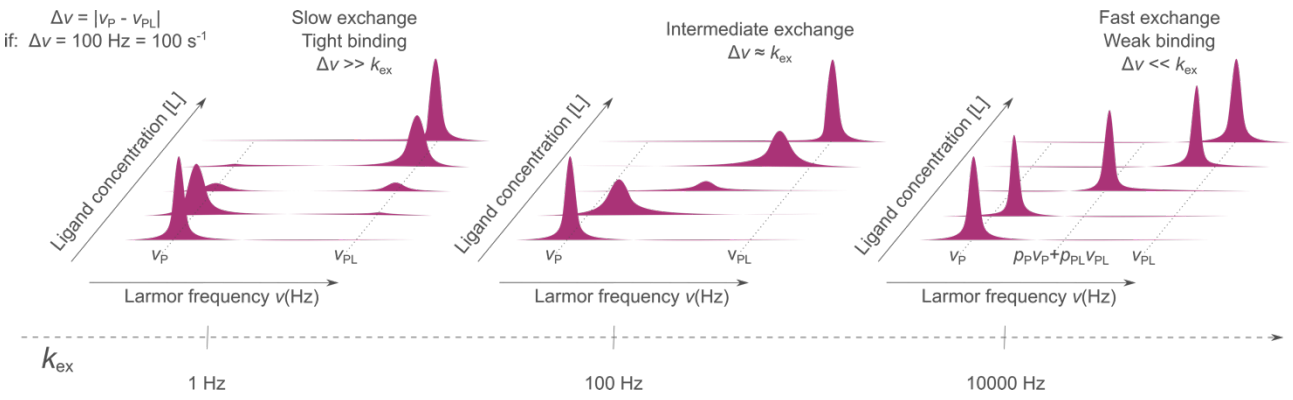


Figure 11 Two-state chemical exchange

The three types of chemical exchange regimes represented by a series of illustrated one dimensional spectra for each scenario. Each series shows how the change of the ratio of P and PL states (varying ligand concentration) changes the appearance of the spectra. ν_P and ν_{PL} denote the resonance frequencies of the P and PL states of the same spin respectively, and their difference ($\Delta\nu$) was chosen to be 100 Hz. The relation between the exchange rate k_{ex} (lowest 'axis') and $\Delta\nu$ determines which exchange regime takes place.

According to the relation between the k_{ex} exchange rate and the $\Delta\nu = |\nu_P - \nu_{PL}|$ difference in resonance frequencies, there are three different scenarios of chemical exchange.

In the purely **slow exchange** regime, the rate constant k_{ex} of the exchange is much smaller than the difference in resonance frequencies $\Delta\nu = |\nu_P - \nu_{PL}|$ ($k_{ex} \ll \Delta\nu$). In this case, distinct NMR signals of the interconverting species are detected at ν_P and ν_{PL} frequencies and with changing A:B ratio of the peak width and the integral values (Figure 11 left panel). In the case of the **fast exchange** ($k \gg \Delta\nu$), the lines of the interchanging nuclei merge into a common signal, which has the frequency of the fractional population-weighted averaged frequencies of the component states ($\bar{\nu} = p_P\nu_P + p_{PL}\nu_{PL}$) (Figure 11 right panel). The **intermediate exchange** is the combination of the two limiting cases. This happens when k_{ex} is similar in magnitude to $\Delta\nu$. The separate resonances first coalesce into a broad signal, then sharpen again while shifting from ν_P to ν_{PL}. (Figure 11 middle panel) The exchange regimes indicate the binding affinity: slow exchange is characteristic for tighter binding, while fast exchange indicates a weaker interaction. These three states are often used for illustrating the phenomenon of chemical exchange and its effect on NMR signals. However, chemical exchange cannot simply be categorized into three states and differences can be subtle depending on the exchange rate. Thus, upon ligand titration, the receptor resonances could also exhibit changes of peak intensity and chemical shift, which lie in between fast-to-intermediate or intermediate-to-slow

exchange. Also, different signals of one protein can reflect different exchange behaviors in a single titration.

The perturbation in the chemical shift values upon ligand binding is the result of the changed magnetic environment for the nuclei of the protein and can be quantified by measuring the difference of the two chemical shift values, termed as chemical shift perturbation, *CSP*. This difference for a two-dimensional ^{1}H, ^{15}N correlated spectrum is defined as

$$CSP = \sqrt{\frac{1}{2}[\delta_{\mathrm{H}}^2 + (\alpha \cdot \delta_{\mathrm{N}})^2]},$$

where δ_{i} are the chemical shift differences in the respective dimension and α is a scaling factor calculated as the ratio of the frequency range of ^{15}N and ^{1}HN chemical shifts [168].

For chemical exchange in the purely fast exchange regime, where the δ_{i} is directly proportional to the protein:ligand ratio, the K_{D} dissociation constant values can be determined by fitting the following equation for each amino acid:

$$\delta_{\mathrm{OBS}} = \frac{\delta_{\mathrm{MAX}}\left((n[\mathrm{P}]_{\mathrm{TOT}} + [\mathrm{L}]_{\mathrm{TOT}} + K_{\mathrm{D}}) - \sqrt[2]{(n[\mathrm{P}]_{\mathrm{TOT}} + [\mathrm{L}]_{\mathrm{TOT}} + K_{\mathrm{D}})^2 - 4n[\mathrm{P}]_{\mathrm{TOT}}[\mathrm{L}]_{\mathrm{TOT}}}\right)}{2n[\mathrm{P}]_{\mathrm{TOT}}}$$

Where δ_{OBS} is the observed *CSP* at a given point of the titration, δ_{MAX} is the shift change in saturation, n is the number of (equivalent) binding sites, $[\mathrm{P}]_{\mathrm{TOT}}$ and $[\mathrm{L}]_{\mathrm{TOT}}$ the total protein and ligand concentration [168].

For chemical exchange in the purely slow exchange regime, the integral values of the peaks corresponding to the ligand bound and unbound state can used in the same equation to determine the K_{D}. This is because the peak integrals are directly proportional to the protein:ligand ratio in the slow exchange regime. However, in the slow exchange regime, the peaks can undergo severe line broadening making it difficult to determine the integrals reliably in most cases.

If the exchange regime is not purely slow or fast, it is not possible to determine the K_{D} by fitting the same equation, because in these cases the spectral parameters are not directly proportional to the protein:ligand ratio. For this exchange regime, complicated and often inaccurate line shape fitting or competition binding experiments can be used.

As the peaks in the fast exchange regime shift gradually upon addition of the ligand, the peak position of the bound state is easy to trace by following the shift changes. This does not apply for the slow exchange, as there the peaks reappear directly at the chemical shift positions of the bound state and for this reason the chemical shift assignments have to be repeated in the ligan bound state to identify the peaks [168].

3.1.4 Structure determination by NMR spectroscopy

NMR structure determination is different to X-ray crystallography or cryo-EM in the sense that one does not obtain absolute coordinates, but more indirect structural information. These structural restraints are then used in computer-based structural modeling of the protein or other molecule. There are several structural restraints that can be measured by a variety of NMR experiments, but the most valuable ones are the NOE (Nuclear Overhauser Effect [169]) restraints. NOE is the spread of the magnetization between spins through relaxation processes over space and not bonds. With this, proton nuclei closer than 5 Å can be detected and the peak intensity correlates with the distance. Additional types of experimental distance restraints can be added to the structure calculations: dihedral angle restraints are calculated from backbone chemical shift values [170], torsion angles from spin-spin couplings derived by the Karplus-equation [171], paramagnetic relaxation enhancement (PRE) restraints delivering distance information up to 30 Å with the help of paramagnetic spin labels [162, 172], hydrogen bonds measured through H/D exchange rates or coupling constants, and residual dipolar couplings (RDC) giving information on the relative orientation of bond axes and domains with respect to an alignment tensor [173]. This structural information is complemented with general understanding of properties of chemical bonds and angles, chirality, or other intrinsic spatial arrangement of atoms, all of which are considered in the modeling. The modeling is most commonly computed by molecular dynamics (MD) algorithms with simulated annealing, where a random coil polypeptide is generated and the equation of motion is numerically solved in iterative rounds on Cartesian coordinates or on the torsion angle space taking all experimental restraints into account. To enable all this, complete (or close to complete) chemical shift assignments (backbone and side chain resonances) are required, for which having a stable and concentrated (minimum 0.5 mM) sample is a prerequisite.

3.1.5 Relaxation and dynamics

The ability to study protein dynamics at all time scales is one of the advantages of NMR spectroscopy compared to other structural methods. NMR spectroscopy is suitable for measuring a broad range of dynamic events for which various pulse experiments were developed (Figure 12).

In my thesis work, I used ^{15}N spin relaxation experiments, which allows to determine local and global dynamics at the ps – ns time scale, which is the range for local molecular motions and molecular tumbling. These motions can be assessed by measuring longitudinal (R_1), and transverse (R_2) relaxation rates and heteronuclear NOEs. As previously mentioned, after the pulse, the bulk magnetization returns to the equilibrium state as a result of relaxation mechanisms. This can be separated into two components and measured independently by various NMR experiments:

exponential decay along the z axis or along the x-y plane. Relaxation can arise from different sources, such as two-spin dipolar interactions, where the local field generated by one spin is sensed by the other spin. This depends on the gyromagnetic ratio, distance, and relative orientation of the interacting spins. Chemical shift anisotropy, chemical exchange processes, and paramagnetic relaxation also contribute to the relaxation.

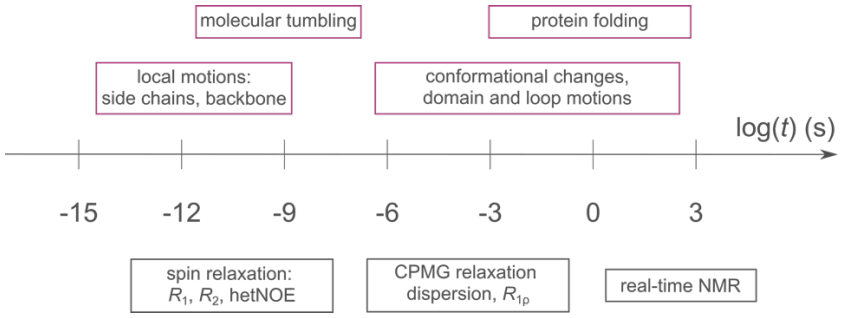


Figure 12 Time scale for dynamic changes of proteins

This scale represents the types of motions with their respective time scales (top) and the matching NMR technique to measure them (bottom).

The recovery of the bulk magnetization in the z-direction is the result of the longitudinal / spin-lattice relaxation mechanisms and can be described by the R_1 relaxation rate, which is the constant of the exponential decay. The polarized spins are in a non-equilibrium state and can interact with the magnetic field induced in the lattice (the molecule and its environment) by thermal motion of the lattice nuclei. The energy of the higher energy polarized state is dissipated in the lattice, causing amplified vibrations and rotations, and temperature increase. The loss of coherence of the bulk magnetization in the x-y plane is due to the transverse or spin-spin relaxation. The factor of this exponential decay is R_2 and is usually faster than R_1 which means that first the x-y component of magnetization disappears and then the z component rises back to its equilibrium state. The source of the spin-spin relaxation is the direct interaction of the neighboring spins causing the dephasing from one another.

The rotational correlation time (τ_c), which is the time it takes for the molecule to rotate 1 radian, is an essential parameter to analyze molecular tumbling motions by NMR relaxation experiments. It can be calculated for each amino acid locally and also globally for individual domains or the whole protein. τ_c values of individual amino acid residues can give important information on the local dynamics and conformational changes. From global correlation times the molecular weight can be

estimated and it applies generally that the higher the correlation time is, the higher the molecular weight is. For monomeric globular molecules in solution $\tau_c \approx 0.6 \cdot MW$.

For calculating the rotational correlation times, the Lipari – Szabo model free approach is frequently used [174]. This can extract from the R_1, R_2, and heteronuclear NOE data the rate of motion globally and internally, and also give information on potential chemical exchange processes. With the assumption of isotropic tumbling for rigid molecules, the equations of the Lipari – Szabo model can be simplified as the following [175]:

$$\tau_c \approx \frac{1}{4\pi\nu_N} \cdot \sqrt{\frac{6R_2}{R_1} - 7}$$

Where ν_N is the ^{15}N resonance frequency in Hz, R_1, R_2 are the relaxation rates and τ_c is the rotational correlation time in seconds. This approximation can be used to directly estimate the rotational correlation time from the R_1 and R_2 values numerically.

4 Materials and methods

4.1 Materials

4.1.1 Devices and consumables

Table 1 List of consumables

Consumable	Manufacturer
Amicon Ultra Centrifugal Filters	Merck Millipore
Disposable 10 mL Polypropylene Columns	Thermo Scientific
G_{50} desalting columns	GE Healthcare/Cytiva
HiLoad Superdex™ 75 pg 26/600	GE Healthcare/Cytiva
HiLoad Superdex™ 75 pg 26/600	GE Healthcare/Cytiva
HiLoad Superdex™ 200 pg 16/600	GE Healthcare/Cytiva
HisPur™ Ni-NTA Resin	Thermo Scientific
HiTrap™ 5 mL Heparin HP	GE Healthcare/Cytiva
Millex-HV Filter, 0.45 µm, PVDF	Merck Millipore
Millex-HV Filter, 0.22 µm, PVDF	Merck Millipore
Mini-Protean TGX Precast gradient gel 4-20%	BioRad
Oak Ridge Centrifuge Tube, 50 mL	Thermo Scientific
Polypropylene centrifuge bottle, 1000 mL	Beckman Coulter
RotiGarose-His/Ni NTA-HPBeads	Carl Roth
Shigemi 5 mm NMR microtube assembly	Shigemi
Spectrum™ Spectra/PorTM 3 RC Dialysis Membrane Tubing	Fisher Scientific
Standard Series NMR tubes	Norell
Superdex 200 Increase 10/300 GL	GE Healthcare/Cytiva
Whatman™ Cellulose Membrane Filters	GE Healthcare

All additional consumables correspond to the usual laboratory standards.

Table 2 List of used devices

Device	Manufacturer
Allegra™ 64R Centrifuge	Beckman Coulter
Avance III HD 600 MHz spectrometer	Bruker
Avance III HD 700 MHz spectrometer	Bruker

Avance III HD 800 MHz spectrometer	Bruker
Avanti JXN-26 (JLA-8.1000 rotor)	Beckman Coulter
Äkta Explorer FPLC	Merck Millipore
Centrifuge 5810	Eppendorf
Heraeus Fresco™ 17 microcentrifuge	Thermo Scientific
ISF-1-W incubator shaker	Kühner
Mini PROTEAN 3 Multi-Casting Chamber	BioRad
Microcal PEAQ-ITC	Malvern Panalytical
MiniDAWN and Optilab	Wyatt Technology
Mini PROTEAN Tetra Cell	BioRad
Mosquito LCP	TTP LabTech
NanoDrop™ 2000	Thermo Scientific
NanoDrop™ OneC	Thermo Scientific
NGC Medium-Pressure Liquid Chromatography System	BioRad
PowerPac™ Basic Power Supply	BioRad
Prometheus NT.48 nanoDSF	NanoTemper Technologies
S20 - SevenEasy™ pH meter	Mettler Toledo
Typhoon Trio Imager 9000	GE Healthcare
Ultrospec™ 2100 pro, UV/Vis spectrophotometer	Amersham Biosciences

All additional devices correspond to the usual laboratory instruments.

Table 3 List of commercial assays and kits

Kit	**Manufacturer**
MinElute Gel Extraction Kit	Qiagen
Plasmid Maxi Kit	Qiagen
Pierce™ BCA Protein Kit Assay	ThermoFisher
Pierce™ Silver Stain Kit	ThermoFisher
QIAprep Spin MiniPrep Kit	Qiagen

4.1.2 Chemical compounds

Table 4 List of used chemical compounds

Compound/Product name	**Manufacturer**
Ammonium-^{15}N chloride	Sigma-Aldrich
cOmplete, EDTA-free protease inhibitor cocktail tablets	Roche
Deuterium oxide	Eurisotop
Dithiothreitol (DTT)	Biomol
D-Glucose-$^{13}C_6$	Sigma-Aldrich
D-Glucose-1,2,3,4,5,6,6-d_7	Sigma-Aldrich
Isopropyl-ß-D-thiogalactopyranosid (IPTG)	Carl Roth
GelRed Nucleic Acid Gel Stain	Biotium
Pf1 phage cosolvent	Asla Biotech
Purple Gel Loading Dye (6x)	NEB
Kanamycin	Carl Roth
PageRuler Low Range Unstained Protein Ladder	Thermo Scientific
PageRuler Broad Range Unstained Protein Ladder	Thermo Scientific
pCp-Cy5	Jena Biosciences

All additional chemicals were purchased from Sigma-Aldrich, Carl Roth or Merck and the quality correspond to the usual laboratory standards.

4.1.3 Composition of buffers

Table 5 Composition of buffers

Medium	**Compound**	**Concentration/quantity**
M9 salts 10x	ddH_2O	900 ml
	NaH_2PO_4	60 g
	K_2HPO_4	30 g
	NaCl	5g
	Adjust pH to 7.4	
	ddH_2O	To 1000 ml
	Autoclave	
SDS 10x	Tris - HCl	250 mM

	Glycine	2.5 M
	SDS	1% w/v
TAE 50x	Tris base	242 g
	Acetic acid	57 ml
	0.5 M EDTA, pH 8.0	100 ml
	ddH_2O	To 1000 ml
TBE 10x	Tris base	108 g
	$B(OH)_3$	55 g
	0.5 M EDTA, pH 8.0	40 ml
	Adjust pH to 8.3	
	ddH_2O	To 1000 ml
Trace elements 100x	ddH_2O	900 ml
	EDTA	5 g
	Adjust pH to 7.5	
	In the following order, wait until each salt dissolved.	830 mg
	$FeCl_3 \cdot 6H_2O$	84 mg
	$ZnCl_2$	13 mg
	$CuCl_2 \cdot 2H_2O$	10 mg
	$CoCl_2 \cdot 6H_2O$	10 mg
	$B(OH)_3$	1.6 mg
	$MnCl_2 \cdot 6H_2O$	To 1000 ml
	ddH_2O	
	Stir overnight at 4 °C	
Lysis buffer	NaCl	500 mM
	Hepes pH 7.2	20 mM
	Imidazole	12 mM
	β-Mercaptoethanol	0.5 mM
	Lysozyme	0.25 mg/ml
	DNaseI	15 µl
	cOmplete protease inhibitor	1 tablet
	ddH_2O	100 ml final volume
His binding buffer	NaCl	500 mM
	HEPES, pH 7.2	20 mM
	Imidazole	12 mM
TEV cleavage buffer	His binding buffer	1000 ml
	β-Mercaptoethanol	2mM
His elution buffer	NaCl	500 mM

	HEPES pH 7.2	20 mM
	Imidazole	200 mM
Heparin binding buffer	NaCl	50 mM
	MES pH 6.0	10 mM
	DTT	2mM
Heparin elution buffer	NaCl	1500 mM
	MES pH 6.0	20 mM
	DTT	2mM
MES/SEC buffer	MES pH 6.5	20 mM
	NaCl	200 mM
	NaN_3	0.02%
	Autoclave	

4.1.4 Composition of media

Table 6 Composition of media

Medium	**Component**	**Quantity**
LB (Luria – Bertani medium)	Tryptone	10 g
	Yeast extract	5g
	NaCl	0.5 g
	ddH_2O	To 1000 ml
M9 Minimal Medium	ddH_2O	850 ml
	Trace elements 100x	10 ml
	Thiamin – HCl	6 mg
	1 M $MgSO_4$	2 ml
	1 M $CaCl_2$	100 µl
	NH_4Cl	0.5 g
	Glucose	4 g
	M9 salts 10x	100 ml
	ddH_2O	To 1000 ml
	Sterile filtration	
	(* NH_4Cl and Glucose with needed isotopic labeling)	
SOC (Super Optimal Broth with Catabolite repression)	Tryptone	20 g
	Yeast extract	5 g
	NaCl	0.5 g
	KCl	0.2 g

	$MgCl_2 \cdot 6H_2O$	2 g
	$MgSO_4 \cdot 7H_2O$	2.5 g
	ddH_2O	To 960 ml
	Autoclave	
	Sterile 50% Glucose solution	40 ml
TB (Terrific Broth)	Tryptone	12 g
	Yeast extract	24 g
	Glycerin	4 ml
	ddH_2O	To 900 ml
	Autoclave	
	Sterile phosphate buffer (0.17 M KH_2PO_4, 0.72 M K_2HPO_4)	100 ml

4.1.5 Enzymes

Table 7 List of enzymes

Enzyme	**Source**
DNaseI	Merck
DpnI	NEB
Green Phusion mix for PCR	In-house (PEP Core Facility)
His_6-TEV protease	In-house (PEP Core Facility)
HF Phusion mix	In-house (PEP Core Facility)
Lysozyme from chicken egg white	Genaxxon bioscience
T4 RNA ligase	NEB

4.1.6 Bacterial cell lines

Table 8 List of cell lines

Cell line	**Producer**
E. coli BL21 (DE3)	NEB
E. coli DH5-Alpha	NEB

4.1.7 RNA oligonucleotides

Table 9 List of RNA nucleotides

Name	Sequence	Source
4-mer-1	AACC	Biomers
4-mer-2	CCUA	Biomers
4-mer-3	UAGG	Biomers
4-mer-4	GGAU	Biomers
4-mer-5	AUUA	Biomers
4-mer-6	UAAG	Biomers
6-mer-1	AACCUA	IDT / Biomers
6-mer-2	UAGGAU	IDT / Biomers
6-mer-3	AUUAAG	IDT / Biomers
6-mer-4	AAGAAC	IDT / Biomers
6-mer-5	UUUUUU	Biomers
8-mer	AACCUAGG	IBA
10-mer	UAGGAUUAAG	IDT / Biomers
14-mer	AACCUAGGAUUAAG	IDT
WT-30-mer	UUUUUUUGAGCACGUGAACCUAGGAUUAAG	Biomers
GG-30-mer	UUUUUUUGAGCACGUGAAGGUAGGAUUAAG	Biomers
UU-30-mer	UUUUUUUGAGCACGUGAAUUUAGGAUUAAG	Biomers

4.2 Methods

4.2.1 Protein cloning, expression, and purification

The sequences of Hrp48 (UniProt code P48809), RRM1 (1 – 88), RRM2 (89 – 173), and the tandem RRM12 (1 – 173) constructs were cloned into the pETM-11 and pETM-22 expression vectors using the restriction-free cloning method. [176]. pETM-11 comprises of an N-terminal His_6-tag and tobacco etch virus (TEV) protease cleavage site, and the pETM-22 expression vector comprises of a His_6-tag and a TRX (thioredoxin) solubility tag and 3C cleavage site. The cloned Hrp48 constructs were tested and the pETM-11 constructs were chosen for expression in Escherichia coli BL21 (DE3). Sxl dRBD3 (123 – 294) was obtained from pET-24d(+) (Novagen/Merck), comprising of a His_6-tag and a TRX (thioredoxin) solubility tag with (TEV)-cleavage site (UniProt: P19339), Unr CSD1 (181 – 252), CSD12 (186 – 344), CSD123, (186 – 414), CSD456 (422 – 677), and CSD789 (757 – 991) were obtained from pETM-11 comprising of a His_6-tag with (TEV)-cleavage site (Uniprot Q9VSK3).

For all experiments, TB or LB medium was used as expression medium, except for protein samples for NMR spectroscopy. For NMR, M9 minimal medium was used, supplemented for isotope labelling with $^{15}NH_4Cl$ for ^{15}N labeled samples, with $^{15}NH_4Cl$ and D-Glucose-$^{13}C_6$ for ^{15}N and ^{13}C doubly labeled samples. Generally, precultures were grown in the same medium as the expression medium overnight at 37°C. Cultures were inoculated with an OD600 of 0.04 and grown until reaching half of the final OD600 of the preculture, then induced by the addition of 0.2 mM IPTG and expressed protein overnight at 18°C.

The harvested cells were resuspended in ice cold lysis buffer and then frozen and sonicated for further lysis. The thawed lysate was centrifuged for 30 mins at 18000×g at 4°C and the supernatant was syringe-filtrated with 0.45µm pore size filter. For the purification the cleared lysate was loaded 3 times on a Nickel-nitrilotriacetic acid (Ni-NTA) gravity flow column that was pre-conditioned with his binding buffer. The column was washed with 20 CV of his binding buffer and eluted with 5 CV of elution buffer. The His_6-tag was cleaved by addition of TEV protease and 20mM β-Mercaptoethanol, incubated on ice for 1 hour, and then the sample was dialyzed into dialysis buffer overnight at 4°C. Next, the sample was dialyzed for an hour against his binding buffer, to remove the β-Mercaptoethanol, which would damage the column. The sample was re-applied to the Ni-NTA gravity flow column and the flow-through was collected. For Unr CSD1, CSD12, and CSD123, the reverse Ni-affinity step was followed by a heparin purification. For this, the flow-through was diluted five times to lower the concentration, and then dialyzed against the low salt heparin binding buffer overnight. The protein was injected onto a HiTrap™ 5 mL Heparin HP, washed with heparin binding buffer and eluted with a heparin elution buffer to remove bacterial RNA contamination. All proteins

were further purified by SEC on a HiLoad 16/600 Superdex S75 or S200 pg column equilibrated with SEC buffer. For Hrp48 RRM2 and RRM12, the Ni-NTA flow-through was not concentrated before SEC and was loaded in multiple 5 ml fractions to avoid aggregation, and the pooled fractions were kept at low concentration (0.1-0.2 mg/ml), and then concentrated directly before further analysis. Protein quality was assessed by Coomassie staining.

4.2.2 Complex formation

For the complex of Sxl-Unr-Hrp48-*msl-2*, the Sxl-dRBD3, Unr-CSD12, or CSD1, Hrp48-RRM12 constructs purified using SEC buffer, and the WT-30-mer or GG-30-mer RNA (Table 9) including binding sites for all three proteins were incubated on ice with a 2:3:2:1 ratio. The ratio was optimized through several SEC-MALLS (size-exclusion chromatography-coupled multiangle laser light scattering) experiments. For this, the protein concentrations were set to 50 µM before mixing. First, the Sxl construct was mixed with the Unr, then this mixture was pipetted quickly onto the GG-30-mer already in a tube. The mixture was incubated on ice for 15 mins, then Hrp48 was pipetted in and mixed. This mixture was incubated for 30-45 mins on ice. The mixture was concentrated with a 3.5 kDa cut-off concentrator unit. Depending on the required quantity, the final volume was ~300 µL for SEC purification on a Superdex 75 10/300 GL column, or ~1 mL for the HiLoad 16/600 Superdex 75 pg column. The identity of the complex peak was confirmed by UV absorption measurement at 260 and 280 nm and also using SEC-MALLS weight determination.

For the ITC experiment with pre-formed Sxl-Unr-*msl-2*, the same method was used, except of the addition of Hrp48. This was only performed with the GG-30-mer, to avoid the duplex formation of the wild-type RNA. After size-exclusion chromatography, the pooled fractions were concentrated and the concentration was measured at 260 nm on a NanoDrop UV-VIS absorption spectrophotometer with an extinction coefficient of 325000 ($M^{-1}cm^{-1}$), that is the sum of the extinction coefficients of the components at 260 nm. For this, with known concentrations of the proteins, I first determined the extinction coefficients of Sxl and Unr at 260 nm.

4.2.3 Size exclusion chromatography-multi-angle laser light scattering

In order to optimize and validate the complex formation, SEC-MALLS experiments were used. For this, different columns were used: a Superdex 200 Increase 10/300 GL or a Superdex 200 Increase 5/150 GL gel-filtration column coupled to the MiniDAWN and Optilab MALLS system from Wyatt Technology. Several different samples were prepared: for each series, the single components (proteins and RNA alone) were measured, then mixtures of these at different ratios. After mixing and

incubation on ice, the samples were centrifuged for 10 mins at 15000 rpm and then 60 µL was loaded for the Superdex 200 Increase 5/150 GL or 100 µL for the Superdex 200 Increase 10/300 GL columns. The minimal concentration was 1.0 mg/mL for each protein. The experiments were performed at room temperature in SEC buffer filtered twice through 0.22 µm pore-size filter. Data analysis was performed using the Astra 7 software (Wyatt Technology).

The SEC-MALLS experiments were performed together with Karine Lapouge (EMBL PEPCORE, Biophysical Characterization).

4.2.4 NMR data acquisition

All NMR spectra were recorded on Bruker Avance III NMR spectrometers with magnetic field strengths of 14.1, 16.4 and 18.8 T, corresponding to proton Larmor frequencies of 600 MHz, 700 MHz or 800 MHz equipped with a room temperature triple resonance probe head (700 MHz), or a cryogenic triple resonance gradient probe head (600 and 800 MHz) at 298 K. The NMR samples were measured in SEC buffer with 5% D_2O for the deuterium lock. The multidimensional experiments were recorded using apodization weighted sampling [177]. Backbone resonance assignment of ^{13}C, ^{15}N-labelled Hrp48 RRM1 was achieved to a completion of 100%, RRM2 of 94%, and for RRM12 to a completion of 95% in the free and of 79% in the 10-mer bound state (excluding prolines) using ^{1}H,^{15}N-HSQC, HNCO, HN(CA)CO, HNCA, CBCA(CO)NH, HNCACB triple resonance experiments [166]. All NMR spectra were processed using NMRPipe [178], analysed using CcpNmr Analysis [179-180], CARA (http://cara.nmr.ch), and Sparky [181]. Backbone torsion angles were predicted from CA and CB chemical shifts for RRM2 with TALOS-N [170].

4.2.5 NMR titration

NMR titrations were performed by recording two-dimensional ^{15}N,^{1}H-HSQC spectra of the labelled protein and stepwise addition of the titrant protein or RNA, followed by measurement of a HSQC spectrum at each step, until reaching saturation (i.e. the appearance of the spectra does not change further) or a 1:1 ratio. For NMR titration experiments, various protein concentrations were used: ^{15}N-labelled Hrp48 RRM1, RRM2, and RRM12 were titrated with different purchased RNA oligos (List of RNA nucleotides), protein–protein interactions were tested by titrating ^{15}N-labelled Hrp48 RRM12 with unlabeled Sxl dRBD3, Unr CSD123, CSD456, and CSD789, details described in Table 10.

CcpNmr Analysis and scripts used in our lab were used to follow chemical shift changes and determine dissociation constants (K_D). Individual K_D values of peaks exhibiting fast exchange [168] were averaged to calculate K_D values of the whole protein. Errors were calculated from the individual fitting errors by error propagation. In the intermediate and slow exchange regime, the data fitting is erroneous and for comparative reasons the qualitative appearance of the spectra was assessed. Chemical shift perturbations were calculated according to $CSP = \sqrt{\frac{1}{2}[\delta_H^2 + (\alpha \cdot \delta_N)^2]}$, where α was adjusted based on the peak positions of the spectra [168].

Table 10 Relative and absolute concentrations used for NMR titrations.

Analyte protein in NMR tube	**Concentration of protein (µM)**	**Titrants**	**Molar ratios**
Hrp48-RRM1	100	6-mer-1, 6-mer-2, 6-mer-3, 6-mer-4, 6-mer-5	1.0:0.0 – 1.0:2.5 or 1.0:3.0
Hrp48-RRM2	50	6-mer-1, 6-mer-2, 6-mer-3, 6-mer-4, 6-mer-5	1.0:0.0 – 1.0:2.5 or 1.0:3.0
Hrp48-RRM12	100,	10-mer, 14-mer, Sxl-dRBD3, Unr-CSD123, Unr-CSD456, Unr-CSD789,	1.0:0.0 – 1.0:2.5
	50	4-mers	1.0:0.0 – 1.0:1.0 or 1.0:1.5
Sxl-dRBD3	100	Hrp48-RRM12	1.0:0.0 – 1.0:2.5

Titration experiments with Hrp48 RRM2 were performed by Clara Hayn, a lab-rotation student I supervised. Chemical shift perturbations were plotted as B-factors on the structures using Pymol [182].

4.2.6 NMR relaxation

Measurements of R_1 longitudinal and R_2 transversal relaxation rates experiments for Hrp48 RRM12 and for the Hrp48 RRM12 – 10-mer RNA complex were acquired at proton Larmor frequencies of 700 MHz at 298 K using standard pulse sequences [175]. The sample concentration was 100 µM and the 10-mer RNA was in 1.5-fold excess. For RRM12 alone, only 2 points were measured because of the instability of RRM12 without RNA. For the R_1 experiment, relaxation delays were 20 and 700 ms; while for the R_2 experiment, they were 16 and 80 ms. For the RRM12 – 10-mer sample, the same experiments were recorded as a control. For better fitting of the datapoints, each experiment was also recorded with 13 relaxation delays. For this, the R_1 delays of 20, 50, 100, 150, 150, 250, 400, 500,

650, 800, 1000, 1300 and 1600 ms were used. For the R2 experiment, 32, 32, 64, 80, 96, 128, 160, 160, 192, 256, 320, 320, 384 ms relaxation delays were used. Peak integration, error estimation and exponential fitting for the relaxation experiments were done using PINT [183-184]. Calculation of the relaxation rates from two points were completed in MS Excel. The experimental rotational correlation times were calculated according to

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

where ν is the Larmor frequency in Hz, T_1 is the ^{15}N-longitudinal relaxation time, and T_2 is the ^{15}N-transversal relaxation time [185]. The errors were calculated by error propagation of the individual fitting errors.

4.2.7 Residual dipolar coupling measurements

Dipolar coupling experiments were used to obtain structural restraints supporting the modeling of the Hrp48 RRM12 – 10-mer complex structure. For this, first measurements under isotropic conditions were recorded, then in anisotropic conditions with the use of alignment medium. The sample of 0.1 mM ^{15}N-labeled Hrp48 RRM12 in complex with UAGGAUUAAG 10-mer was prepared in MES buffer. For the anisotropic conditions, the sample was complemented with Pf1 phage cosolvent (Asla Biotech) at 40 mg/ml final concentration. The deuterium splitting of this sample was 11.6 Hz at 298K. ^{15}N-IPAP-HSQC [186] experiments were acquired for the partially oriented and the isotropic samples to measure ^{15}N-^{1}H dipolar couplings. RDCs were determined as the difference of the coupling in the anisotropic and isotropic conditions. Module 2.0 [187] was used for analyzing and filtering out outlier RDC values for single domains of RRM12. The analysis was performed together with Bernd Simon with NmrPipe DC (General-Purpose Dipolar Coupling Analysis and Protein Backbone Chemical Shift Prediction) [178].

4.2.8 Structural modeling of RRM12 – 10-mer

The script used to generate the modeled structures was generated by Bernd Simon (NMR facility; EMBL Heidelberg). A model of the Hrp48 – RNA complex was generated by combining the X-ray structure of the hnRNP A1 (UP1) bound to DNA (2up1.pdb [101]) and the AF2 models of Hrp48-

RRM1 and RRM2. The AF2 models superimpose well to the hnRNP A1 X-ray structure (RMDS 0.726 RRM1 and 0.944 RRM2 respectively) and there are no clashes with the DNA. In the hnRNP A1 X-ray structure, the DNA binds hnRNP A1-RRM1 and RRM2 of two different protein molecules in the crystal, forming a 2:2 complex. To generate a monomer structure of Hrp48-RRM12 bound to DNA, the AF2 Hrp48-RRM1 to hnRNP A1-RRM1 + DNA and AF2 Hrp48-RRM2 to the corresponding symmetric molecule in the crystal were aligned. To generate a template file with Hrp48-RRM12 bound to RNA, we changed the DNA sequence to the correct RNA sequence keeping all identical atoms at their original position. The missing atoms and the broken linker bonds were generated by simulated annealing minimization of the model in cns-1.2 with all RNA atoms with known coordinates and the AF2 coordinates of Hrp48-RRM1 (residues 1-86) and Hrp48-RRM2 (residues 95-173) fixed [188-190].

4.2.9 Crystallization, data collection, and structure determination

Crystallization trials were performed using 3-lens 288-well crystallization plates using the sitting drop method, in which two sample concentrations/conditions can be tested with the same crystallization buffer. For this, 5-6 different 96-condition commercial screens were used at 4°C and at 20°C. Each well contained 0.1 µl sample and 0.1 µl crystallization condition solution, set up with a Mosquito LCP liquid handling robot by Brice Murciano at the Crystallization Facility of EMBL Heidelberg. Several conditions yielded crystals for Hrp48 RRM1. For the final crystal that gave the highest resolution in X-ray diffraction, RRM1 of 5 mg/mL concentration in the 200 mM NaCl, 30 mM NaPi, 2 mM DTT, pH 6.5. buffer was mixed in 1:1 ratio (100 nL:100 nL) with 0.2 M K_2SO_4 and 20% (w/v) PEG 3350 reservoir solution at 4°C. The sitting drop vapor diffusion method was used, and rod-shaped crystals started to nucleate overnight and kept growing further for 7-10 days. Crystals were soaked in mother liquor supplemented with 30% glycerol as a cryoprotectant prior to freezing.

Diffraction datasets were recorded at P-13 beamline at the German Electron Synchrotron (DESY), Hamburg, Germany. The crystals diffracted up to 1.15Å resolution.

The structure of RRM1 was solved by *ab initio* molecular replacement with the human hnRNP A1 RRM1 structure (pdb 1HA1, 40% sequence identity) as model using Phenix Phaser-MR [191]. The initial model was built using Phenix AutoBuild [192] and manual adjustments were executed with Coot [193]. The structure was further improved in iterative rounds of manual correction with Coot and restrained refinement with phenix.refine [194].

4.2.10 MD modeling of RRM1 – 6-mer-3

The MD modeling was performed by Miroslav Krepl from Jiří Šponer research group (Institute of Biophysics of the Czech Academy of Sciences, Brno).

To simulate the spontaneous binding process of the 5′-UUAAG-3′ motif of the 6-mer-3 RNA (AUUAAG) to the RRM1 domain, we have constructed a system with the RNA positioned ~20 Å away from the protein. The initial coordinates of the protein and RNA were obtained from the X-ray structure of the RRM1 domain and by NAB (Nucleic Acid Builder), respectively [195]. The initial conformation of the 6-mer-3 RNA corresponded to an A-RNA helix with the complementary strand removed. We have used the LEaP module of AMBER 20 [195] to prepare the topology and coordinate files. The OL3 [196] and ff12SB [197] force fields were used to describe the RNA and protein, respectively. To allow the spontaneous binding of the single-stranded RNA, we have applied the stafix potential (factor 0.5) to weaken the excessive RNA-RNA interactions [198]. The RNA and the protein were immersed in an octahedral box of SPC/E water molecules [199] with minimal distance of 14 Å between the solutes and the box border. We have added the KCl ions [200] to neutralize the systems and to obtain ion concentration of ~0.15 M. Prior to the production simulations, the systems were minimized and equilibrated [201]. The production simulations were then performed in constant pressure ensemble. Monte-Carlo barostat and Langevin thermostat were used to control the pressure and temperature, respectively [195]. We have performed six 10-μs-long independent MD simulations, with different trajectories obtained by utilizing random seed numbers. The resulting trajectories visualized the RNA at different stages of binding to the RRM1. We have subsequently manually selected a structure with the RNA stably bound at a location close to all the protein residues which exhibited chemical shift changes in NMR experiments, and used this structure as our working model for the RRM1/6-mer-3 protein-RNA complex structure. Long-term stability of this structure was subsequently verified in two independent 10-μs-long MD simulations.

4.2.11 Isothermal titration calorimetry

ITC measurements were performed using a Malvern MicroCal PEAQ-ITC calorimeter at 20°C. For all experiments, the protein was loaded in the cell and RNA or protein–RNA complex in the syringe (see Table 11). The samples were dialyzed against MES buffer (20 mM MES, 200 mM NaCl, 0.02% NaN3, pH 6.5) buffer overnight, adjusted to the appropriate concentrations, centrifuged at 15000 r.p.m. for 10 minutes, transferred to the final tubes for measurement, and degassed for 5 minutes. The titrations were accomplished with 13 – 25 injections corresponding to 1.5–3 μL injection volumes, in

order to achieve optimal sampling with regard to the enthalpy change. All experiments were initialized with an 0.4 µL injection followed by the standard injections, each of them lasting 3 seconds with 150 second delays in between. The sample stirring was set to 750 rpm, instrument feedback to high, and the reference power was 10 µcal/s. Further details about the concentrations and experimental setup in individual titrations are listed in Table 11.

For data evaluation, base-line corrections and non-linear curve fitting the provided Malvern MicroCal PEAQ-ITC analysis software was used.

Table 11 Detailed parameters and results of ITC measurements.

Exp. name	Replicates	Conc. syringe (μM)	Conc. cell (μM)	*N* (sites)	*N* (error)	K_D(μM)	K_D (μM, error)	Δ*H* (kcal/mol)	–*T*Δ*S* (kcal/mol)	Injection no.	Injection vol. (μl)	Shown in figure
Hrp48-RRM1 + 6-mer-1	2	6-mer-1 766–800	RRM1 30.0	2.4, 3.1	$45 \cdot 10^{-3}$, $121 \cdot 10^{-3}$,	1.46, 3.48	0.40, 1.23	-7.2±0.3, -5.3±0.4	-25.6 -25.4	19	2.0	Figure 14
Hrp48-RRM1 + 6-mer-2	3	6-mer-2 189–550	RRM1 17.5–20.0	0.8, 0.8, 0.9	$35 \cdot 10^{-3}$, $76 \cdot 10^{-3}$, $94 \cdot 10^{-3}$	8.68, 3.99, 7.38	1.25, 2.18, 2.87	-76±6, -65±12 -85±16	48 35 56	19/25	2.0/1.5	Figure 14
Hrp48-RRM1 + 6-mer-3	4	6-mer-3 235–601	RRM1 23.0–25.0	0.8, 0.6, 0.4, 0.7	$17 \cdot 10^{-3}$, $18 \cdot 10^{-3}$, $13 \cdot 10^{-3}$, $17 \cdot 10^{-3}$	5.04, 6.08, 2.61, 3.38	0.63, 0.50, 0.37, 0.40	-105±5, -100±5 -107±6 -74±3	75 71 76 43	19/25	2.0/1.5	Figure 14
Hrp48-RRM1 + 6-mer-4	3	6-mer-4 650–1400	RRM1 17.5–30.0	1.7, 1.6, 2.4	$32 \cdot 10^{-3}$, $12 \cdot 10^{-3}$, $37 \cdot 10^{-3}$	1.01, 0.78, 3.03	0.26, 0.09, 0.49	-9.7±0.3, -10.4±0.1 -7.3±0.2	-24.0 -23.9 -23.7	19/25	2.0/1.5	Figure 14
Hrp48-RRM12 + 8-mer	1	Region 5 1000	RRM12 60.0	No binding detectable by ITC.						19	2.0	
Hrp48-RRM12 + 10-mer	3	10-mer 200	RRM12 15.3–20.0	0.9, 0.9, 0.9	$4.9 \cdot 10^{-3}$, $6.1 \cdot 10^{-3}$, $5.4 \cdot 10^{-3}$	0.27, 0.32, 0.34	$21 \cdot 10^{-3}$, $31 \cdot 10^{-3}$, $27 \cdot 10^{-3}$	-173±2, -165±2, -169±2	136, 129, 133	25	1.5	Figure 21
Hrp48-RRM12 + GG-30-mer	3	GG-30-mer	RRM12	Complex binding mode of multiple binding sites.						19/25	2.0/1.5	Figure 30
Hrp48-RRM12 + GG-30-mer–Sxl-dRBD4–Unr-CSD12 complex	4	complex 283–300	RRM12 19.0–29.3	0.5, 0.6, 0.5	$18 \cdot 10^{-3}$, $14 \cdot 10^{-3}$, $9.7 \cdot 10^{-3}$	0.85, 1.80, 1.91	0.22, 0.31, 0.25	-124±7, -126±6, -120±4	90, 94, 128	19/25	2.0/1.5	Figure 30

4.2.12 Fluorescent labeling of RNA

The GG-30-mer RNA oligo used for EMSAs was 3′ end-labeled with pCp-Cy-5 (Cyanine 5). For this, the following reaction mixture was combined in 20 µL final volume: 100 pmol RNA, 200 pmol pCp-Cy-5, 2 µL T4 RNA ligase (10x, 20 U), 2 µL DMSO, 1 mM ATP, 1 mM DTT, 10 mM $MgCl_2$, 50 mM Tris-HCl (pH 7.5). The reaction was incubated at 16 °C overnight.

Subsequently the reaction was cleaned up by NaOAc/EtOH precipitation. To the reaction mixture, 4 µL (0.2 × volume) of 3M sodium-acetate, 1.5 µL of GlycoBlue Coprecipitant (Invitrogen), and 59 µL (2.5 × volume) of ethanol cooled to -20°C was added and mixed well by vortexing. The RNA was precipitated overnight at -70 °C, centrifuged at 4 °C and 13000 r.p.m. for 30 minutes, washed two times with -20°C 70% ethanol and once with 100% ethanol.

The final amount and the labelling efficiency were measured by spectrophotometry using the NanoDrop™ OneC.

4.2.13 Electrophoretic mobility shift assays

Electrophoretic mobility shift assays (EMSAs) were used to determine RNA-binding affinities in a semi-quantitative mode [202]. Recombinantly purified Hrp48 RRM12, Sxl dRBD3, or/and Unr CSD12 were mixed with 20 nM Cy-5 labeled GG-30-mer probe in a buffer consisting of 50 mM Tris-HCl (pH 8.4), 100 mM NaCl, 1 mM EDTA, 250 mg/ml BSA, 10% glycerol, 0.05% Triton X-100 and 1 mM DTT in 12.5 µL reactions and incubated on ice for 30 min, protected from light. The concentration of the proteins is indicated on the figure. The RNA – protein complexes were resolved on a 6% native 1 × TBE polyacrylamide gel for 35 min at 200 V. The gels were imaged at a Typhoon Trio Imager 9000 (GE Healthcare).

4.2.14 *In-vitro* translation assays

The *in-vitro* translation assays were performed by Tanit Guitart (Fátima Gebauer Laboratory, CRG Barcelona), except for the mutations of the plasmids that were carried out by me.

Plasmids

Wild type construct BLEF, as previously described [65], is composed of 69 nt of *msl-2* 5′ UTR sequence including site B and 46 nt of the *msl-2* 3′ UTR including sites E and F. For introducing mutations in the 3' UTR, primers were created by QuikChange Primer Design (https://www.agilent.com/) and used for restriction-free cloning to create changes in the Region 5 and Region 6. The mutant constructs are described in Chapter 5.4.

Recombinant protein expression and purification

Sxl dRBD4 (amino acids 122–301 of *Drosophila melanogaster* SXL) was expressed in *Escherichia coli* as an N-terminal GST-tagged fusion protein and purified as described [63]. The protein was dialyzed against buffer D (20 mM HEPES pH 8.0, 20% glycerol,1 mM DTT, 0.01% NP-40, 0.2 mM EDTA).

RNA *in vitro* transcription

BLEF mRNA derivatives were synthesized using T3 RNA polymerase (Ambion) and contained a 5′ m^7GpppG cap and a poly(A) tail of 73 residues. mRNAs were purified by phenol-chloroform extraction and G_{50} columns (GE Heathcare). All mRNAs used in the same experiment were synthesized and quantified in parallel, and the concentration and quality confirmed by separation in agarose gels.

In vitro translation assays

In vitro translation assays were performed as described [62]. Briefly, 17 ng of *msl-2* Firefly reporter mRNA and 10 ng Renilla luciferase mRNA, used as an internal control, were incubated with increasing amounts of GST-dRBD4 in a final volume of 10 µL. The reaction contained 40% *Drosophila* embryo extract, 60 µM amino acids, 16.8 mM creatine phosphate,

80 ng/µL creatine kinase, 24 mM HEPES pH7.5, 0.6 mM $Mg(OAc)_2$ and 80 mM KOAc. The reaction was incubated at 25°C for 90 min, and the Firefly and Renilla activities were measured using the Dual Luciferase kit (Promega).

4.2.15 Data presentation

Data graphs were plotted and fitted using GraphPad Prism 5.03 and the appearance adjusted in Inkscape 1.0.1 to the final form. Visualization of structures as molecular images was performed with PyMol 2.5.1. Figures were prepared in Inkscape 1.0.1.

5 Results

In structural biology, it is common practice to embark on projects with the divide-and-conquer approach, and study systems by first focusing on smaller elements that can be studied in great detail, and gradually putting together the parts to understand the system as a whole. Employing this idea, I started my thesis work by characterizing the single domains of Hrp48 before studying the tandem domain construct. In parallel, I also investigated the assembly and of the complex of Hrp48, Sxl, Unr, and *msl-2*.

5.1 Structure and RNA binding studies of Hrp48-RRM1

Initially, I optimized constructs for expression and solubility for RRM1, RRM2 and RRM12. To this end different boundaries were tested based on predicted domain arrangements and sequence alignments. The constructs were screened based on solubility and stability and tested by 2D-^{1}H,^{15}N-HSQC NMR experiments, where constructs with peak dispersion were chosen, indicating a properly folded protein. All the optimized constructs embrace extensions compared to the predicted boundaries and are listed in Figure 13 A.

Structure determination of Hrp48-RRM1 was completed successfully by X-ray crystallography. I was able to crystallize Hrp48-RRM1 and the resulting crystals diffracted to 1.15 Å resolution, which allowed me to solve the crystal structure using molecular replacement based on the hnRNP A1-RRM1 structure (PDB: 1HA1 [203]). The RRM1 domains of Hrp48 and hnRNP A1 share a 62% sequence similarity and 43% sequence identity (Supplementary Figure 1). The structure of RRM1 follows the canonical RRM fold with four β-strands packed against two α-helices (Figure 13 B, C, D). The RNP motifs consist of the canonical amino acids, except of RNP-2 where at the fifth position, the canonical asparagine is replaced by a glycine (G13) (Figure 13 E), a feature also present in hnRNP A1 (Supplementary Figure 1 and Supplementary Figure 2). In addition, loop 1 carries an aromatic residue (W16), which is present in 25% of human RRMs and is usually involved in RNA-binding [81].

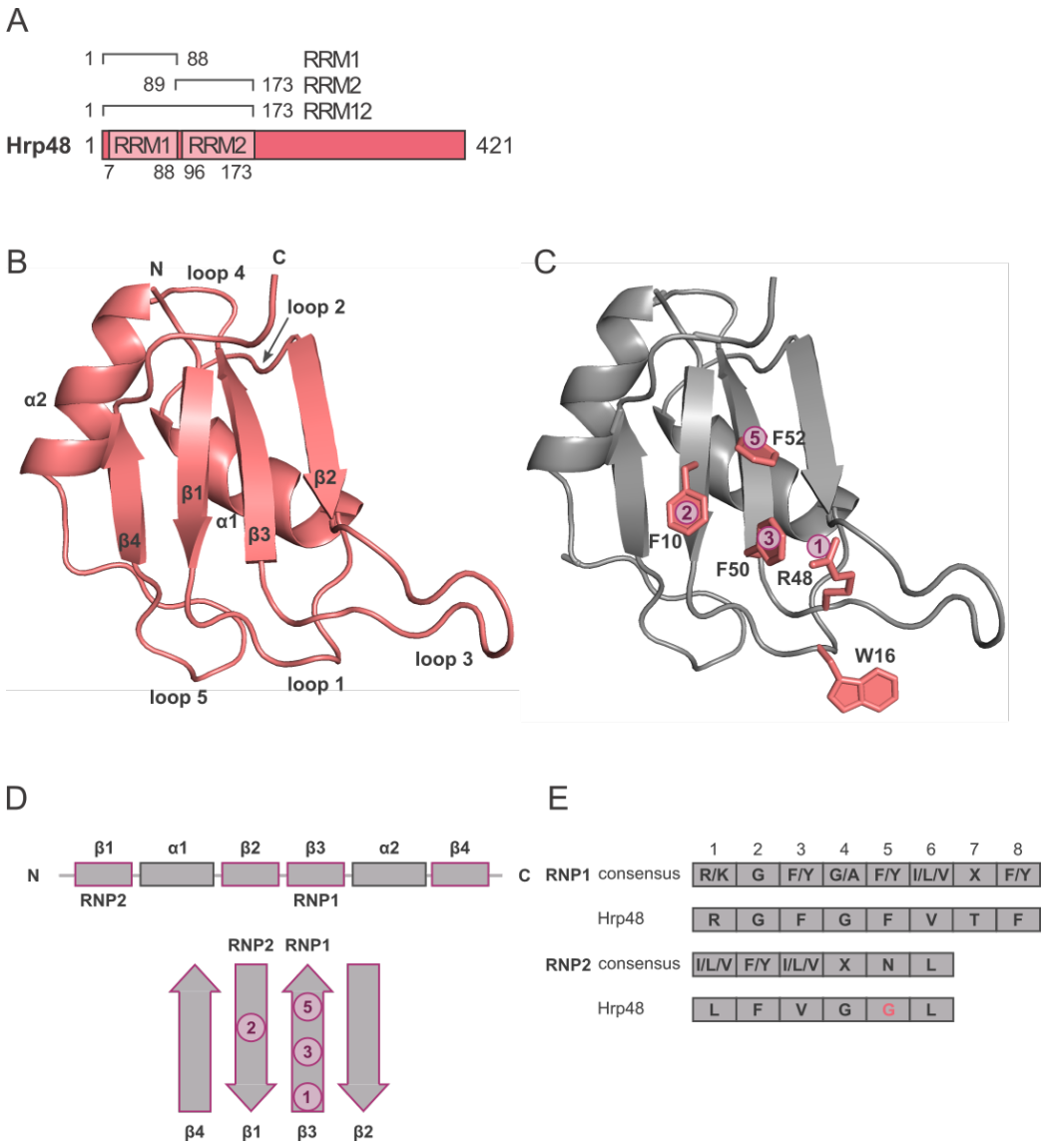


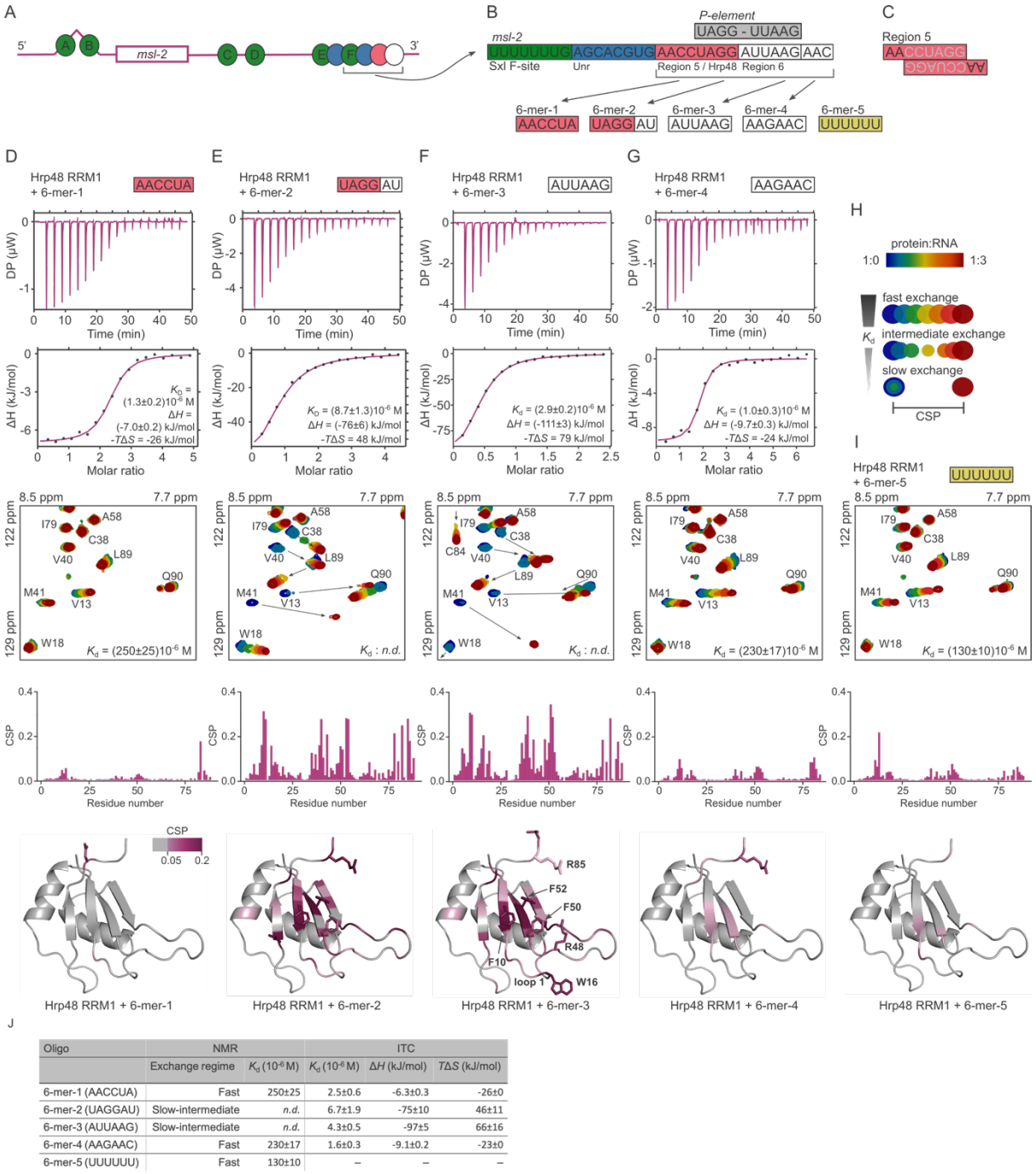
Figure 13 Crystal structure of Hrp48-RRM1

A: The optimized domain boundaries of Hrp48. **B-C:** Crystal structure of RRM1 with secondary structure elements (**B**) and highlighted RNA-binding residues (C). **D:** Domain arrangement of Hrp48-RRM1. E: RNP consensus sequences compared to Hrp48-RRM1 RNPs.

In order to characterize the RNA-binding properties of Hrp48-RRM1, I used isothermal titration calorimetry (ITC) and ^{1}H, ^{15}N-HSQC NMR titration experiments to measure chemical shift perturbations (CSP) of protein resonances upon titration with RNA. The previously reported *msl*-2 putative binding site of Hrp48 is directly downstream of the Sxl F and Unr binding sites and has been labelled as Region 5 (Figure 14 A and B) [72]. The last four nucleotides of Region 5 and the downstream Region 6 together (Figure 14 B) are almost identical to the Hrp48 binding site on the P-element transposase mRNA (UAGGAUUAAG

(*msl-2*) and UAGGUUAAG (*P-element*), Figure 14 B). To this end, I divided the 17 nucleotides downstream to the Unr binding site to 6-mer RNA oligomers to test binding on the level of single RRM domains (Figure 14 B). This resulted in four 6-mers with the following sequences: 6-mer-1: AACCUA, 6-mer-2: UAGGAU, 6-mer-3: AUUAAG and 6-mer-4: AAGAAC. A reference oligomer (UUUUUU, U6-mer or 6-mer-5) that has an unrelated sequence to the tested 17-nucleotide long region was also added as a control (Figure 14 B). In addition, formation of RNA duplexes was also avoided, because Region 5 encompasses a palindromic motif (Figure 14 C), that is not present in any of the 6-mers. This RNA self-association was confirmed by one-dimensional ^{1}H-NMR experiment with an 8-mer construct where the palindrome is present (Supplementary Figure 3).

Initially, I tested the RNA binding of RRM1 by ITC for each 6-mer. The binding of the control U6-mer (6-mer-5) was not strong enough to be detected by ITC, demonstrating that Hrp48-RRM1 binds RNA with base specificity. From ITC dissociation constants (K_D) the optimal RNA motif for RRM1 could not be determined, because they were on the same scale for all four 6-mers, thus did not clearly suggest which of the four 6-mers is the optimal motif for RRM1 binding (Figure 14 D - G). However, there was a large difference in the enthalpic and entropic contributions to the affinity of binding of 6-mer-1 and 4 compared to 6-mer-2 and 3 (Figure 14 D – G and J). This could be a result of different binding modes and could hence be an indication of whether one of the motifs would be better suited for structure determination or give insights into which of the motifs would be bound by RRM1 in the context of a longer RNA and RRM12 binding.



J

Oligo	NMR		ITC		
	Exchange regime	K_d (10^{-6} M)	K_d (10^{-6} M)	ΔH (kJ/mol)	$T\Delta S$ (kJ/mol)
6-mer-1 (AACCUA)	Fast	250±25	2.5±0.6	-6.3±0.3	-26±0
6-mer-2 (UAGGAU)	Slow-intermediate	*n.d.*	6.7±1.9	-75±10	46±11
6-mer-3 (AUUAAG)	Slow-intermediate	*n.d.*	4.3±0.5	-97±5	66±16
6-mer-4 (AAGAAC)	Fast	230±17	1.6±0.3	-9.1±0.2	-23±0
6-mer-5 (UUUUUU)	Fast	130±10	–	–	–

Figure 14 RNA-binding of Hrp48-RRM1

A: Schematic model of the full-length *msl-2* mRNA with reported binding sites for Sxl (green), Unr (blue) and Hrp48 (pink). The white circle marks the region we studied in addition to the earlier identified binding sites. **B:** Zoomed in region of the 3' UTR. The Region 5 is the sequence suggested previously as the *msl-2* Hrp48 binding site [72]. The Hrp48 *P-element transposase* binding site shows high similarity to Region 5 and Region 6 of *msl-2.* The 6-mer constructs were used in this study to refine the Hrp48 binding site. 6-mer-5 was used as a control. **C**: Region 5 RNA oligomer can self-associate through duplex formation in isolation. **D – G, I:** Biophysical characterization of RNA binding of RRM1 by the 6-mers. **Top (D – G)**: isothermal titration calorimetry (ITC) data of Hrp48-RRM1 titrated with 6-mer-1, 2, 3 and 4. **Middle**: (**D – G, I**): a zoomed-in region of the ^{1}H, ^{15}N-HSQC NMR titration experiments of ^{15}N-labeled Hrp48-RRM1 all the 6-mers. **Middle**: CSP plots of the NMR titrations. **Bottom**: mapping of the CSP data of the titrations on the RRM1 crystal structure. The scaling for the coloring is the same for each titration. **H:** Illustration of the effect of the exchange rate on the appearance of two-

dimensional peaks. The exchange regimes indicate the relative binding affinity: the slower exchange is the tighter the binding. The chemical shift perturbation (CSP) is the weighed distance of the two-dimensional peaks. **J**: Table of NMR and ITC results of interaction studies between RRM1 and the 6-mers. The determination of the K_D by fitting the NMR data in the slow-intermediate exchange regime was not straightforward. Averaged ITC data was derived from replicate experiments, the complete data is shown in Table 11. Errors were calculated with error propagation except for $T\Delta S$, where the standard deviation of the averaged values is shown, because of the lack of experimental error. 6-mer-5 binding was not strong enough to allow reliable fitting of ITC data.

In order to complement the ITC results and to obtain further insights into RNA binding specificity of Hrp48-RRM1, NMR titration experiments were acquired. A 2D-^{1}H,^{15}N-HSQC spectrum was recorded for each titration step (gradual titration with RNA) to observe chemical shift perturbations. The resonances of residues involved in the interaction with RNA exhibit stronger CSPs compared to the ones not engaged in RNA binding. (Nevertheless, allosteric effects upon interaction can also bring about strong CSPs on residues that are not in close proximity to the RNA.) According to the CSPs, RRM1 binds to all five 6-mers, but with different affinities (Figure 14 D – G, I). For the titrations with 6-mer-1, 6-mer-4 and U6-mer the peaks of RRM1 move according to the fast exchange regime on the NMR time scale, whereas for 6-mer-2 and 6-mer-3 the peaks exhibit strong CSPs and the intensities also change, characteristic for the intermediate-to-slow exchange regime (Figure 11 and Figure 14 H). CSPs of the binding events in the fast exchange regime can be fitted to derive the dissociation constant, which are considerably higher than the ones determined by ITC. The intermediate-to-slow exchange regime does not allow for reliable determination of the K_D values because of problematic data quantification and curve fitting. However, the fast exchange indicates weaker binding than the slow exchange regime, that usually accounts for sub micromolar affinities [168]. Thus, from the NMR data, one could qualitatively assume that 6-mer-2 and 6-mer-3 are bound stronger by Hrp48 than 6-mer-1, 6-mer-4 and 6-mer-5, with 6-mer-3 featuring the highest number of residues in the slow exchange. Although not entirely consistent with ITC data, we concluded from the NMR titrations that 6-mer-3 is the best Hrp48-RRM1 binder due to strongest CSPs over most residues (Figure 14 F and J).

To visualize the RNA-binding interface and the differences between the RRM1 – 6-mer interactions, I mapped the CSP values on the crystal structure (Figure 14 D – G, I). For this, I completed the backbone assignment of RRM1 to identify the peaks on the 2D-^{1}H,^{15}N-HSQC spectra. The RNA-bound spectra of Hrp48-RRM1 in complex with 6-mer-2 and 6-mer-3 has to be re-assigned with triple-resonance experiments and doubly labelled samples, because the chemical shifts of the peaks in the slow exchange regime cannot be followed simply by

comparing the titration spectra. The CSP mapping shows for the 6-mer-3 and RRM1 interaction, that the RNA binds along the four β-strands, resembling the canonical RNA binding mode of RRM domains (Figure 14 F). All canonical RNA-binding amino acids (F10 on RNP-2 and R48, F50, F52 on RNP-1, Figure 14F) are involved in the interaction. In addition, the W16 on loop1 and residues in the N-terminal region (R85) exhibit strong CSPs, suggesting a role in RNA-binding.

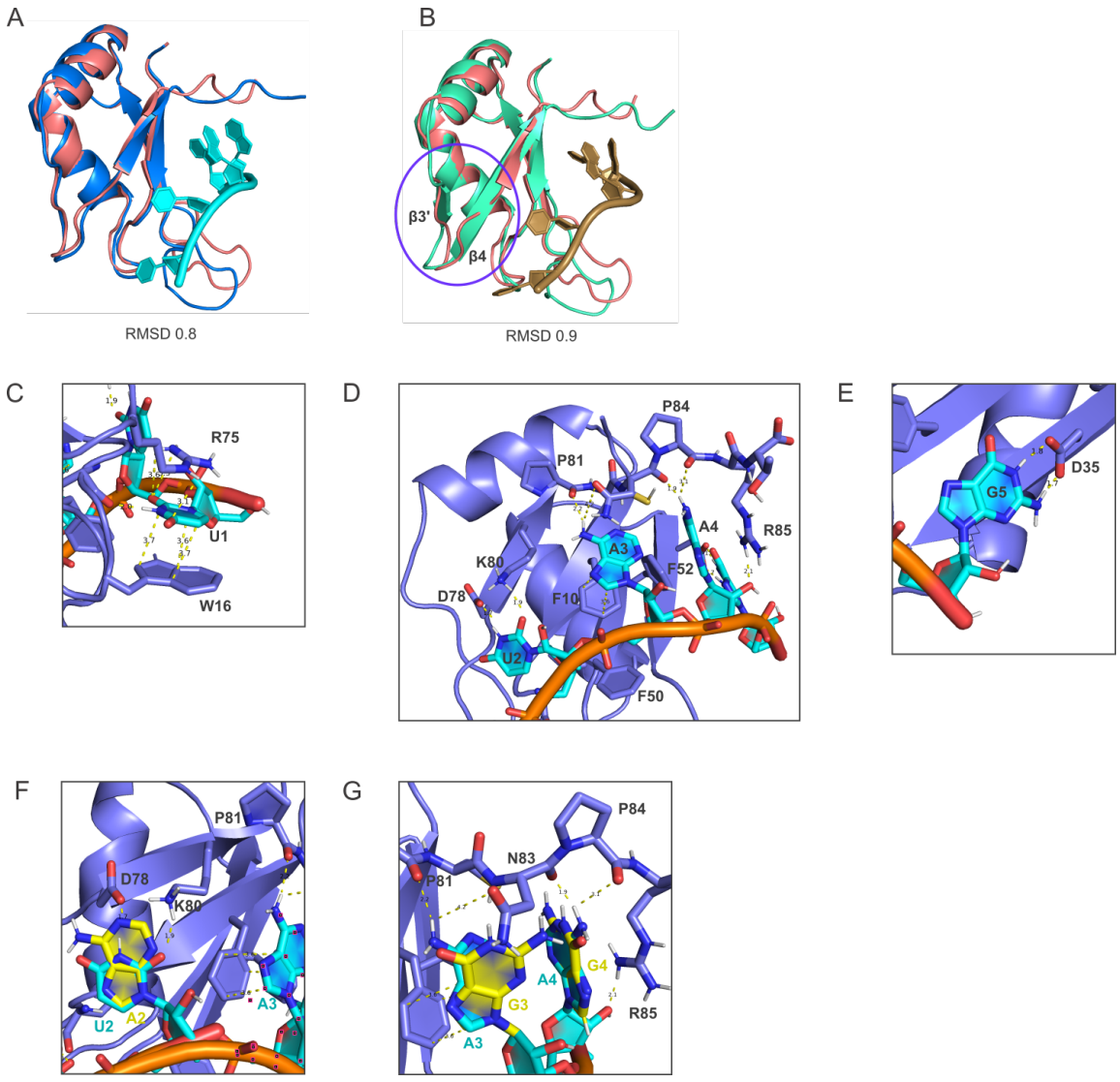


Figure 15 MD simulation of Hrp48-RRM1 – 6-mer-3 binding

A and B: two different frames of the MD simulations, in good agreement with the crystal structure (RMSD of 0.8 and 0.9 respectively). In some of the frames (B), a β3'-strand forms and β4 is also elongated in the N-terminal direction. **C-E:** canonical RRM – RNA contacts in the complex. C: U1 is sandwiched between W16 and R75. D: P81 and P84 restricts the C-terminal in a bent conformation, and R85 fixes this with a direct contact to A4 sugar ring. Along this bend several protein – RNA contacts form. E: G5 is recognized by D35 on the β2-strand. **F-G:** Mutagenesis of the nucleotides

specifically recognized *in silico.* WT: cyan, mutant: yellow. F: D87 and K80 recognizes U2, but an A2 mutagenesis shows clashes with these amino acids. G: At position 3 adenine is recognized and a guanin could not be and acceptor of the hydrogen bonds with P81 backbone and N83. The C-G panels show depictions of the MD frame presented on A.

Next, I aimed to solve the structure of Hrp48-RRM1 bound to 6-mer-3. For this I tried co-crystallization of RRM1 with the RNA. However, the crystals did not contain RNA, but only the protein. NMR structure determination was also unsuccessful, because of the lack of detectable intermolecular NOE peaks that would provide structural restraints for structure calculation. Our collaborator, Miroslav Krepl in the Jiří Šponer research group (Institute of Biophysics of the Czech Academy of Sciences, Brno) has developed a method to simulate structures of RRM domains binding to single-stranded RNA, for which goal-specific force-field modifications were developed [204]. To obtain an MD model of the RRM1 – 6-mer-3 complex, the crystal structure of RRM1 and the UUAAG motif of 6-mer-3 was used as input. This shortening of the 6-mer-3 by one nucleotide was necessary for the simulation, which is optimized for 4-mer or 5-mer motifs. Initially six independent MD simulations were performed for 10 μs obtaining different trajectories. The resulting trajectories visualized the RNA at different stages of binding to the RRM1 and we selected manually one trajectory based on previous knowledge on RRM domain binding modes and my CSP data (residues exhibiting at least 0.1 ppm CSP were considered). The resulting models might not mirror the correct binding register for RRM1 – 6-mer-3 recognition, but can give an insight into a possible binding mode. From the chosen trajectory of 10^6 frames every 1000^{th} frame was selected to study (this adds up to 1000 selected frames altogether). From around the 600^{th} selected frame both the protein and the RNA conformation fluctuates around similar coordinates. Subsequently, two structures were selected with the RNA stably bound at a location close to all the protein residues which exhibited chemical shift changes in NMR experiments (Figure 15 A and B). Most frames towards the end of the simulation possess a similar secondary structure arrangement as depicted in Figure 15 A which is identical to the crystal structure of RRM1. In some of the frames however, a β3'-strand forms and parallel to this, β4 gets elongated (Figure 15 B). Based on the analysis of the protein – RNA contacts this does not have a significant implication on the binding mode. I chose the frame without the non-canonical β3'-strand (on Figure 15 A) to investigate the RNA binding. The first uridine of the UUAAG motif (U1) is sandwiched between residues W16 and R75 with unspecific contacts (Figure 15 C). Also, W16 remains flexible during the MD simulation based on comparing several frames (data not shown). Nevertheless, this residue exhibits very strong CSPs upon binding with the 10-mer (Figure 14

F). F10, F50 and F52 form the canonical, non-sequence-specific RRM – RNA contacts with A3 and A4. Amino acid side chains and backbone amide groups of the β4-strand and the C-terminal extension form several contacts with the U2, A3 and A4 nucleotides (Figure 15D). This C-terminal extension is rich in prolines (P81 and P82) which arranges the conformation in a loop and the direct contact between R85 and the sugar ring of A4 fixes this loop to the RNA. G5 does not interact with this back-folded peptide chain but establishes hydrogen bonds with D35 (Figure 15 E). To gain insight into the sequence specificity and the correct binding register, I mutated some of the nucleotides to the neighboring residues in Pymol (Figure 15 F and G, mutated nucleotides are colored yellow). For example, U2 was mutated to A, as the third nucleotide is an adenine (Figure 15 F). The small pocket formed by D78 and K80 cannot accommodate a purine base and therefore this mutation would cause a clash. In addition, G13, that is located beneath the aromatic ring of U2 shows strong CSPs not just when titrated with 6-mer-3 but also with U6-mer (6-mer-5, greatest shift on the CSP plots on Figure 14 I), implying U-specific recognition of this part of the RRM due to steric clashes with other bases. It is also possible that the following A3 is specifically recognized. A mutation of A3 to a guanine would abolish the hydrogen bonds forming between P81, N83 and the hydrogen-bond donor site of the purine ring (Figure 15 G). A pyrimidine mutation would not suffice the distance between the RNA backbone and the binding pocket, suggesting specific recognition of an adenosine at this position. This in silico analysis together with the NMR titration suggests that the binding register is correct in the MD simulation.

5.2 Biophysical and RNA binding characterization of Hrp48-RRM2

Structure determination of RRM2 by X-ray crystallography and NMR spectroscopy was attempted with no success. Due to the self-interacting nature of RRM2, only low (100 µM) protein concentrations could be used for NMR spectroscopy backbone experiments that makes the assignment very demanding and challenging. I completed the backbone chemical shift assignment for RRM2 to 94% for the non-proline residues. For structure determination by NMR, further experiments are needed that cannot be performed at this concentration.

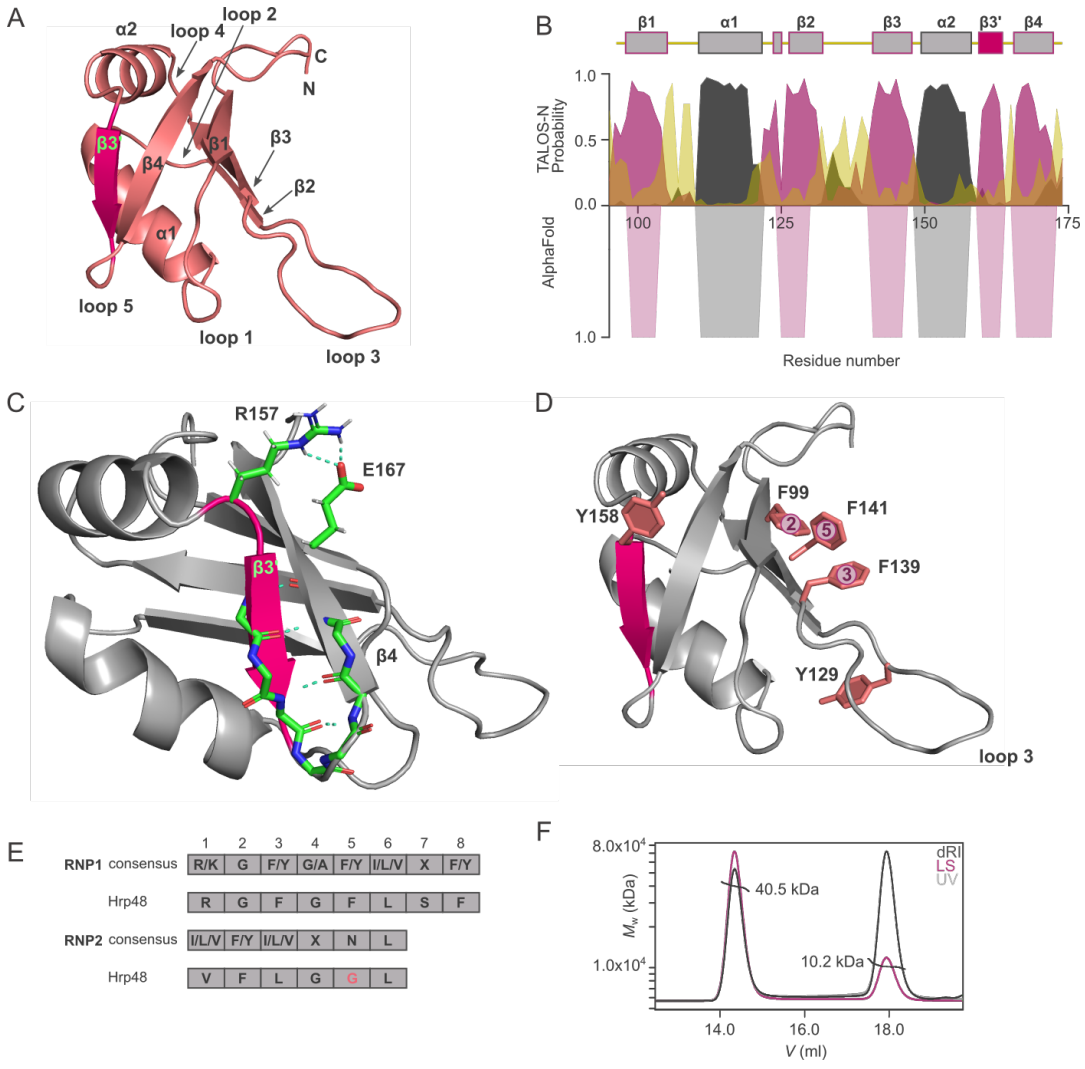


Figure 16 Structural model and domain arrangement of Hrp48-RRM2

A: AlphaFold2 structural model of RRM2 with secondary structure elements, β3' highlighted in pink. **B**: The β – α – β – β – α – β – β fold prediction of AF is compared to experimental NMR data analyzed by TALOS+ (Prediction of Protein Backbone Torsion Angles from NMR Chemical Shifts) [170] α-helix: grey, β-strand: purple, loops: yellow. **C:** The non-canonical β3'-strand is stabilized by a salt bridge formed between R157 and E167 and sever H-bonds between β3' and β4. **D:** Highlighted RNA-binding residues of RRM2 and additional aromatic residues. **E:** RNP consensus sequences compared to Hrp48-

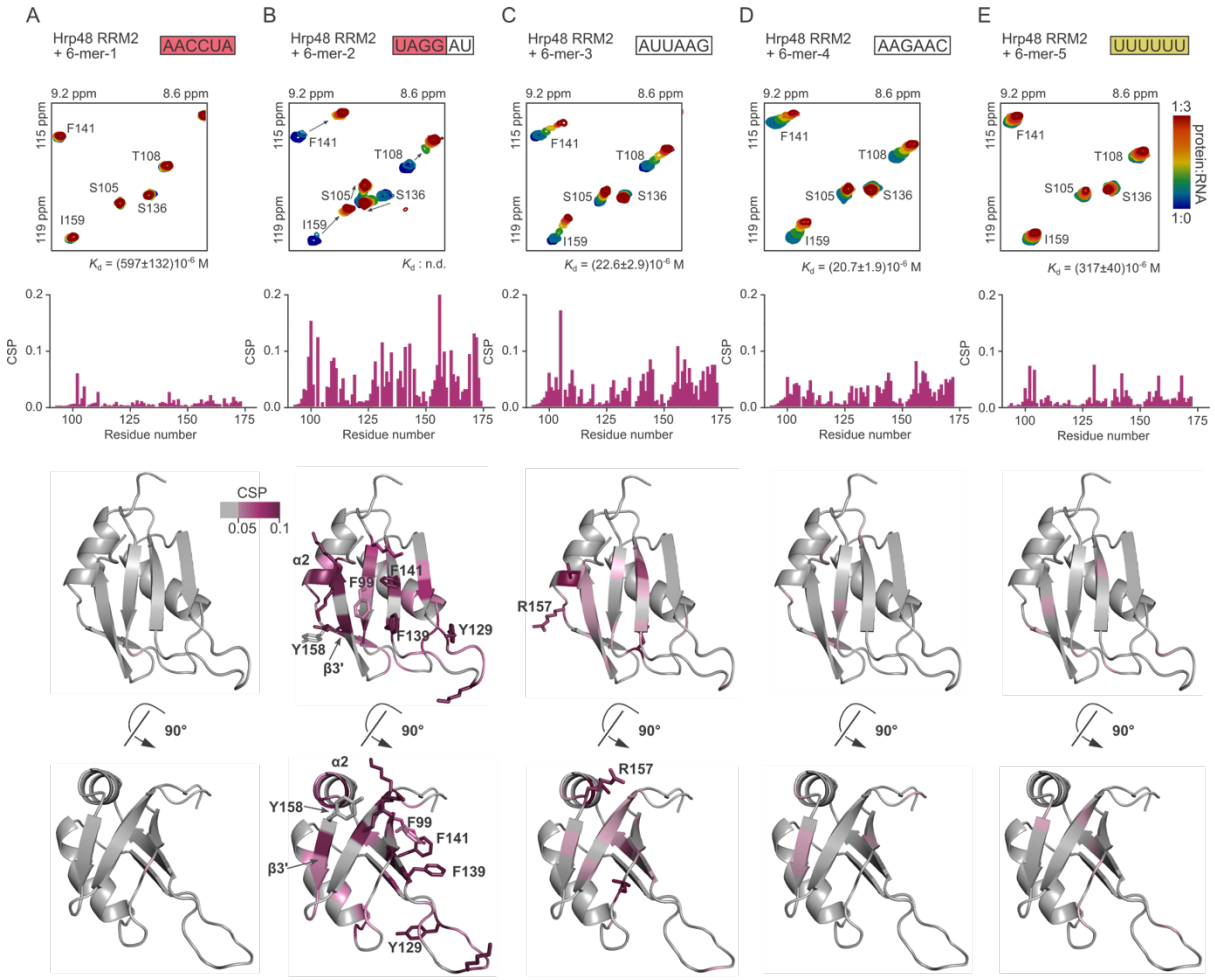
RRM2 RNPs. **F:** SEC-MALLS of RRM2 reveals tetramerization tendency of RRM2. Dark grey: differential refractive index (dRI), purple: light scattering (LS), light grey: UV absorption.

Therefore, I used AlphaFold2 to predict the structure of RRM2 as this method has been shown to achieve experimental accuracy [122]. Interestingly, the predicted structure revealed an extra β-strand between the last α-helix and β-strand, deviating from the canonical β – α – β – β – α – β fold of RRMs (Figure 16 A). To validate this model, I predicted the dihedral angles by TALOS+ based on my NMR chemical shift assignment of RRM2 [170]. TALOS+ (Torsion Angle Likelihood Obtained from Shift and Sequence Similarity) is a database system which empirically predicts protein dihedral torsion angles from the protein sequence and six different types of chemical shift values (HN, H_α, C_α, C_β, CO and N). The prediction is based on known structures with assigned chemical shifts, and the observation that the typical secondary chemical shifts for different residues (the deviation from the random coil average chemical shifts) correlate with the secondary structure [170]. The TALOS+ prediction of Hrp48-RRM2 supported the presence of the additional β-strand (β3') in solution (Figure 16 B), and indirectly the validity of the AlphaFold2 model.

The β3'-strand formation is stabilized by the formation of a salt bridge between R157 and E167 (Figure 16 C) and additional hydrogen bonds between β3' and β4, which also elongates the β4-strand in the N-terminal direction. The β3'-strand provides a platform for Y158 and also loop 3 contains an aromatic residue (Y129), which could be involved in RNA binding apart from the canonical F99, F139 and F141 in the RNP-motifs (Figure 16 D). The RNP-motifs of RRM2 adhere to the canonical sequences with the same exception as for RRM1: residue Nr. 5 of RNP-2 is a glycine instead of an asparagine (Figure 16 E).

During purification and analysis of NMR data, I observed that RRM2 tends to oligomerize. Size-exclusion chromatography coupled with multi-angle laser light scattering (SEC–MALLS) confirmed that this isolated domain forms tetramers in an isolated *in vitro* context at 10-20 µM and above (Figure 16 F). Even at these low concentrations, no interpretable ITC curve was possible to measure, presumably because of the self-association. Therefore, NMR titration experiments were utilized in the same way as for RRM1 to assess the RNA binding specificity and affinity of RRM2 (Figure 17 A-E, complete spectrum series presented on Supplementary Figure 5). The NMR titration experiments were performed together with my internship student, Clara Hayn. Upon titration with 6-mer-1 and control RNA 6-mer-5 no or only weak CSPs could be observed. Titration with 6-mer-3 and 6-mer-4 resulted in stronger CSPs and the peaks showed the pattern of fast and fast-to-intermediate exchange, whereas 6-mer-2 titration induced

the largest CSPs and chemical exchange in the intermediate-to-slow regime (Figure 17 A-E Top and Middle). Fitting the CSP data to obtain K_D values and qualitative assessment based on the CSP patterns, 6-mer-2 is the optimal binding sequence for RRM2. For 6-mer-2 I did not consider the fitted K_D because a nonlinear regression of the mixed slow-intermediate exchange data is not reliable.



F

Oligo	NMR	
	Exchange regime	K_d (10^{-6} M)
6-mer-1 (AACCUA)	Fast	597±132
6-mer-2 (UAGGAU)	Slow-intermediate	*n.d.*
6-mer-3 (AUUAAG)	Intermediate	22.6±2.9
6-mer-4 (AAGAAC)	Intermediate	20.7±1.9
6-mer-5 (UUUUUU)	Fast	317±40

Figure 17 RNA-binding of Hrp48-RRM2

A-E: Biophysical characterization of Hrp48-RRM2 RNA-binding. **Top:** Zoomed-in regions of the ^{1}H,^{15}N-HSQC NMR spectra of ^{15}N-labelled Hrp48-RRM2 titrated with 6-mer RNA constructs to three-fold excess. RRM2 does not bind 6-mer-1 and weakly binds 6-mer-5. 6-mer-3 and 6-mer-4 binding is in the fast exchange regime on the NMR time scale, and 6-mer-2 in the intermediate–slow exchange regime. **Middle:** CSPs represented over the sequence, with the same scale in order to illustrate the

differences. **Down:** Mapping of the CSPs on the structure of RRM2. The scaling for the coloring is the same for each titration. **F**: Table of NMR results of interaction studies between RRM2 and the 6-mers. The determination of the K_D by fitting the NMR data in the slow-intermediate exchange regime did not allow reliable fitting of the NMR data.

Next, the CSP values of the 6-mer – RRM2 titrations were mapped on the RRM2 structure (Figure 17 A-E Bottom). This shows that the canonical β-strands and the aromatic residues of the RNP motifs are involved, typical for RRM domains. However, apart from the putative RNA binding site, the α2-helix (6-mer-2 and 6-mer-3) and the non-canonical β4-strand (6-mer-2, 3, 4, 5) seem to also be involved in RNA binding. The peaks of Y158 of the β3'-strand disappeared in most titrations, which is also a sign of interaction, but neighboring residues showed strong CSPs indicating a role of this region in RNA binding (for example I159, Figure 17 Top or R157 Figure 17 Bottom). Upon titration with 6-mer-2, all aromatic residues of the RNPs (F99 on RNP-2 and F139, F141 on RNP-1), Y129 on loop 3 and peaks corresponding to residues of non-canonical β3'-strand and α2-helix exhibit strong CSPs (Figure 17 B Bottom). This might be due to direct interaction with RNA or due to allosteric effects. The table in Figure 17 F summarizes the characterization of RNA binding by RRM2.

Previously in this section I have shown the tetramerization tendency of the RRM2 domain of Hrp48 (Figure 16 F). My preliminary data suggests a link between oligomerization and RNA binding. On the ^{1}H, ^{15}N-HSQC NMR spectrum of RRM2 recorded at a high (500μM) concentration (Figure 18 A) some of the peaks exhibit doubling, which could correspond to the two different forms of RRM2: the monomer and the tetramer. I mapped these peaks on the AF2 model in order to visualize the putative tetramerization interface (Figure 18 B) and found, that all β-strands including the α2-helix and the non-canonical β3'-strand have residues that show peak doubling. Thus, the putative tetramerization and RNA binding interfaces overlap (CSP mapping of the RRM2 – 6-mer-2 titration on the RRM2 structure, Figure 18 C and D), suggesting a competition between self-association and RNA-binding.

In addition, self-association is also present in the case of the RRM12 construct and diminishes upon RNA binding. Comparing the light-scattering SEC-MALLS chromatograms of RRM12 and RRM12 in complex with 10-mer (Figure 18 E) shows that the oligomeric peak of RRM12 is reduced upon RNA binding. Additional studies directed to understand the link between RNA binding and self-associations could shed light on a potential mechanism with functional relevance.

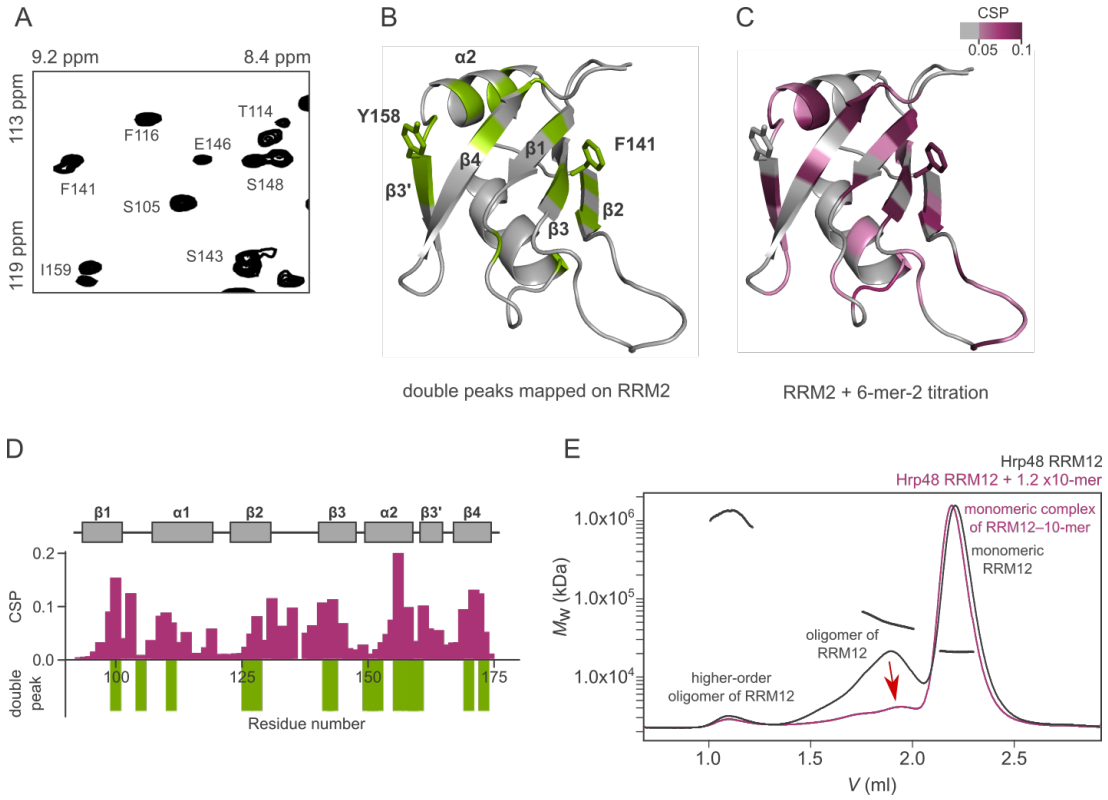


Figure 18 Self-association induced by RRM2

A: Zoomed-in section of the ^{1}H, ^{15}N-HSQC NMR spectrum of Hrp48 RRM2 at 500 µM concentration. **B:** AF2 model of RRM2 with residues show peak doubling highlighted in green. **C:** AF2 model of RRM2 with CSP mapping of the 6-mer-2 titration experiment. **D:** CSPs observed in the RRM2 – 6-mer-2 titration plotted over the residues (purple) and residues exhibiting peak doubling (green, arbitrary unit). **E:** SEC-MALLS light scattering chromatograms of Hrp48 RRM12 (grey) and Hrp48 RRM12 with 1.2 x excess of the 10-mer. The red arrow indicates the change of intensity in oligomeric peak upon addition of RNA.

5.3 Structural, biophysical and dynamics characterization of Hrp48-RRM12 and the Hrp48 – 10-mer complex

Structure determination of Hrp48-RRM12 by X-ray crystallography was attempted with no success despite numerous crystallization screens. As I have shown in the previous section 5.2, Hrp48-RRM2 tends to strongly tetramerize (Figure 16 F), and this self-association tendency applies also for Hrp48-RRM12, probably induced by the RRM2 domain. For this reason, I could not obtain a stable high concentration sample needed for NMR experiments for structure determination, which renders NOE-based structure calculation impossible. Nevertheless, the backbone chemical shift assignment of RRM12 was completed for 95% of residues for the free and 79% of residues for the RNA-bound state of non-proline residues. As substitution for an experimentally determined high-resolution structure, I used the AlphaFold2 prediction model of RRM12 as a model of the RNA-unbound state [122]. The validity of this model for RRM2 has been confirmed in the previous section, and for RRM1 I found that my crystal structure aligns well with the prediction (backbone RMSD = 0.45Å). The structural features of RRM1 and RRM2 have been discussed in the previous sections. In addition to these, AF2 predicts for Hrp48-RRM12 a short helix (α0) for the N-terminal residues of RRM1 that are not part of the crystal structure (Figure 20 A and C).

Previously, for RRM1 the 6-mer-3 and for RRM2 the 6-mer-2 was deduced as the optimal binding motif of the tested 6-mers. Based on this finding on the separate domains, a 10-mer RNA (UAGGAUUAAG, Figure 19 A) encompassing both 6-mers was used to test the RNA binding of the tandem RRM12 construct of Hrp48. This sequence element is four nucleotides further downstream as previously reported [72], and consequently it does not include the palindromic motif, allowing *in vitro* investigations without the risk of forming dsRNA and perturbance of the RNA-binding assessment by RNA conformational change. With the intention to test whether RNA binding of the tandem domains is stronger than that of the individual RRMs, I performed an NMR titration and ITC experiments to determine the binding affinity. The NMR titration data shows resonances in the slow exchange regime on the NMR timescale for most of the peaks (Figure 19 B) indicating tight binding in the nanomolar range. Both domains exhibit strong CSPs, indicating that both RRMs bind the 10-mer. However, the RRM1 residues exhibit somewhat stronger CSPs than the RRM2 residues (Figure 19 C). Here, it must be mentioned, that some parts of RRM2 could not be assigned in the RNA-bound state, and these residues might also exhibit CSPs upon RNA binding.

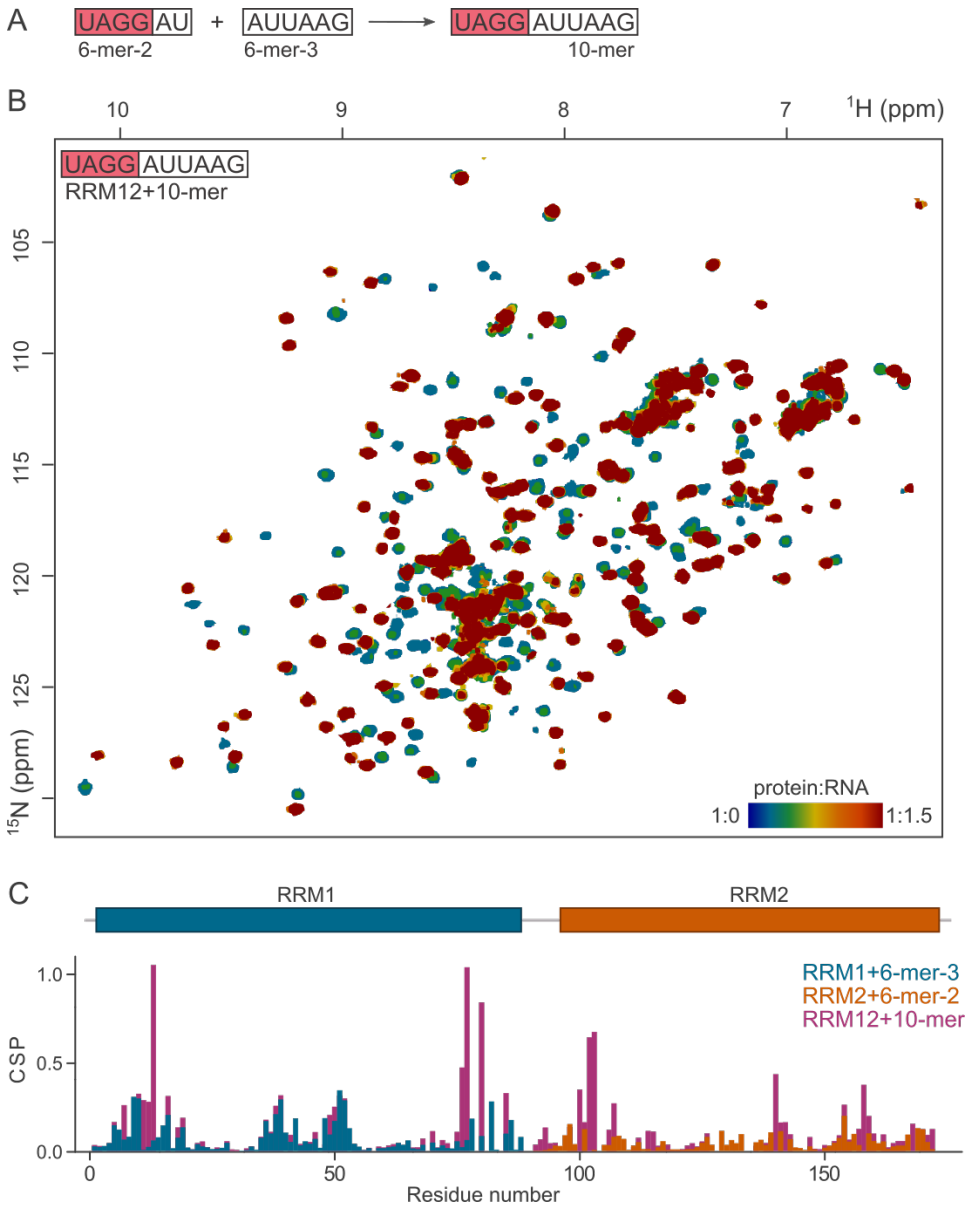


Figure 19 NMR titration experiment of Hrp48-RRM12 with 10-mer

A: The combination of 6-mer-2 and 6-mer-3 sequences covers a 10-mer. **B:** The 10-mer was used in a ^{1}H, ^{15}N-HSQC NMR titration experiment. **C:** CSP plots of the RRM1 – 6-mer-3, RRM2 – 6-mer-2 and RRM12 – 10-mer titrations. Above: the domain arrangement of Hrp48-RRM12.

The CSPs of Hrp48-RRM12 upon 10-mer binding are considerably larger than for the isolated domains during the 6-mer titration experiments, but the pattern of CSPs versus the residue numbers of the single-domain experiments overlap well with the CSPs of the 10-mer titration (Figure 19 C). The strongest CSPs were observed on the β-sheet, especially on β1- and β4-strands of RRM1, and β1, β3 and β3' for RRM2 (Figure 20 A). In addition, between α2 and β4 of RRM1 some residues (T76 and I77) also exhibit strong CSPs (Figure 20 A), which might

also be explained by allosteric effects, eventually similar to an N'-end direction elongated β4 and β3' formation in the case of RRM1 MD simulation (Figure 15).

I mapped the CSP values of the 10-mer – Hrp48-RRM12 titrations on the RRM12 AF2 structure model (Figure 20 B, C, D). This visualizes the RNA-binding interface along the β-strands and that the aromatic residues of the RNP motifs are involved (F10, W16, F50, F52 of RRM1, and F141 and Y158, Figure 20 C and D respectively). However, for RRM2 other canonical RNA-binding aromatic residues are less strongly affected than in the case of the 6-mer-2 titration. Importantly, the linker between RRM1 and RRM2 does not seem to be involved in the RNA binding.

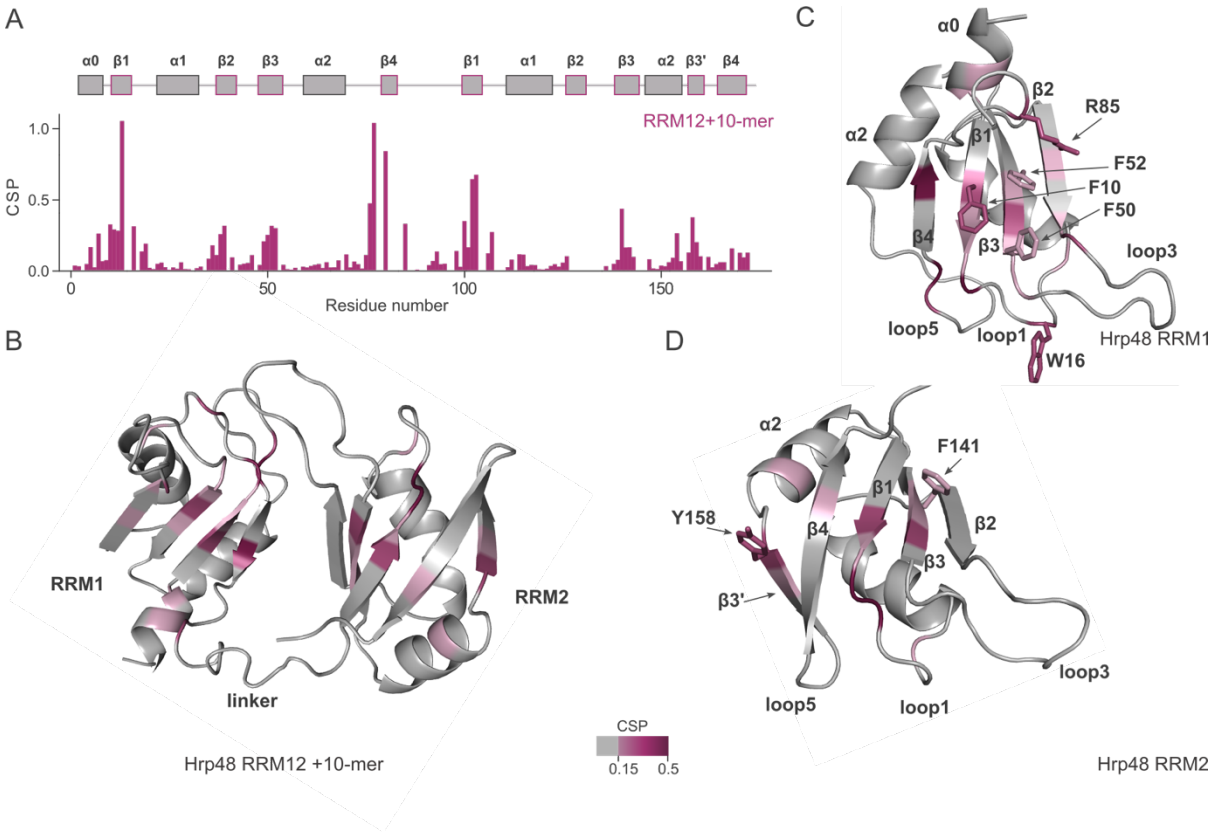


Figure 20 CSP values of the 10-mer – RRM12 titration mapped on the AF2 structural model of Hrp48

A: CSP plot of the RRM12 – 10-mer titrations. Above: the secondary structure arrangement of Hrp48-RRM12 based on the AF2 structure. **B:** RRM12 tandem domains with the mapped CSP values. **C** and **D:** RRM1 (C) and RRM2 (D) shown separately of the same structure and CSP mapping as in B. Secondary structure elements and side-chains of strongly interacting residues are shown.

The tight binding qualitatively assessed by the NMR titration is supported by ITC, as the determined dissociation constant is 273 nM (Figure 21 A). This strongly indicates that the tandem RRMs bind *msl-2* simultaneously and that the interaction is synergistic. The affinity of

Hrp48-RRM12 to the 10-mer is about 10-fold stronger than the affinities of the single RRMs to their respective 6-mers (Figure 14 J and Figure 17 F).

In order to further characterize the complex, I performed SEC-MALLS experiments on RRM12, the 10-mer and the complex of these and carried out a comparative evaluation (Figure 21 B) and a conjugate analysis (Figure 21 C). Comparing the UV-chromatograms of 10-mer and RRM12 with 1.2-times excess of the 10-mer shows a clear shift of the main peak. The molecular weights provided by the MALLS analysis suggested an increase in molecular weight of the RRM12 compared with the RRM12 with 1.2-times excess 10-mer (Figure 21 C). The conjugate analysis of the latter chromatograms could also differentiate the complex from free RRM12 and the 10-mer (Figure 21 C). The data suggests the formation of a 1:1 complex in solution, and which together with the NMR data confirms that both domains bind one RNA 10-mer at the same time.

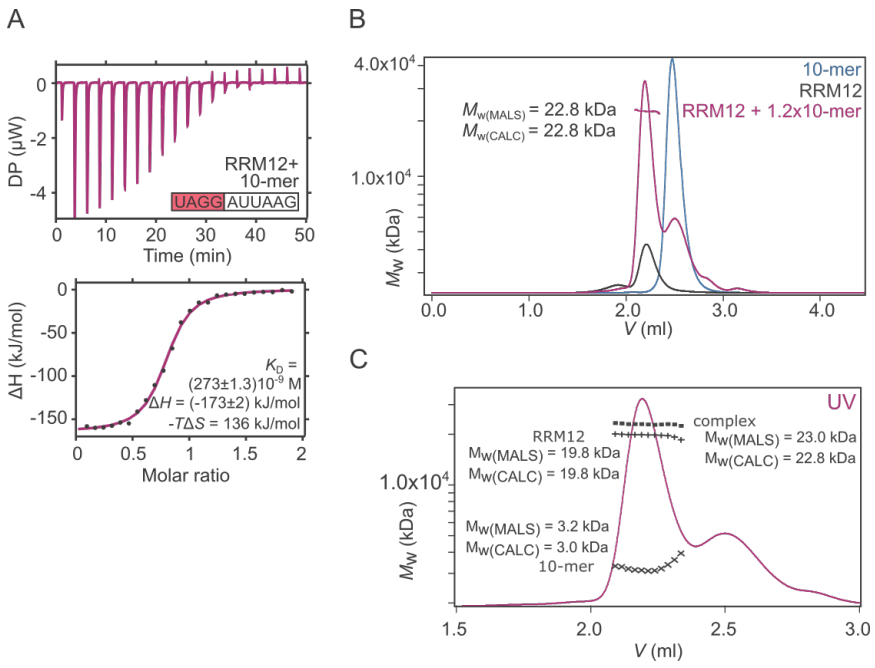


Figure 21 Biophysical characterization of Hrp48-RRM12 RNA binding

A: Isothermal titration calorimetry of Hrp48-RRM12 with the 10-mer. **B:** SEC–MALLS UV absorption chromatograms of the free RRM12 (grey), the 10-mer (blue) and their complex (purple). The marker indicates the molecular weight of the main peak of the complex chromatogram. **C:** SEC–MALLS UV absorption chromatogram and conjugate analysis of the Hrp48-RRM12 – 10-mer complex. Markers indicate the molecular weights of the different species (RRM12, 10-mer, RRM12 – 10-mer complex).

Having this strong RNA binder optimized, I attempted to obtain a crystal structure of an Hrp48-RRM12 – RNA complex. I managed to grow small crystals, but despite extensive

crystallization trials I could not optimize their size and they did not diffract. Together with Bernd Simon from our laboratory, we resorted to structure modelling based on homology, AF2 structure prediction and NMR data. In several cases, AF2 tends to predict the RNA bound conformation of RNA binding proteins such as the mouse RBM20 bound to the UCUU RNA motif (pdb: 6so9) [205], *Drosophila* Unr-CSD789 bound to poly-A RNA (pdb: 7zhh, Figure 8 E) [69] and *Drosophila* Sxl-dRBD3 in complex with *msl-2* or *transformer* mRNA (pdb: 4qqb, 1b7f, Figure 8 H) [25, 68]. We were wondering, whether the structural features and the domain-domain arrangement (distance and orientation of both RRMs with respect to each other) in the AF2 model of RRM12 could resemble the solution conformation in the RNA bound state.

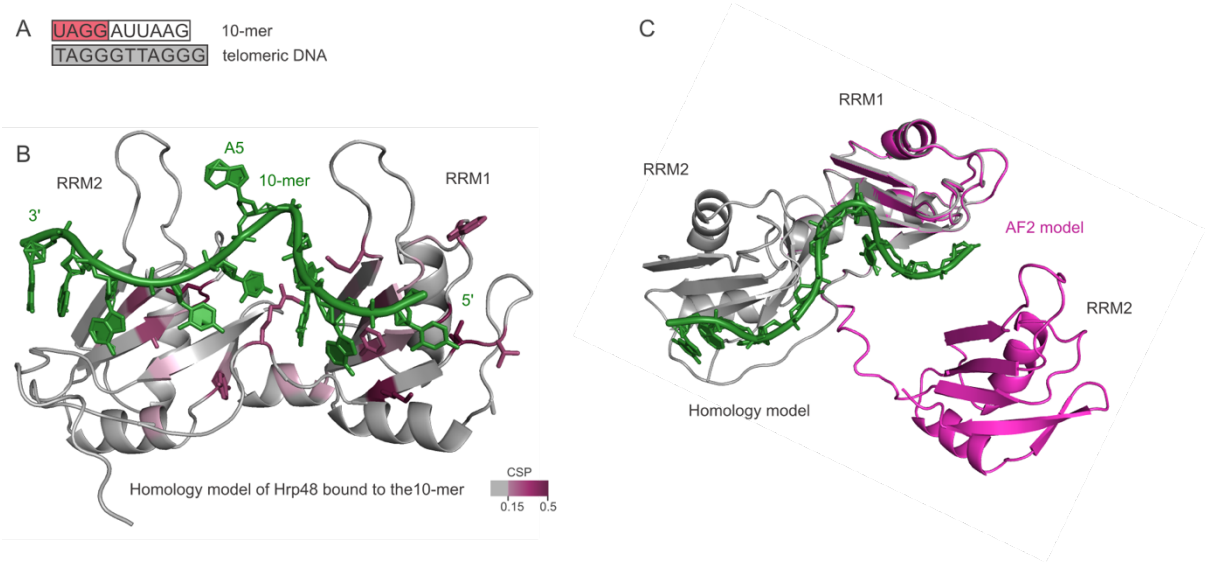


Figure 22 Homology model of the RNA-bound state of Hrp48-RRM12

A: Comparing the 10-mer and the telomeric DNA. **B:** Homology model of the Hrp48-RRM12 – 10-mer complex based on the crystal structure of Hrp48-RRM1, the AF2 prediction of Hrp48, and the UP1 structure bound to the telomeric DNA. RRM12: grey, with CSPs mapped, 10-mer: green. **C:** Superimposition of the homology model (grey) and the AF2 prediction (pink).

For this, Bernd Simon generated a homology model of tandem domains RRM12 Hrp48 based on the 2up1 pdb structure of hnRNP A1, homologous to Hrp48 (Supplementary Figure 1 and Supplementary Figure 2). hnRNP A1, that has been shown to be able to bind both DNA and RNA can recognize sequences similar to those Hrp48 binds, in the hnRNP A1 structure the TAGGGTTAGGG sequence of telomeric DNA (Figure 22 A). The hnRNP A1 structure was chosen as homology model of the different published hnRNP A1 structures based on the similarity of the bound nucleic acid sequence [101]. The modelling was performed by combining

hnRNP A1 and the AF2 models of RRM1 and RRM2 of Hrp48. The AF2 models superimpose well with the hnRNP A1 X-ray structure (RMSD: 0.726 for RRM1 and 0.944 for RRM2). In the crystal, the DNA is in contact with two different protein molecules through RRM1 and RRM2. To generate a monomer Hrp48-RRM12 with DNA, the AF2 RRM1 to hnRNP A1 RRM1 with DNA and AF2 RRM2 to the corresponding symmetric molecule in the crystal were aligned. To generate a template file with Hrp48-RRM12 bound to RNA, the DNA sequence was replaced to the correct RNA sequence keeping all identical atoms at their original position. The missing atoms and the broken linker bonds were generated by simulated annealing minimization of the model in CNS with all RNA atoms with known coordinates and the AF2 coordinates of RRM1 (residues 1-86) and RRM2 (residues 95-173) fixed.

Having obtained this model, I compared the CSP values derived from the 10-mer titration experiments for qualitatively validating the RNA binding. For this, I mapped the CSP values on the generated model (Figure 22 B). Most of the CSPs are in good accordance with the protein – RNA interface, but the relative orientation of the two domains is very different to the one observed in the AF2 predicted structure (Figure 22 C), implying that AF2 does not predict the RNA-bound conformation of Hrp48.

In order to analyze the dynamics of the RRM12 – 10-mer complex, I also measured R_1 and R_2 relaxation rates and calculated the rotational correlation times (τ_c) of Hrp48-RRM12 in the free and RNA bound states (Figure 23 A, B, C). The two domains of RRM12 in the apo-state tumble independently from each other, as the τ_c is 5.6 ns (Figure 23 A, for both domains separately as well). The rotational correlation time is directly proportional to the molecular weight (rule of thumb, τ_c of a monomeric protein is 0.6 time the molecular weight for isotropic tumbling, assuming a globular, sphere-like shape of monomers) which is in good accordance with the molecular weights of 10 kDa of the single RRMs. This also indicates, that there is no interaction between RRM1 and RRM2 in the apo-state.

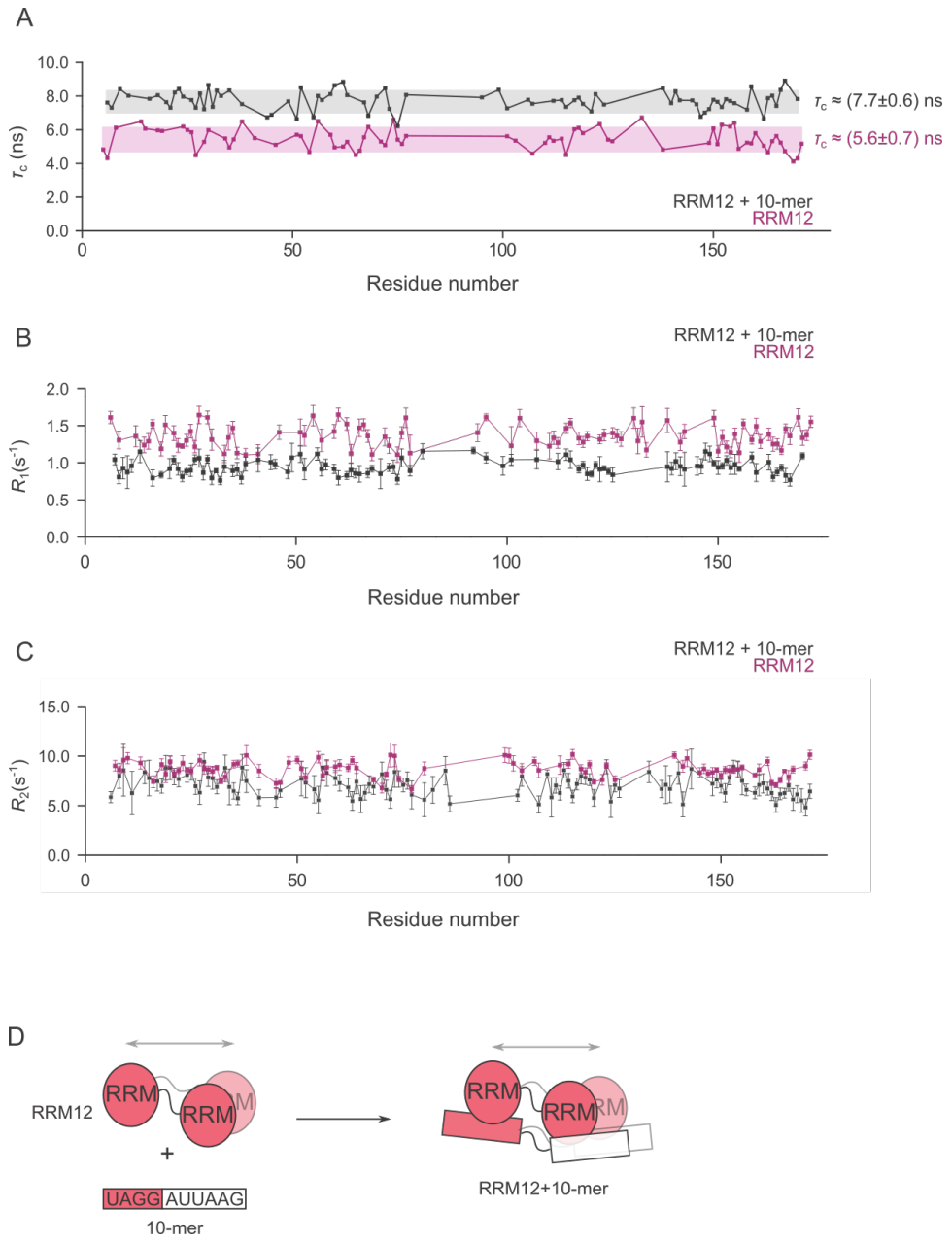


Figure 23 Dynamics of the RRM12 – 10-mer complex

A: Rotational correlation times of RRM12 residues in the apo state (purple) and bound to RNA (grey). The rectangular space is centered around the average τ_c and the height represents two times the standard deviation of the datapoints. **B, C:** Longitudinal (R_1) and transversal (R_2) relaxation rates of the residues in the apo state (purple) and bound to RNA (grey). **D:** Model of the hypothesis of the RNA binding of Hrp48. In the RNA free state, the two domains do not interact with each other and tumble independently. Upon RNA binding, the two domains can still tumble independently while both are bound to RNA.

Upon RNA binding to Hrp48-RRM12 the τ_c increases uniformly by 2.1 ns for both domains. This is unexpected, considering the findings about the RRM12 – 10-mer complex. If both domains bind RNA simultaneously to form a rigid 1:1 complex, where both domains are

assumed to be in a fixed conformation relative to each other, the complex would behave as a 23 kDa molecule. The τ_c vales would reflect this as an about two-fold increase and the expected value would be around 14 ns. The possibility, that the RRM12 was in the RNA unbound state and that I sampled the RNA-unbound data is to be excluded, as the chemical shifts of the peaks reflected the RNA-bound state which is very different from the apo-state chemical shifts. An explanation could be that both domains are still able to move relative to each other while bound to the 10-mer. This would be possible only, if one or more nucleotides in the middle of the 10-mer do not interact with Hrp48-RRM12, which could leave flexibility in the relative orientation of the two RRMs (Figure 23 D). Of note, this has to my knowledge so far not been reported for tandem RRM domain proteins bound to RNA. Also, crystal structures cannot report on this and the lack of success in crystallizing this complex could be due to this remaining mobility between both RRM domains in the RNA bound state.

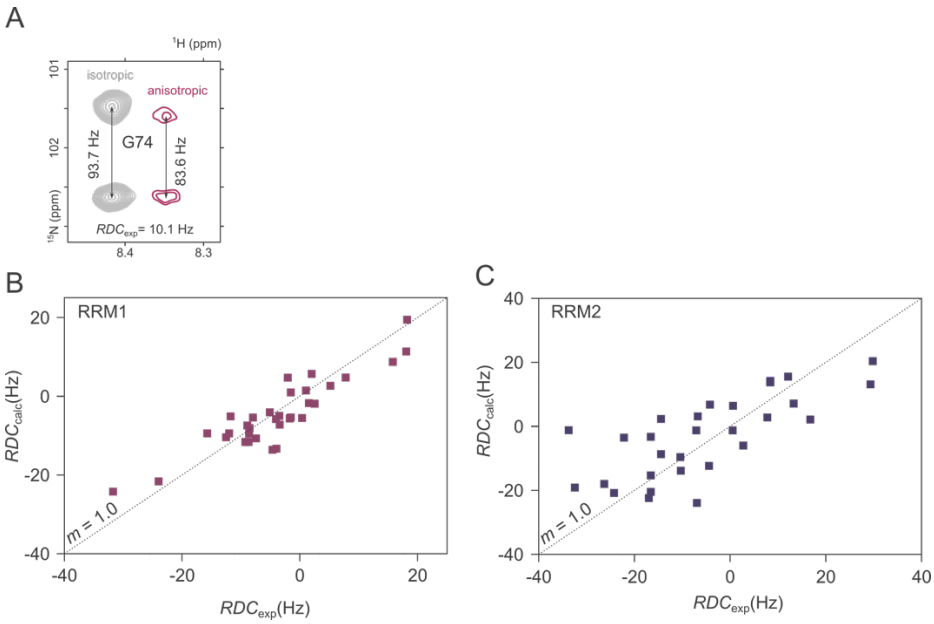


Figure 24 Residual dipolar coupling measurements on Hrp48-RRM12 – 10-mer complex

A: Calculating RDC values on the example of the G74 residue of Hrp48-RRM12. Grey: spectrum measured in isotropic conditions, purple: spectrum measured in anisotropic conditions. RDC is the difference of the coupling constants measured in isotropic and anisotropic conditions. **B and C:** Measured and back-calculated RDC values of RRM1 and RRM2. If the back-calculated values are in a good agreement with the experimental RDCs, the values deviate around a linear with a slope of 1.

With the aim to further understand the overall structure and dynamics of the complex, I measured N-H residual dipolar couplings (RDCs) of the Hrp48-RRM12 – 10-mer complex. RDCs are usually used to give long- range orientational restraints for structure determination.

RDCs in ideal solutions are averaged out (isotropic conditions) due to molecular motions, but aided by alignment medium, the molecules can be partially aligned (anisotropic conditions). The RDCs are the difference of the dipolar coupling constants measured in isotropic and anisotropic conditions (Figure 24 A), and give information on the angles of the N-H bond vectors relative to the external magnetic field which can be used to determine the orientation tensors of the protein. In a complex, or multi-domain protein, the tensors can be aligned relative to each other to define the relative orientation of the domains or different proteins in the complex, if they can be treated as rigid bodies. This means, if there is a fixed relative orientation, the RDC is a powerful method for determining it. To this end, I first determined the RDCs of RRM12 in complex with the 10-mer, using Pf1 phage as alignment medium. Using the Module software [187] and the model of the complex, I fit the RDC values on the structure model of RRM1 and RRM2 separately (Figure 24 B and C) and removed outliers based on the difference of the measured and back-calculated RDCs. These pre-filtered RDCs were then used for fitting the domains and the relative orientation in the DC program of NmrPipe by Bernd Simon. The separate domains could be fit and the orientation tensor was determined, but we did not manage to find the relative orientation. This would further confirm the lack of a fixed relative orientation of the domains, which is in good accordance with the relaxation data, suggesting that in the 1:1 complex of Hrp48-RRM12 and the 10-mer both domains bind the RNA, but remain flexible with regards to the orientation between both RRM domains in the complex.

In order to test this hypothesis, short 4-mer motifs were tested for binding in ^{1}H, ^{15}N-HSQC NMR titration experiments with Hrp48-RRM12 (Figure 25). The 4-mers were designed to cover the whole Region 5 and Region 6 and consequently the 10-mer motif (Figure 25 A). The assessment of the results was performed based on the magnitude of the CSP at 1:1 protein:RNA ratio. This shows that the 4-mer-3 (UAGG) and the 4-mer-6 (UAAG) motifs are clearly recognized by RRM12 (Figure 25 B and C). Interestingly, these two motifs mark the two ends of the 10-mer RNA oligomer, **UAGG**AU**UAAG**. In addition, the two 4-mers in-between, 4-mer-4 (GGAU) and 4-mer-5 (AUUA) do not show strong CSPs, implying at most only non-specific weak RNA binding. This is also in accordance with our model of RNA-bound Hrp48-RRM12, where A5 is not in contact with the protein (Figure 22 B). These findings support the hypothesis, that RRM12 binds the 10-mer close to the 3' and 5' ends, and the middle of the 10-mer is not significantly involved in the interaction. Furthermore, the observation that 4-mer-1 (AACC) and 4-mer-2 (CCUA) also only weakly / unspecifically interact with RRM12, validate

that the Hrp48 binding motif starts with the 4-mer-3 (UAGG), further supporting that the 10-mer is the optimal binding sequence of Hrp48-RRM12 on the *msl-2* mRNA.

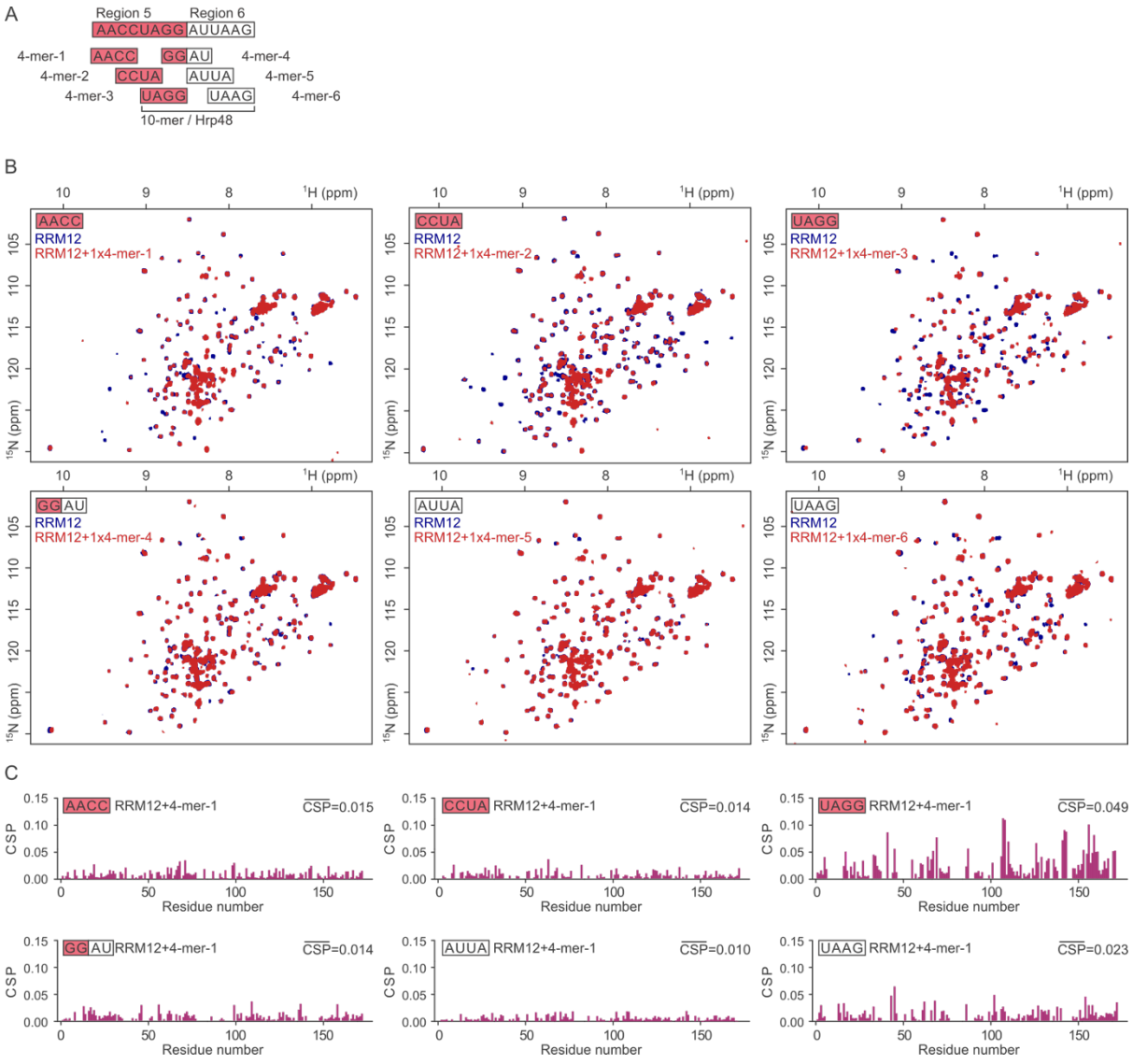


Figure 25 NMR titration experiments of Hrp48-RRM12 with the 4-mer RNA

A: Refinement of the RNA motif bound by Hrp48-RRM12. The sequences of the 4-mers used for the tests are shown. **B:** ^{1}H, ^{15}N-HSQC NMR spectra of free RRM12 (blue) and RRM12 with one equivalent 4-mer (red). **C**: The chemical shift perturbation of the RRM12 residues upon addition of one equivalent 4-mer. The average CSP was calculated from the CSP values greater than the standard deviation of all CSP values. The standard deviation was calculated for each experiment separately.

5.4 The Hrp48 binding site is required for optimal *msl-2* translational repression

To validate the refined Hrp48 binding site of *msl-2* mRNA, *in vitro* translation assays were performed using different mutants of the *msl-2* 3' UTR (Figure 26 A). The level of translation was measured in *Drosophila* embryo extracts containing the mRNA with the luciferase reporter and increasing amounts of recombinant Sxl-dRBD4. (dRBD4 is the Sxl construct, that is sufficient for full translation repression activity (Figure 7B).) The effect of the mutations was assessed by following the change in the luciferase activity, as previously described [62].

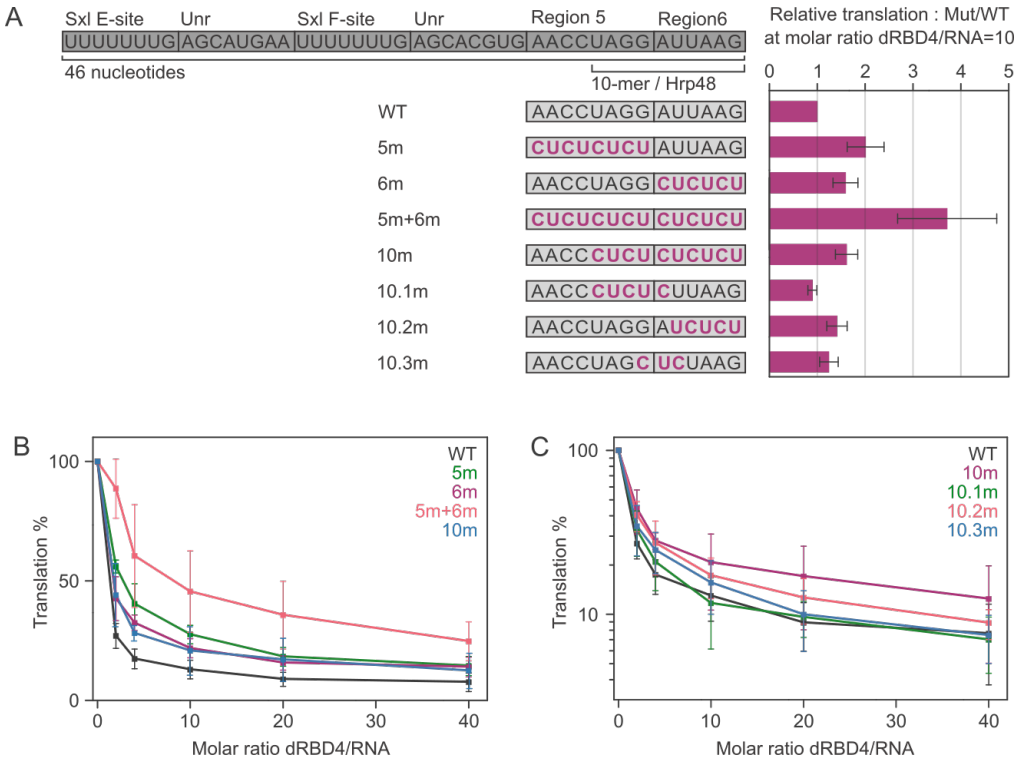


Figure 26 In vitro translation assays

A: Schematic representation of the 46 nucleotides of the 3' UTR important for translational repression. In vitro translation assays were performed for various constructs modified at Region 5 and Region 6 of the *msl-2* mRNA. The CU-repeat substitutions are indicated with bold purple font. This panel shows the relative translation activity at a Sxl-dRBD4/RNA ratio of 10. Co-translated Renilla luciferase was used as an internal control and as a reference for normalization of the reporter Firefly luciferase signal, and the data was also normalized to the wild type (WT) signal activity. The standard deviations of the three replicate experiments are represented by error bars. **B-C:** *In vitro* translation assay results for all used Sxl-dRBD4/RNA concentration ratios. The data was plotted as relative translation (in %) assuming that the initial data point with no Sxl has 100% activity. The standard deviations of three replicate experiments are represented by error bars. On panel C the y axis is logarithmically scaled, because of the modest difference between the translation efficiency mutants.

The minimal region necessary for the 3' UTR translational repression comprises 46 nucleotides spanning the E site to Region 6 (Figure 26 A) [72]. Both Region 5 and 6 are required for optimal translation repression [65, 72]. To validate my findings on the *msl-2* RNA binding of Hrp48, I designed and cloned the 8 different sequences tested in the reporter studies. The assays and the data processing were conducted by our collaborator, Tanit Guitart from the Fátima Gebauer laboratory. Several RNA constructs were tested for *in vitro* translation activity in Luciferase reporter assays (Figure 26). I used CU-repeats to create substitution mutants of specific parts of the 3'UTR. 5m is a mutant in which Region 5 is replaced and was previously tested and identified as the Hrp48 binding site [72]. In the 6m mutant Region 6 is substituted, and the 10m mutant covers the refined Hrp48-RRM12 10-mer binding site (10m). We also investigated the impact on translation repression upon substituting the entire region (Region 5 + Region 6, termed 5m+6m) with CU-repeats.

The 5m+6m mutant has the strongest de-repression effect and when compared at a Sxl-dRBD4/RNA concentration ratio of 10 (Figure 26 A and B), the effect of 5m+6m is about the same as 5m and 6m together. This could be a result of two different events happening at Region 5 and Region 6 simultaneously, and not cooperatively. The 10m mutant has a smaller effect than 5m+6m, suggesting that the first four nucleotides of Region 5 might have a different and additional role, than Hrp48-binding. Moreover, the translation repression activity of the 10m and the 6m mutants are similar, proposing that the 6m region is a key for *msl-2* recognition and translational repression activity of Hrp48. Indeed, Hrp48-RRM1 binds the isolated Region 6 with the highest affinity (AUUAAG, the 6-mer-3, Figure 14 F). Above an Sxl-dRBD4/RNA concentration ratio of 10, the effects of 5m, 6m and 10m are similar and 5m+6m remains to affect the translation repression significantly stronger (Figure 26 B).

In order to study the relevance of different parts of 10m on translation, mutations of sections of the 10-mer (10.1m, 10.2m and 10.3m) were also tested (Figure 26 A and C). The results of these mutants did not show high variations, and for this reason these results might not be representative. Because of the low magnitude in change, the graphs are represented on a logarithmic scale. What we can say, is that from mutations in the 10-mer region, 10.2m has the strongest de-repression effect, which could underline the previous result that the 3'end of the 10-mer is more important for Hrp48 binding and translation than the 5' end (Figure 26 C).

Overall, these assays emphasize the importance of Region 6 and the 10-mer in the *msl-2* translation repression, but further cell-based studies would be needed to cross-validate the effects with the binding of Hrp48.

5.5 Hrp48-RRM12 binds independently of Sxl-dRBD3 and Unr-CSD12 to *msl2*-mRNA

After gaining mechanistic insight into the Hrp48 binding mode and base specificity my aim was to assess whether Hrp48-RRM12 would form a complex with Sxl, Unr on the *msl-2* mRNA. It has been shown previously that Sxl-dRBD3 and Unr-CSD1 do not interact without RNA, but protein-protein contacts between the two proteins are established upon RNA binding [68]. My aim was to use an extended RNA comprising of binding sites for all three proteins and to see if Hrp48 incorporates into a complex with Sxl and Unr, and whether their binding to *msl-2* is cooperative.

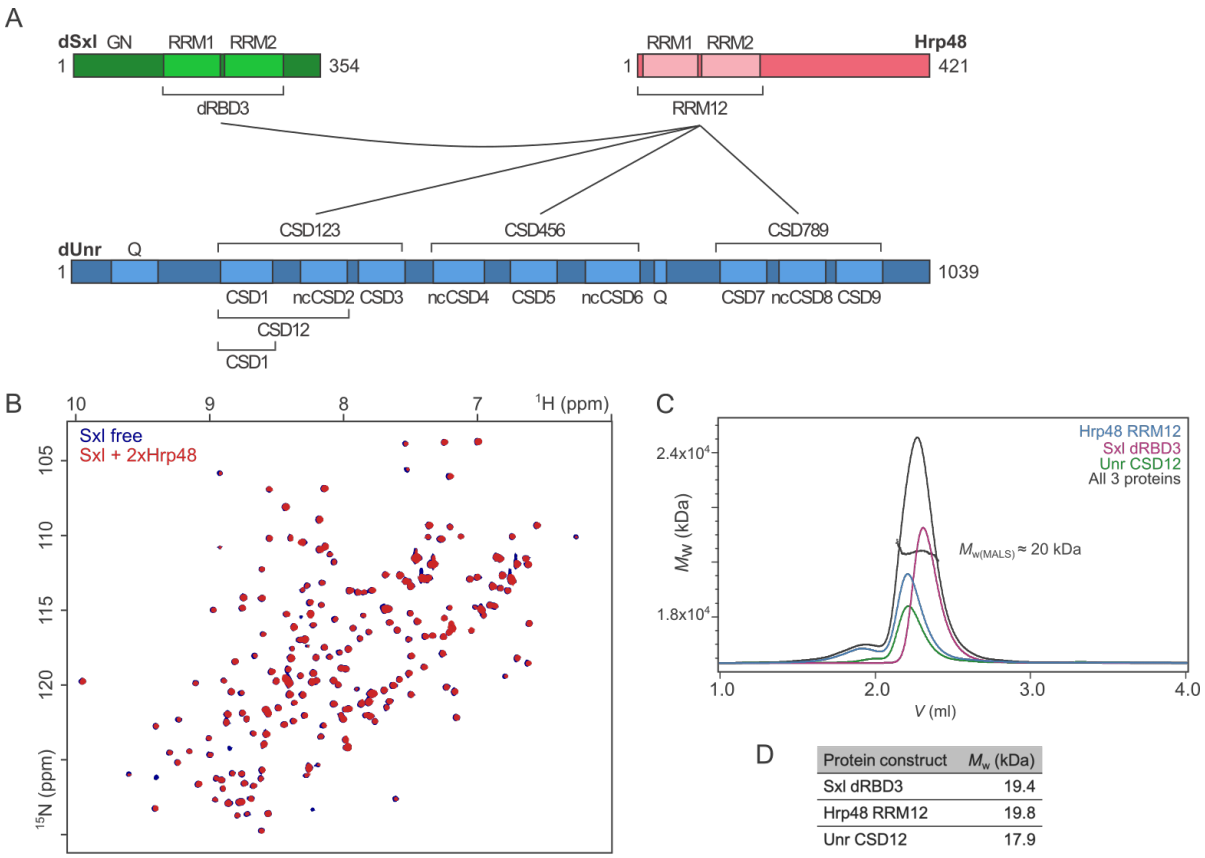


Figure 27 Sxl, Unr and Hrp48 do not interact in the absence of RNA

A: Domain arrangement of Sxl, Unr and Hrp48 with the indicated constructs used to test binding between these proteins. The connecting lines represent which constructs were tested which one another in $^{1}H,^{15}N$-HSQC NMR titration experiments. **B:** $^{1}H,^{15}N$-HSQC NMR spectra of ^{15}N labelled Sxl-dRBD3 free protein (blue) and Sxl mixed with two-fold excess Hrp48-RRM12 (red). **C:** SEC–MALLS UV absorption chromatograms of the proteins Hrp48-RRM12 (blue), Sxl-dRBD3 (purple), Unr-CSD12 (green) and the proteins at the same concentrations, corresponding for a 1:1:1 ratio, injected together (grey). The grey marker indicates the molecular weight of the peak in the grey chromatogram. **D:** Table of the molecular weights of Sxl-dRBD3, Hrp48-RRM12 and Unr-CSD12.

Initially, I tested whether Hrp48 interacts with Sxl or Unr in the absence of RNA. Hrp48-RRM12, Sxl-dRBD3 and different constructs of Unr were expressed unlabeled and ^{15}N labelled and purified to test whether they interact directly in absence of RNA by $^{1}H,^{15}N$-HSQC NMR experiments (Figure 27 A and B, Supplementary Figure 6). The domain boundaries of Unr were based on previous work in our lab, where four additional non-canonical CSDs were identified in between the five predicted canonical CSDs (Figure 7 and Figure 27) [114]. $^{1}H,^{15}N$-HSQC spectra were recorded of ^{15}N labelled Hrp48-RRM12 alone, and after addition of unlabeled Sxl-dRBD3 or different constructs of Unr (Supplementary Figure 6). I also performed the reverse experiment with ^{15}N-labelled Sxl-dRBD3 and unlabeled Hrp48-RRM12 (Figure 27 B). No significant CSPs or decrease in intensities upon addition of the titrant are observable, indicating that Sxl-dRBD3 and Unr do not interact with Hrp48-RRM12 in the absence of RNA. The slight decrease in intensity for some of the peaks might be a result of unspecific interactions or aggregation of the proteins. I also used SEC–MALLS to confirm the lack of stable interaction between the three proteins (Figure 27 C). For this I only tested the protein constructs (Hrp48-RRM12, Sxl-dRBD3, Unr-CSD12) that were used for the complex formation with the RNA (see later). Overlay of SEC–MALLS chromatograms indicate that the chromatogram of all proteins injected together is the sum of the single-protein chromatograms and there is no shift of the peak towards shorter retention times. MALLS-analysis indicates that the molecular weight of the peak of the combined injection is around 20 kDa, which corresponds to the molecular weights of the single protein components (Figure 27 D). This corroborates the findings that Hrp48-RRM12 does not interact with structured domains of Sxl and Unr in the absence of RNA.

With the intention of testing whether Sxl, Unr and Hrp48 jointly interact with *msl-2*, which has been proposed earlier [72], I designed a 30-mer RNA construct (WT-30-mer, Figure 28 A) based on published data for Sxl and Unr [68] and the findings of the Hrp48 binding site (see Section 5.3). This RNA oligomer combines the regions of Sxl-dRBD3 and Unr-CSD1 binding (F site) and the Hrp48 binding site (10-mer) (Figure 28 A). I also designed a different RNA construct, GG-30-mer, in which I introduced a mutation to abolish the palindromic sequence and preclude the self-association of the RNA. For this only two nucleotides had to be mutated (Figure 28 A). The wild-type and the GG-mutant construct were also tested in *in vitro* Luciferase reporter assays in embryo extracts. No significant differences in de-repression activity was observed

for the mutants compared to the wild-type translation levels, which supports the utilization of the mutants in biophysical studies (Figure 28 B).

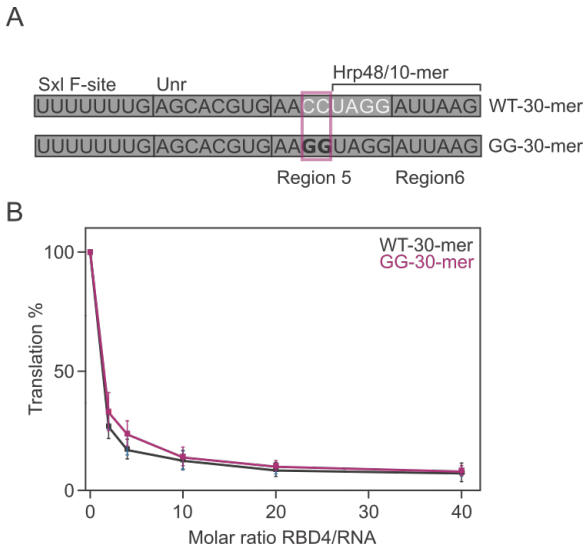


Figure 28 Mutation abolishing the palindromic sequence

A: The RNA constructs used for complex formation. Both constructs contain one binding site for each protein. The GG-30-mer is a mutant in which the palindromic sequence is abolished, to decouple the dimerization from protein binding in biophysical experiments. **B:** Mutation of C3 and C4 of Region 5 does not impair the translation repression activity. In vitro translation assays were performed for two constructs modified at C3 and C4 to UU or GG, highlighted on the graph by purple font color. The assays were performed with increasing amounts of recombinant Sxl-dRBD4. Co-translated Renilla luciferase was used as an internal control and as a reference for normalization of the reporter Firefly luciferase signal. The data was plotted as relative percentage with considering the data point of no Sxl as the 100% activity. The standard deviation of three replicate experiments are represented by the error bars.

For complex formation studies involving RNA, two Unr constructs containing the CSD1 domain were used (CSD1 and CSD12), as CSD1 engages in direct interaction with Sxl and both proteins jointly and sequence-specifically bind the F site. Initially, only the isolated protein constructs and the RNA separately were injected to obtain the reference molecular weights, which also helped to assess the quality of the components I aimed to use (Supplementary Figure 7). All protein samples were pure and their measured molecular weight was in accordance with the actual molecular weight. Hrp48-RRM12 exhibited a small degree of self-association, which I observed earlier and this was thus assigned to the tetramerization tendency mediated by the RRM2 domain (Figure 16 F and Supplementary Figure 7 C).

I started by optimizing the formation of the ternary complex of Sxl-dRBD3, Unr-CSD12 and GG-30-mer, as this was expected to be straightforward based on the literature background [68].

For this, the two proteins were mixed together and incubated on ice with the GG-30-mer, with different molar ratios, and I obtained the highest amount of complex at 1:2:3 RNA:dRBD3:CSD12 proportion, on Figure 29 A the grey chromatogram main peak corresponds to this complex (measured *MW*: 48.0 kDa, calculated: 47.1 kDa). Next the Hrp48 amount was optimized, and for this a 2-fold excess was used respective to the RNA. On Figure 29 A the main peak of the purple chromatogram corresponds to the quaternary complex (see also on Figure 29 B, with all detectors), and due to the low resolution on the SEC-column the *MW*-determination was affected by peak overlap (measured *MW*: 63.9 kDa, calculated: 66.9 kDa). The peak in the overlapping region with shorter retention time corresponds to the oligomerized protein, whereas the peak with higher retention time corresponds to excess free protein. I also tested complex formation with the WT-30-mer, and the same protein:protein and protein:RNA ratios were used for the WT-30-mer complexes as for the GG-30-mer. WT-30-mer was injected with Sxl-dRBD3 and Unr-CSD12 to pre-form the ternary complex, I found that the molecular weight is almost twice as high as the molecular weight expected for a quaternary complex (83.2 kDa instead of 47.1 kDa, Figure 29 C). An explanation for this could be the duplex formation due to the palindromic sequence located within the WT-30-mer, which does not happen in the case of the GG-30-mer (Figure 29 A). For this reason, I used GG-30-mer in ITC and gel-mobility shift assays (see below) to avoid additional interactions complicating the biophysical characterization of the complex. However, upon addition of Hrp48, the monomeric quaternary complex was detected as the *MW* and the peak shifted back to the expected range (measured *MW*: 66.3 kDa, calculated: 66.9 kDa Figure 29 D). The peak shift upon addition of Hrp48 to lower molecular weight means that Hrp48 binding opens the double-stranded RNA. Therefore, Hrp48-RRM12 forms a stable quaternary complex with Sxl-dRBD3, Unr-CSD12 and the 30-mer RNA, both with the wild-type sequence and the GG substitution mutant. The complete comparison of the complex formation with the different 30-mer and schematic illustration of the complexes is represented on Supplementary Figure 8.

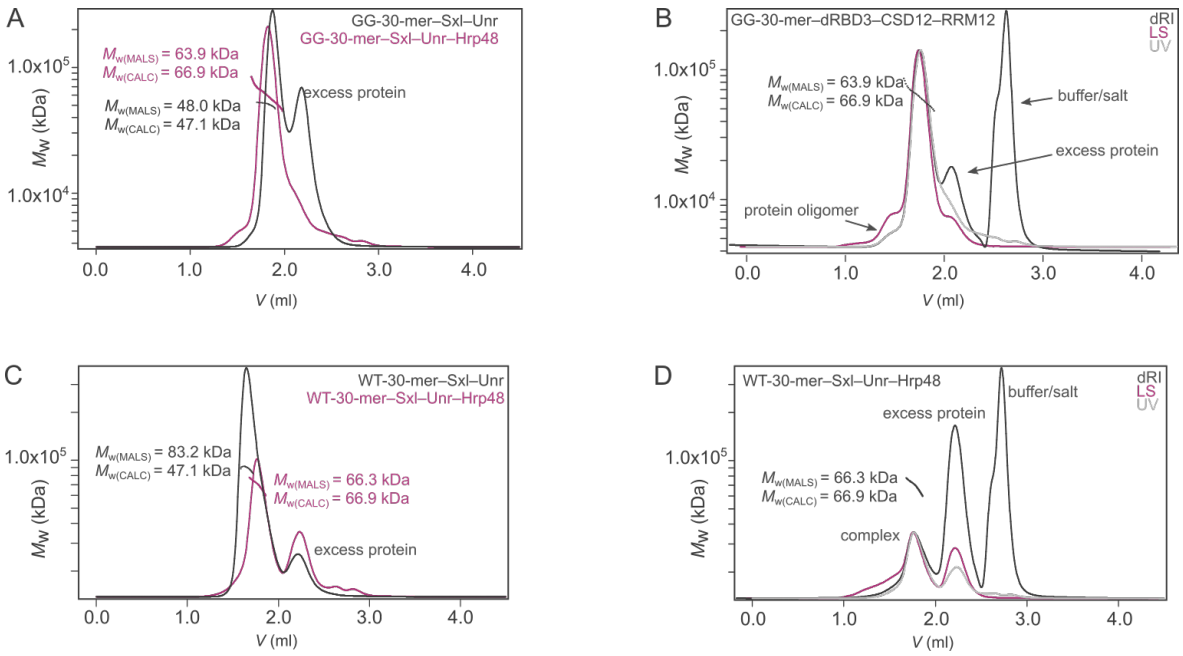


Figure 29 Complex formation of Hrp48 with Sxl, Unr and *msl-2*

A-D: SEC–MALLS chromatograms of the different complexes of *msl-2*, Hrp48, Sxl and Unr. The markers indicate the molar weights of the molecules/complexes in the corresponding peaks. **A:** SEC–MALLS UV chromatograms of dRBD3–CSD12–GG-30-mer (purple) and RRM12–dRBD3–CSD12–GG-30-mer (grey) complexes using proteins in excess. **B:** SEC-MALLS chromatograms of the GG-30-mer-Hrp48-RRM12-Unr-CSD12-Sxl-dRBD3 complex. (LS), Dark grey: differential refractive index (dRI), purple: light scattering, light grey: UV absorption. **C:** SEC–MALLS UV chromatograms of dRBD3–CSD12–WT-30-mer (purple) and RRM12–dRBD3–CSD12–WT-30-mer (grey) complexes using proteins in excess. The WT-30mer in complex with dRBD3 and CSD12 forms a dimer through the palindromic region of the WT-30-mer. This opens up when RRM12 is added and a 1:1:1:1 complex forms. **D:** SEC-MALLS chromatograms of the WT-30-mer-Hrp48-RRM12-Unr-CSD12-Sxl-dRBD3 complex. (LS), Dark grey: differential refractive index (dRI), purple: light scattering, light grey: UV absorption.

Unr-CSD1 was tested similarly for complex formation (Supplementary Figure 9). Here, I used the same protein-RNA and protein-protein ratios as for Unr-CSD12. Unr-CSD1 stably forms a ternary complex with both 30-mers and Sxl-dRBD3. However, with the WT-30-mer a dimer of the ternary complex is formed due to the palindromic motif, similar to CSD12 (Supplementary Figure 9 A) and with the GG-30-mer a monomeric ternary complex (Supplementary Figure 9 C). After addition of Hrp48-RRM12, the quaternary complex forms with the WT-30-mer (Supplementary Figure 9 C), but also a higher molecular weight peak appears (109 kDa), that I could not assign to any of expected molecular weights. A high molecular weight peak (123.8 kDa) also appears for the GG-30-mer, Unr-CSD1, Sxl-dRBD3 and Hrp48-RRM12 complex formation, where the molecular weight of the main peak, possibly

corresponding to the quaternary complex, has a greater offset from the expected molecular weight (45.9 kDa instead of 57.4 kDa, Supplementary Figure 9 D). For this reason, CSD12 was chosen for further studies on the complex formation.

Having established that the three proteins form a stable complex with *msl2*-mRNA, it still remains unclear whether binding of Hrp48-RRM12 is cooperative or independent of the Sxl-Unr moiety. Therefore, I wanted to assess whether Hrp48-RRM12 binding is synergistic with Sxl-dRBD3 and Unr-CSD12 for complex formation with the 30-mer RNA. For this, I performed ITC titrations and electrophoretic mobility shift assays with the GG-30-mer RNA construct to avoid RNA duplex formation, which complicates data interpretation and distinction between whether the read-out changes are due to RNA binding or RNA self-association. ITC experiments with GG-30-mer (syringe) titrated into Hrp48-RRM12 (cell) reproducibly resulted in a complex ITC curve (Figure 30 A). The positive heat change is not expected for RNA binding, but could be explained by RNA conformational change upon dilution. This was excluded by performing the control experiment of titrating the 30-mer-GG RNA into buffer (Figure 30 B). This indicates no effect of dilution on the RNA which would complicate the evaluation of the ITC data. Another plausible explanation for this would be that Hrp48-RRM12 binds the RNA multiple times but I could not fit the data reliably. However, it was possible to decouple the binding events by electrophoretic mobility shift assay (Figure 30 C) and found that the affinity of the first binding event is between 100 nM and 250 nM, which is comparable to the Hrp48-RRM12 binding to the 10-mer RNA (273nM, Figure 21 A). This indicates that the first event is the binding to the 10-mer motif within the GG-30-mer. The later binding event takes place at very high Hrp48-RRM12 concentrations, and possibly binds at the Unr-CSD1 or the Sxl-dRBD3 binding site. The data on Hrp48-RRM1 and RRM2 show weak interactions also with the U6-mer RNA (6-mer-5, for RRM1 Figure 14 I, J and for RRM2 Figure 17 F) and as the Sxl F binding site encompasses seven uridines (Figure 28 A), Hrp48 could also bind weakly to this region of the 30-mer. Due to the low affinity of the second binding event, this was considered as biologically irrelevant and as an artefact arising from the *in vitro* experimental arrangement. ITC titrations of the pre-formed complex of GG-30-mer, Sxl-dRBD3 and Unr-CSD12 was titrated into Hrp48-RRM12 (Figure 30 D). This ITC curve corresponds to a one-site binding, thus the binding of Sxl-dRBD3 and Unr-CSD12 to the GG-30-mer RNA abolished the second binding event and I could measure the K_d of Hrp48-RRM12 binding to this preformed complex. Surprisingly, the affinity was about 10-fold weaker than when Hrp48-RRM12 interacts with the 10-mer RNA, which precludes synergistic binding of

Hrp48-RRM12 to GG-30-mer in the presence of Sxl-dRBD3 and Unr-CSD12. Reverse titration experiments, which are useful controls, were excluded due to the oligomerization tendency of Hrp48-RRM12 at concentrations required for the ITC syringe.

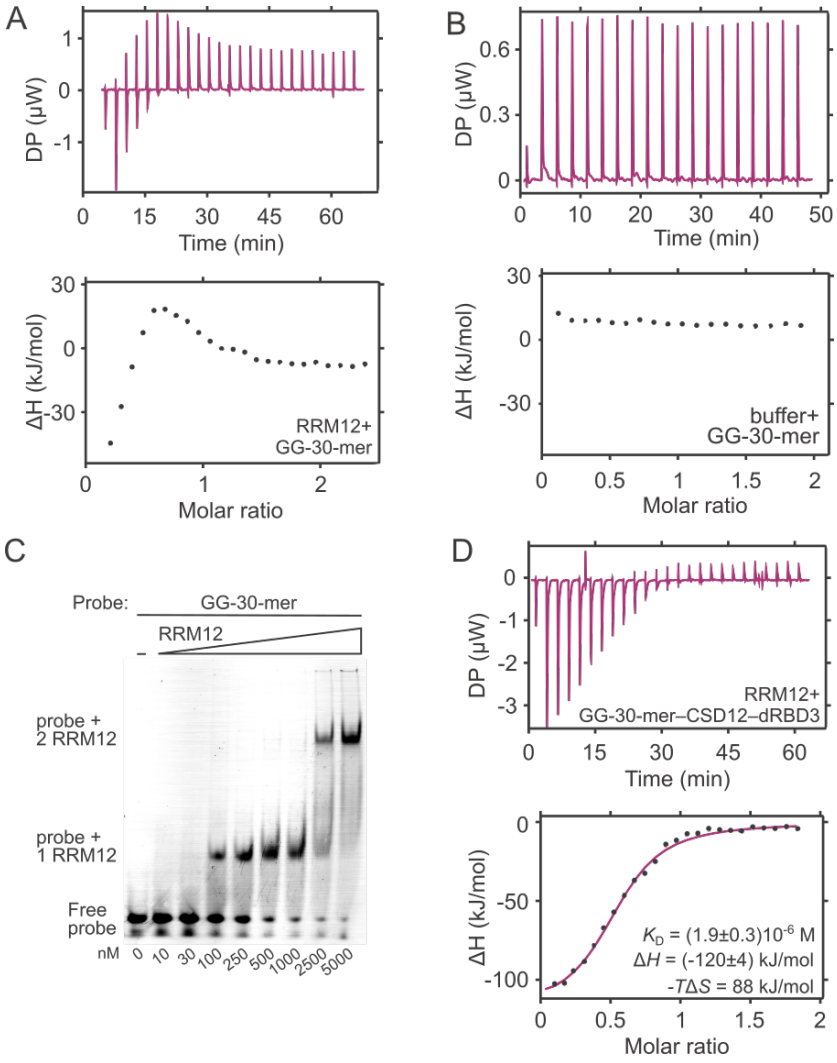


Figure 30 Biophysical characterization of the complex formation

A-C: Isothermal titration calorimetry of GG-30-mer titrated into Hrp48-RRM12 (**A**). **B:** Control ITC experiment: 30-mer RNA titrated into the buffer. **C:** Gel mobility shift assay using the GG-30-mer and increasing amounts of Hrp48-RRM12. **D:** GG-30-mer–Unr-CSD12–Sxl-dRBD3 complex titrated into Hrp48-RRM12.

One of the main aims of my PhD project was to determine the structure of the quaternary complex of *msl-2*, Sxl, Unr and Hrp48 to gain insight into the cross-interactions of the different components. This would also extend our knowledge on the function of Hrp48 in the 3' UTR RNP complex. In spite of extensive trials, I did not manage to grow crystals of the complex

and the structure determination of the complex was unsuccessful. This can be explained by the possible flexible nature of the complex. The lack of binding cooperativity between the Sxl-Unr moiety and Hrp48 suggests a beads-on-the string type of arrangement for the complex. In this scenario, Hrp48 would bind to *msl-2* without forming contacts with Sxl or Unr, thereby staying flexible relative to the two other proteins. NMR experiments directed to understand this dynamic character of the complex were also unsuccessful despite my repeated efforts. For NMR studies on molecules and complexes of this size (around 70 kDa), deuterium labeling is needed to achieve good NMR signal-to-noise ratios [206]. Using media for deuterium labeling can have a drastic negative effect on recombinant protein expression levels and sample quality. It was really hard to express Hrp48 in this medium and in the end, after building the complex, I could not measure interpretable spectra because of poor sample quality.

Taken all together, I showed that the structured domains of Sxl, Unr and Hrp48 do not interact with each other in the absence of RNA, but they form a complex along with *msl-2.* The binding of Hrp48 in this complex is not cooperative, compared to the synergistic behaviour of Sxl and Unr. Further efforts are needed to understand this assembly and determine the overall structure of the complex. This would ensure a better understanding of the function of this RNP complex and provide unprecedented insight into translation regulation.

6 Discussion

6.1 RNA binding of the RRMs of Hrp48

In my thesis work, I identified the binding sites for the individual RRM domains of Hrp48 by mapping the region of *msl-2* using 6-mer RNA oligoribonucleotides (Figure 14). For this, I dissected the putative binding motif and additional nucleotides in the downstream region of the *msl-2* mRNA using NMR and ITC titration experiments. This resulted in four different 6-mers, that partially overlap in sequence. With this approach one has to consider the possibility, that slicing of the sequence is not fine enough. This would mean that the part of the sequence important for recognition is not present as a whole in any of the 6-mers. Testing the *msl-2* mRNA-recognition of Hrp48 with a 6-mer in single nucleotide resolution would however define 6-mers that include the palindromic motif and induce double-stranded RNA formation (Figure 31 A). The double-stranded RNA-formation can compromise or completely inhibit RNA-binding. To avoid palindrome formation, only a limited number of 6-mers were possible to design (Figure 31 A). There was no other possibility between 6-mer-1 and 6-mer-2, and so I continued with this resolution for the whole tested region, resulting in the four 6-mers.

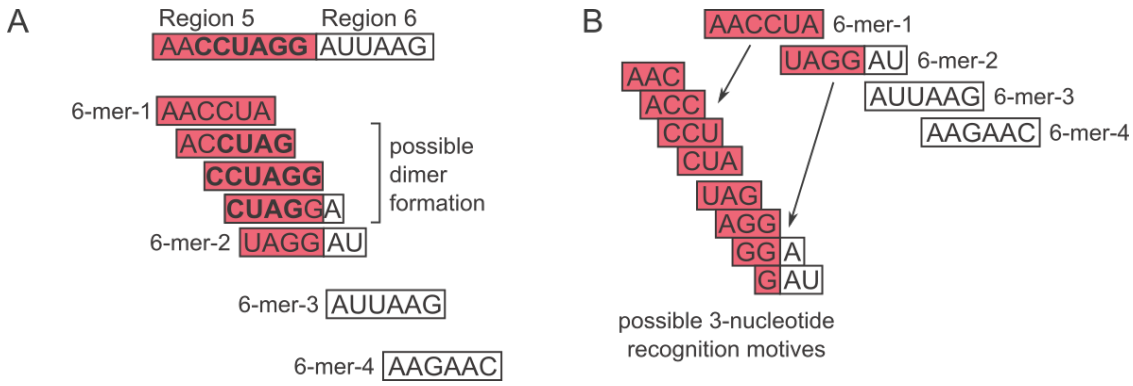


Figure 31 Dissection of the RNA-binding site

A: The number of possible 6-mers is limited by the palindromic sequence element, which needs to be avoided to study the RNA-binding decoupled from RNA conformational change. Motifs that are palindromic are marked with bold letters. **B:** The four 6-mers cover all possible trinucleotides in the region studied for RNA-binding.

On the other hand, it is known that an RRM domain usually only recognizes two-four nucleotides specifically at its RNP motifs and can bind additional ones at the 5' and 3' ends

non-specifically [94]. If we consider all possible motifs of three nucleotides at the tested region, the designed 6-mers cover all variations (Figure 31 B). This method of locating RNA recognition sites within longer sequences has also been utilized for other single-stranded RNA-binding protein domains, such as Unr-CSD1 [68]. If the number of specifically recognized nucleotides can be assumed, as is the case of RRM domains, the number of RNA oligomers used for testing can be reduced. In addition, when I designed the 6-mers, I could have used shorter motifs as 5-mers or 4-mers. However shorter motifs require more RNA oligos to be designed and tested to cover the sequence. Using too long motifs does not give enough information on the location of the binding site.

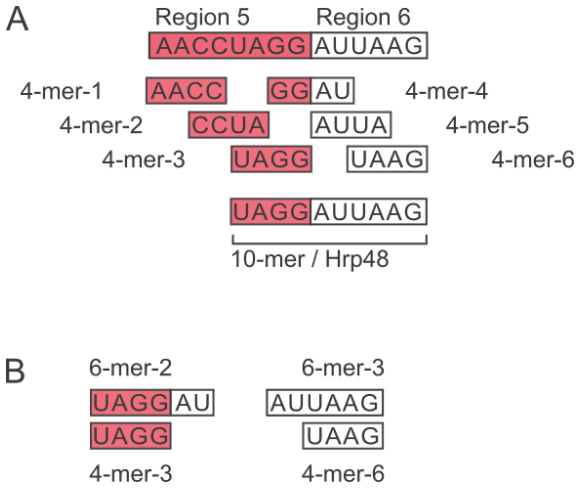


Figure 32 RNA motifs recognized by Hrp48 in the *msl-2* mRNA.

A: The schematic representation of Region 5 and Region 6 probed with the 4-mers for RNA binding motifs of Hrp48 RRM12. 4-mer-3 and 4-mer-6 align with the 3' and 5' ends of the 10-mer. **B:** Putative motifs of RNA-binding within the 6-mer-2 (4-mer-3) and 6-mer-3 (4-mer-6).

With this approach, I identified 6-mer-2 (UAGGAU) and 6-mer-3 (AUUAAG) as the strongest interactors for RRM1 and 6-mer-2 for RRM2. In addition, RRM2 binds 6-mer-3 with ≈20 μM affinity (Figure 17). Based on these findings and the high similarity with the *P-element* binding site, I designed a 10-mer (UAGGAUUAAG) containing both 6-mer-2 and 6-mer-3 sequences as the putatively optimal binding partner for tandem Hrp48-RRM12 (Figure 19). With a similar approach as the one used to identify the cognate 6-mer motifs for the single RRM domains, 4-mers were used to probe 10-mer optimality of sequence for Hrp48-RRM12 binding. Here I showed, that the 4-mer-3 (UAGG) and 4-mer-6 (UAAG) were the strongest interactors. These two 4-mers mark the two ends of the 10-mer (Figure 32 A) and are also present as smaller

motifs within the 6-mers (Figure 32 B). These results, and the high affinity and simultaneous binding of RRM12 to the 10-mer support my assumption that the identified 10-mer has the optimal boundaries for *msl-2* recognition by Hrp48-RRM12.
These findings support the assumption, that RRM12 binds 10-mer close to the 3' and 5' ends, and the middle of the 10-mer is likely not involved in the interaction. Yet it has to be considered that the center part of the 10-mer as a whole construct and not dissected to 4-mers might form protein-RNA contacts. This is unlikely based on the NMR relaxation and RDC data, which suggest flexibility between RRM1 and RRM2. If the whole 10-mer including the center is bound in the complex, the two domains could not have such a big flexibility relative to each other. This could only be tested with mutated or isotopically labelled RNA and overall structure determination of the complex. Based on the data presented, I would claim that the UAGG and UAAG motifs at the 3' and 5' ends are specifically recognized by Hrp48-RRM12.

As mentioned previously, the Hrp48 RRM12 binding to *msl-2* results in the formation of a flexible complex. I also hypothesized that potentially one or two of the central nucleotides of the 10-mer barely interact with Hrp48. However, it could be that this is only one of different Hrp48 RNA-binding modes. For example, the *P-element* binding site (**UAGG**U**UAAG**) is missing one residue in the middle, when comparing to the 10-mer (**UAGG**AU**UAAG**). There is no dynamical information available of the Hrp48–*P-element* complex in the literature, but it seems likely that this complex would be more rigid than the one in the *msl-2* system, as the shorter RNA would draw the two RRM domains closer. Based on this, it could be hypothesized that the linker between the 3' and 5' motifs could be longer or shorter, adding further sequences that Hrp48 can bind. In addition, different linker lengths can have a direct effect on the binding affinities as well, as it has been shown previously [86]. Further studies with various RNA constructs and increasing the pool of known interacting RNA sequences by pull-down assays would take a step further in understanding the RNA binding of Hrp48 and possibly other multi-domain RNA binding proteins.

It also has to be noted, that all RNA-binding studies were performed in MES buffer, that contains 200 mM NaCl. Lower salt concentrations lead to instability for all constructs (RRM1, RRM2 and RRM12) at least at the concentrations used in my experiments. In order to retain consistency across the different ITC, SEC-MALLS and NMR studies, I did not change the buffer composition. The high salt content in the buffer can influence RNA binding especially in the formation of salt bridges and other polar interactions, because high ionic strength of the

buffer can compromise RNA-protein interactions. For example, it has been shown for the human homologue, hnRNP A1 in salt sensitivity studies, that about 30% of the binding Gibbs free energy comes from ionic interactions [207]. This could imply that the K_D values of Hrp48 are apparent K_D values with a masking effect of the high ionic strength, and the K_D values at physiological conditions are lower than I measured.

6.2 Non-canonical RRM folds of Hrp48 and their role in RNA-binding

The AF2 model supported by the TALOS+ prediction revealed that RRM2 exhibits a non-canonical RRM-fold with the additional β3'-strand between the α2-helix and the β4-strand, which is stabilized by hydrogen bond contacts between residues R157 and E167 and additional H-bonds between β3' and β4 (Figure 16 C and Figure 33 B). The CSP data suggests that the β3'-strand, that includes an aromatic residue (Y158) is involved in RNA binding (Figure 17). Hrp48 RRM1 in the RNA-bound state might also form this β3'-strand in a smaller population, based on the MD simulation (Figure 15 B and Figure 33 C). In these frames β4 is elongated in the N-terminal direction, which might be stabilized by the formation of the β3'. Although there is no direct contact formed between the RNA and residues of β3', R75 within the elongated part of β4 interacts with the RNA (Figure 33 C). This non-canonical RRM fold has also been observed in numerous other proteins along with its role in RNA-binding [80-81, 94]. Sometimes an additional β3''-strand is present as an N-terminal extension of the β4 parallel to the β3' (Figure 33 E). In the case of Hrp48 RRM1 and RRM2, this additional β3'' was not observed, but the β4 is elongated in the N-terminal direction in RRM1 (Figure 33 C). For example, the RRM2 domain of hnRNP A1 exhibits the same β3'-strand and elongated β4, but there are no contacts formed between the RNA and β3' or β4 (Figure 33 D) [93]. The human Fox-1, an RNA-binding protein regulating RNA-splicing, exhibits both the additional β3' and β3''-strands in an NMR structure (PDB: 2ERR), and the β3'' is in direct contact through its R184 residue with the bound RNA-motif (Figure 33 E) [96]. These examples show that the formation of the β3' and β3'' strands or the elongation of β4 can extend the RNA-binding interface, but in other cases it is not necessarily involved in RNA-binding. In fact, they can also be involved in protein-protein interactions, such as for the RRM domain of eIF3b, for which a direct interaction with the eIF3j peptide through the β3' and β3'' strands is strongly suggested based on CSP data [104].

These are just a few examples on the wide range of structural and interaction diversity of the RRM domains and this enormous scope of variation is, what allows this highly abundant motif to perform its countless biological RNA-related functions [94].

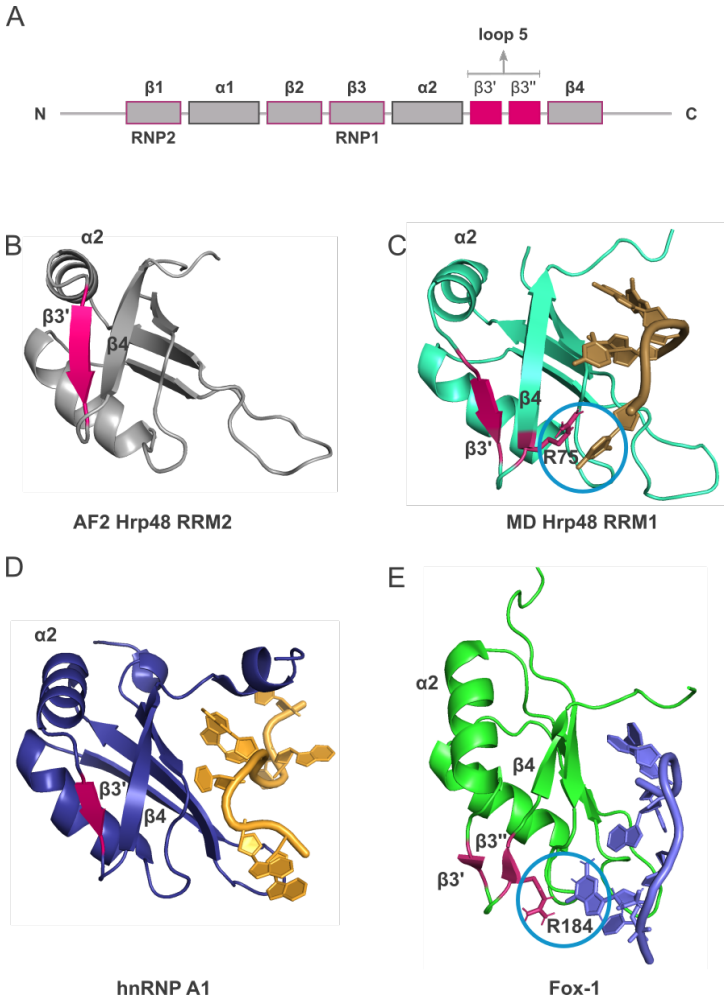


Figure 33 Non-canonical RRM folds of Hrp48

A: Domain arrangement of RRM domains with the placement of the β3' and β3''-strands (highlighted in pink). **B-E:** Uncanonical β3' and β3''-strands visualized in different RRM-structures highlighted in pink. **B:** AF2 prediction of Hrp48 RRM2 with the β3'-strand highlighted. **C:** An MD frame of the Hrp48 RRM1 – 6-mer-3 simulation exhibiting a β3'-strand. The blue circle highlights the interaction between R75 and the RNA. **D:** Crystal structure of hnRNP A1 in complex with pri-miRNA-18a with the β3'-strand highlighted (pdb: 6dcl). **E:** NMR structure of Fox-1 in complex with UGCAUGU RNA with the β3' and β3''-strands highlighted (pdb: 2err). The blue circle highlights the interaction between R184 and the RNA

6.3 Self-association and RNA binding of Hrp48

I have found that the self-association tendency of Hrp48 is possibly mediated by its RRM2, and that its self-association and the RNA-binding properties might be linked. To test whether this was possible, I compared the RNA-binding and tetramerization interfaces of RRM2, and performed SEC-MALLS experiments with Hrp48-RRM12 and RNA (Figure 18). These results confirmed the tetramerization, which might hinder RNA binding due to the potential burial of the RRM2 RNA-binding interface. The tetramerization tendency might also be involved in the formation of higher oligomeric assemblies and eventually phase separation of Hrp48, as it has been shown in the context of *oskar* mRNA localization [121]. Additional studies directed to understand the link between RNA binding and self-association could shed light on a potential mechanism with functional relevance. Mutational analysis of Hrp48-RRM2 could be used to test whether the RNA-binding and tetramerization can be decoupled from one another. Mutants could be tested by SEC-MALLS and ^{1}H, ^{15}N-HSQC NMR for changes in the self-association tendency and for RNA-binding by NMR titrations. One mutant would need to have impaired RNA-binding, while the other would lose the tetramerization function only, to decouple RNA-binding and tetramerization. The two mutants could be subsequently assessed by reporter assays, to test if RNA-binding alone, or tetramerization is sufficient for translational repression activity.

The relationship between self-association and RNA-binding has been studied for other systems as well. The Human antigen R (HuR) consists of three RRM domains and recognizes AU-rich RNA elements. RRM1 and RRM2 are connected by a short linker and RRM3 is attached by a long flexible linker to the tandem RRM12 moiety. All three RRMs engage in RNA-binding, which promotes compaction [208]. The HuR-RRM3 domain undergoes concentration-dependent dimerization mediated by its α1-helices, and the dimer formation extends the RNA-binding interface and increases its RNA-binding affinity [105, 208], which is crucial for its biological function [208]. The cytoplasmic polyadenylation element (CPE) binding protein exists as four different paralogs, of which CPEB1 is the most diverged from the other three (CPEB2-4) [209]. All of the variants consist of two conserved C-terminal RRM domains responsible for the interaction with the CPE nucleotide sequence. It has been shown that CPEB1 dimerizes through its RRMs, thereby forming an inactive pool of CPEB1. Dimerization prevents the RNA-binding of CPEB1 and thus fulfills its role in translational control [210]. The authors also suggested that the dimer formation fine-tunes the amount of available monomeric and active

CPEB1 levels [210]. In another study, it was found that the other paralog, CPEB4, also tends to dimerize. Both the RRM1 domain alone and the tandem RRM12 construct can form dimers, however this only represents a small population of the protein in solution. Interestingly, mass spectrometry analysis of RRM1 and RRM12 demonstrated no cross-interaction between RNA-binding and dimerization, and suggested that the RNA-binding and dimerization interface are separate for CPEB4 [106].

6.4 Entropy-enthalpy compensation in ITC

Based on the ITC titration experiments with the different 6-mers of *msl-2* and RRM1, I determined a comparable K_D for all four 6-mers (Figure 14 J and Figure 34). However, these results are not in agreement with the NMR titration experiments, that showed varying binding strength for the different 6-mers (for 6-mer-1 and 6-mer-4 weaker binding than for 6-mer-2 and 6-mer-3). Furthermore, for the ITC data I observed different enthalpic (ΔH) and entropic ($T\Delta S$) contributions to the Gibbs free energy change of the binding ($\Delta G = -RT \ln(K_D)$). For the 6-mers that showed weaker binding in the NMR titration experiments (6-mer-1 and 6-mer-4), the enthalpy change was small and the entropic contribution was moderate, whereas for the stronger binders (6-mer-2 and 6-mer-3) the heat change is much greater and it is compensated by a significant entropy change. The largest heat release was observed for 6-mer-3. This could be explained by a different binding mode for the different 6-mers.

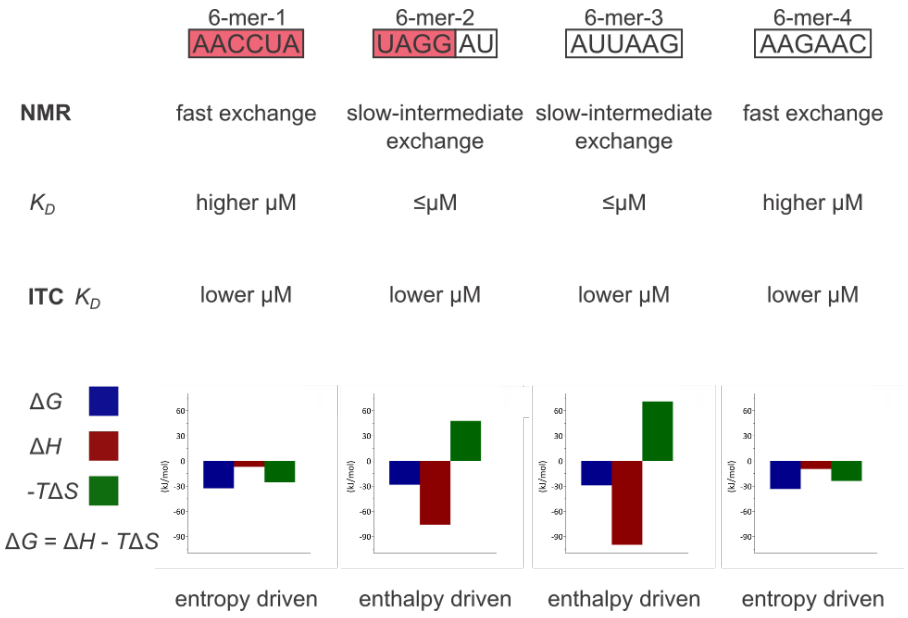


Figure 34 Binding affinities and thermodynamic contribution of RNA binding of RRM1

The outline of the results of the NMR titrations ITC experiments are shown. In the bottom part of the figure the thermodynamic contributions of the binding events are represented on heatplots with identical scaling for all 6-mers. Blue: Gibbs free energy change, red: enthalpy change, green: entropy change.

A similar phenomenon termed as entropy-enthalpy compensation is frequently observed in calorimetric studies of proteins with ligands, such as SAR (Structure-activity relationship)

screenings [211]. In these cases, either the small molecule ligand is functionalized differently or the protein is changed by mutations, but these modifications are not reflected in the ITC binding affinity and so the Gibbs free energy changes. For example, introduction of a hydrogen bond partner in the ligand could increase the enthalpy change, while it is compensated with an entropic penalty, and there is no overall change in the affinity, in spite of the additional intramolecular bond. This was observed in the case of a potent HIV protease inhibitor where changing a thioether functionality by a sulfonyl group caused no net change in the affinity, albeit the formation of a new hydrogen bond [212]. In another study, different HIV protease mutants were tested, including one with drug resistance. Here, an entropy-enthalpy compensation was observed with several approved HIV protease inhibitors, suggesting critical changes in the ligand recognition mode [213]. This phenomenon creates a difficulty in drug discovery, as it makes it hard to distinguish between the actual potency of the tested compounds based on the direct assessment of the K_D.

What causes this compensatory effect? The formation of more or stronger van der Waals contacts and H-bonds upon changes in the ligand functional groups leads to an increased magnitude of the negative enthalpy change. Formation of additional bonds can in return lead to decreased flexibility of both the protein and the ligand, and consequently to a reduced conformational entropy [211, 214]. Another potential factor is the water solvation. While water molecules bound to structured and rigid parts of the protein and ligand are dynamically constrained, dynamics of water molecules bound to flexible parts are more similar to bulk water molecules [214]. The more dynamically restricted the solvating water molecules are, the smaller the solvation entropy becomes. These two factors of entropy-enthalpy compensation have been extensively studied by NMR relaxation experiments, X-ray crystallography, and molecular dynamic simulations [211, 215].

In my RNA-titration experiments the change between the tested ligands were substantial, when compared to changing a single functional group in a small molecule ligand. Still, ITC was not able to resolve the differences in binding on the level of the K_D values while I observed significant differences in the NMR titration experiments. In line with the idea behind the entropy-enthalpy compensation, I assumed that the stronger binder 6-mer-2 and 6-mer-3 interact with RRM1 specifically, thereby forming more intramolecular interactions resulting in a substantial heat release, while the conformational space of the complex is reduced. The weaker binder 6-mer-1 and 6-mer-4 still interact with RRM1, but the binding is less specific and the formed complexes remain more flexible. It would be interesting to measure NMR

relaxation and determine conformational entropies of the different 6-mer complexes to see whether there is a correlation with the ITC results.

6.5 Complex formation at the 3' UTR of *msl-2*

In order to optimize the complex formation of Hrp48, Sxl, Unr on the *msl-2* mRNA, SEC–MALLS experiments were used with different RNA and protein constructs. The results suggest the duplex formation of the WT-30-mer, that contains the palindromic motif. Addition of Hrp48 unwinds the duplex and the monomeric complex can form (Figure 29 D). The duplex formation is possibly an artificial effect of my *in vitro* experimental procedure and high sample concentrations and may not have a biological relevance. This could be confirmed by *in vitro* translation assays with the GG and UU mutants (Figure 35). This allows testing the effect of both purine (GG) and pyrimidine (UU) mutations. As these modifications do not affect translation repression activity notably (Figure 35), the duplex formation is possibly not relevant biologically. It is only complicating the biophysical studies at the concentrations I used.

Another interesting aspect of the *in vitro* translation assays is that the 5m mutant, the CU-substitution of the initially proposed Hrp48 binding site (region 5) has a somewhat greater de-repressing effect than the 10m, the CU-substitution of the Hrp48 binding site I identified (Figure 26 A and B). This could be explained by the fact, that a part of Region 5 is not just important for Hrp48 binding, but for additional protein-RNA interactions. The RRM1 domain of Sxl dRBD3 has a reported contact with the first adenosine of Region 5, thus 5m has an effect on both Unr-Sxl binding, and Hrp48 binding (Figure 35 B and C). It also cannot be excluded, that the few nucleotides between the Sxl-Unr and the Hrp48 binding sites are involved in additional interactions, because the reporter assays in the embryo extracts involve the full-length proteins expressed in *D. melanogaster* cells, except for the recombinant Sxl dRBD4. Thus, additional factors or protein regions present in the reporter assays can affect the translation activity, which is hard to address based on the biophysical and structural data.

I showed in this study that Hrp48-RRM12 forms a flexible complex with the 10-mer RNA and that the binding of Hrp48 to the *msl-2* mRNA is not cooperative with Sxl and Unr. In addition, the structured domains of Sxl, Unr do not interact with Hpr48-RRM12 in the absence of RNA. These suggest an overall flexible complex formation in which Hrp48 binds in 3 nucleotides distance downstream to Sxl and Unr on the *msl-2* mRNA.

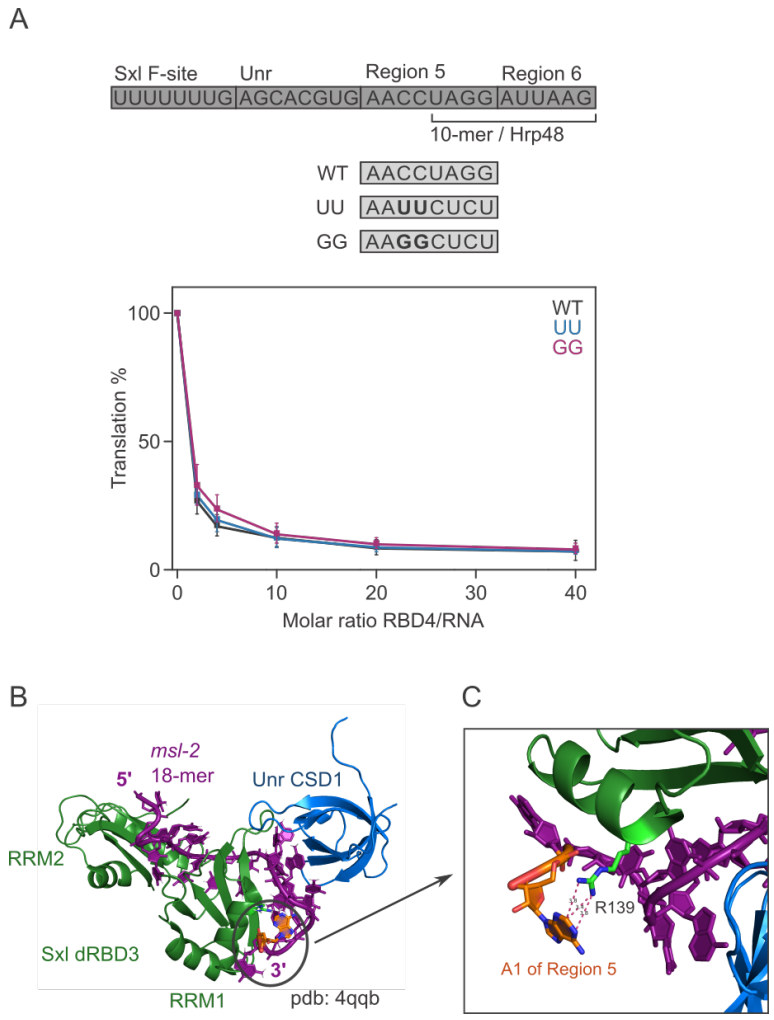


Figure 35 Discussion of the reporter assays

A: Schematic representation of the mutations used to cancel the effect of dimerization. **B:** Crystal structure of the ternary complex of *msl-2* – Sxl-dRBD3 – Unr-CSD1, pdb: 4qqb [68] **C:** R139 of Sxl-dRBD3 directly interacts with the first nucleotide of Region 5.

Clearly, it has to be considered here too that not the full-length proteins were used in this study. Extended protein regions of Sxl, Unr or Hrp48 not involved in my experiments could serve additional functions in RNA recognition, protein – protein interaction and stabilization of the RNP complex. It has been shown, that while dRBD3 (aa 123 – 294, Figure 7 B) is sufficient for the RNA-binding activity of Sxl, the C-terminal elongated dRBD4 (aa 122 – 301, Figure 7 B) is the minimal construct for the optimal translation repression activity [63]. Preliminary smFRET data of Marco Payr in our laboratory suggest that a C-terminal extended construct stabilizes more the RNP complex formation with Unr-CSD1 on the *msl-2* 3' UTR than dRBD3. He used dRBD5, a construct even longer than dRBD4, which also contains a C-terminal

putative α-helix based on AlphaFold predictions. Additional secondary structure elements such as this putative α-helix could account for extended number of interactions, stabilizing the complex formation. Additionally, it is known that Unr-CSD1 (aa 182 – 252, Figure 7 B) is sufficient for the high-affinity interaction and complex formation with Sxl and *msl-2*. However, for the full translation repression activity, a much longer construct is needed, enclosing the N-terminal Q-rich domain and the first three CSD domains (aa 1 – 397) [112]. These findings indicate, that for both Unr and Sxl, the constructs I used for complex formation (Sxl-dRBD3, Unr-CSD1 and CSD12) do not account for the whole biologically relevant regions. For Hrp48, there were no studies conducted to date directed to the understanding of the role of different domain boundaries on the translational repression activity.

It has been shown, that the SG-enriched low complexity C-terminal domain of Hrp48 is important for the phase separation in the *oskar* mRNA localization [121]. Here, along with Bruno, Hrp48 acts as a scaffold protein and induces granule formation which undergoes a liquid-to-solid phase transition. Both Sxl (GN-rich) and Unr (Q-rich) have domains that could induce aggregation too [216]. This means it cannot be excluded, that phase separation also plays a role in the *msl-2* translational repression, and this question would have to be addressed in further studies.

Taken together, in my thesis work I could show that along with Hrp48-RRM12, the short constructs of Sxl (dRBD3) and Unr (CSD1 and CSD12) do form a complex on the *msl-2* 3'UTR. It would be interesting to test the extended protein constructs in complex formation experiments. The additional regions and protein domains might help to stabilize the complex, and it would also increase the molecular weight paving the way for structure determination by electron microscopy. Alternatively, cross-linking or native mass spectrometry [217], complemented with small-angle scattering experiments could give an insight into the arrangement of the 3' UTR regulatory RNP complex.

7 Conclusions and outlook

In this work, I have presented the structural and biophysical characterization of the RNA-binding protein Hrp48 alone and in context with the *msl-2* 3' UTR RNP complex.

The structure of the RRM1 domain of Hrp48 has been solved, and the AlphaFold2 structure of RRM2 has been validated based on my NMR data. The binding motifs of the single RRM domains and the tandem construct of Hrp48 have been discovered and used for structural modelling of the protein-RNA complex structures. Hrp48 RRM12 forms a 1:1 complex with *msl-2* mRNA in solution and both domains are involved in RNA binding. Studies directed to understand the dynamic behaviour of Hrp48 show that in the apo-state the RRM domains tumble independently, and upon RNA binding the domains remain flexible with respect to each other. The biological relevance of the identified *msl-2* binding site has been confirmed in the cellular context. I also showed, that the structured domains of Sxl, Unr and Hrp48 do not interact with each other in the absence of RNA, but they form a complex along with *msl-2*. The binding of Hrp48 in this complex is not cooperative, compared to the synergistic behaviour of Sxl and Unr, that had been reported previously [68].

Modulating the translation efficiency contributes to myriads of biological processes and is often orchestrated by trans-acting factors binding to autoregulatory mRNA sequences at the UTRs. Hrp48 acts as a bifunctional modulator, examples show its involvement in both translational enhancement (*Hsp83*) and translational repression processes (*msl-2*), in concert with other RNA-binding factors [72, 139]. It is remarkable how the same protein can perform in both roles. The current hypothesis is that the function is decided based on the additional cis- and trans-acting factors involved in the regulative process and Hrp48, Unr, Sxl and pAbP modulate the function of one another to perform in *msl-2* translation repression jointly. In order to understand the interplay between these factors of this highly interesting model system, further studies are needed to decode their interaction on the mechanistic and cellular levels. Particularly interesting is the contact with the translational machinery through Hrp48, which also remains to be targeted in additional studies. Understanding the underlying mechanisms of the *msl-2* translational repression would give an unprecedented insight into translation initiation control.

8 Acknowledgements

As a chemist entirely new to structural biology, I came a long exciting and challenging way to learn about the theory, the methods and different other aspects of this field. I would like to thank everyone I mention in the Acknowledgements to support me on my journey because all of them contributed to my PhD work.

First and foremost, I am thankful to my thesis supervisor Janosch Hennig for providing me the opportunity to join his lab and for all the effort he has put into this project. I am grateful for all the theoretical and practical knowledge that he has transferred to me, and for staying enthusiastic all the way through my thesis work.

I would also like to thank my TAC committee members, Matthias Hentze, Julia Mahamid and Rebecca Wade for taking their time to discuss and give thoughtful input on my project during the annual meetings. I would also like to thank Stefan Pfeffer for reading my thesis and participating in my thesis defence committee.

I am grateful to Bernd Simon for all the insightful practical help and advice on several NMR technical problems and data analysis issues.

I would also like to thank the different core facility members for their help in my experiments. Special thanks to Karine Lapouge for the extensive support on the biophysics side and the constructive discussions on my project, and for being a great person to be around. I am also very grateful to Brice Murciano who set up numerous crystallization screens and also provided amusing company.

This work would not have been possible without our collaborators. Many thanks to Miroslav Krepl at the Jirí Šponer lab for the MD simulations on the complex structures. I am also grateful to Tanit Guitart at the Fátima Gebauer lab for performing the reporter assays.

I would like to thank to the current and previous members of the Hennig Group for guiding me into the world of molecular and structural biology, which has not always been a straightforward journey. I am especially grateful to Cecilia, Lucia, Marco, Pravin, Kevin and Tamara for the valuable professional remarks on my thesis work and for the family-like and highly inspirational working atmosphere that they have provided. Special thanks to Cecilia and the two Kevins for proofreading my thesis.

I could have not undertaken this journey without my closest friends, Kevin and Léa, who supported me on all levels during the ups and downs of navigating my project, writing this thesis and personal life. I would also like to thank Alina, Ana, Marina, Jesus, Lucia, Anna and Michelle, who supplied a warm, entertaining and encouraging background during the course of my years of PhD and through the process of research and writing this thesis. Because of you, I enjoyed these four years at Heidelberg a lot and this place could really become my home.

Finally, this endeavor would not have been possible without my loving family, especially my Mum, who provided me with constant support through all peaks and valleys of my PhD. Thank you.

9 Publications

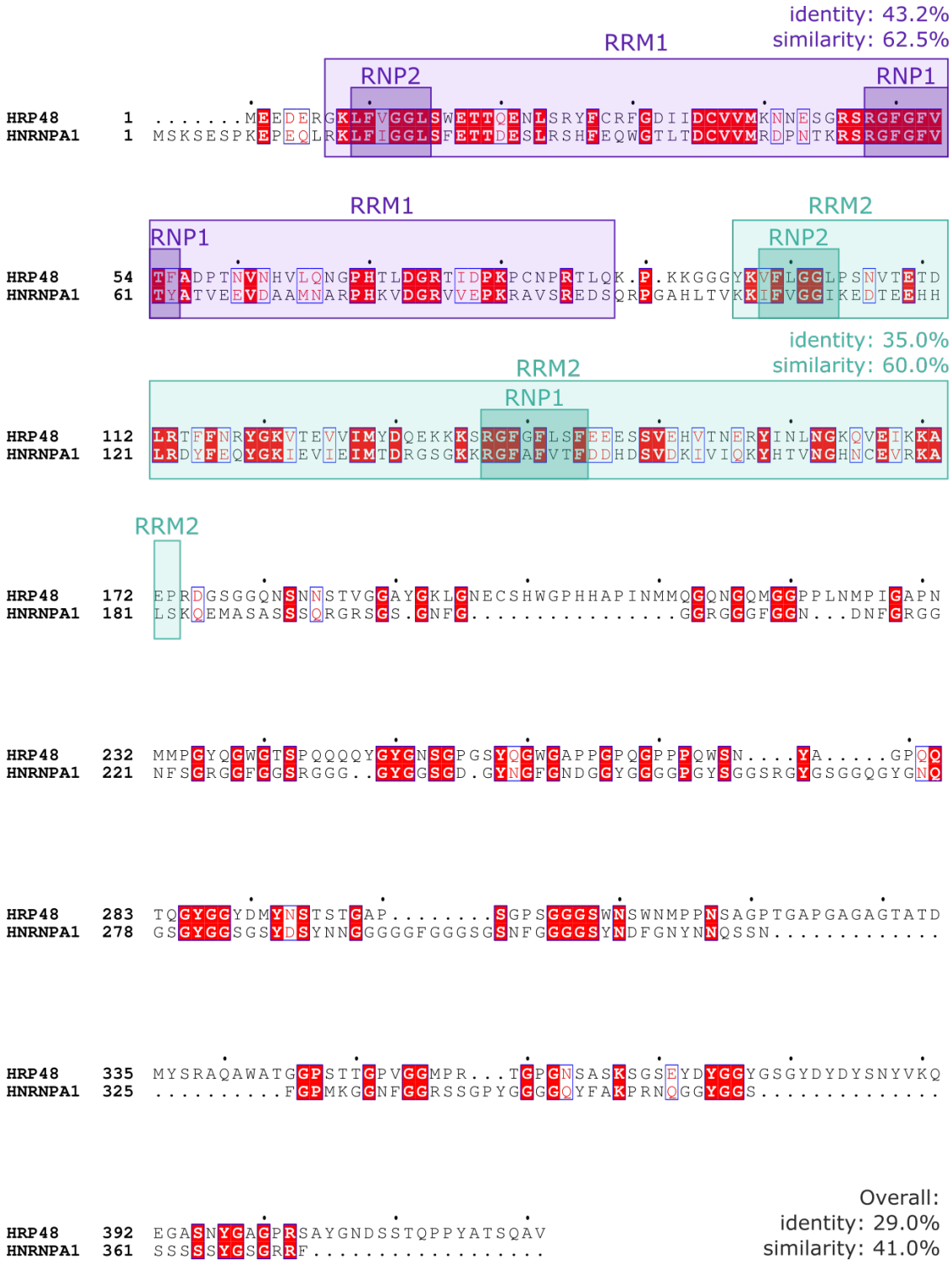
Results of my undergraduate research, not related to the current thesis.

[1] Z. Szakács, Z. Sánta, A. Lomoschitz, C. Szántay, Self-induced recognition of enantiomers (SIRE) and its application in chiral NMR analysis. *TrAC Trends in Analytical Chemistry* **2018**, 109, 180-197.

[2] D. V. Horváth, F. Domonyi, R. Palkó, A. Lomoschitz, T. Soós, Regioexhaustive Functionalization of the Carbocyclic Core of Isoquinoline: Concise Synthesis of Oxoaporphine Core and Ellipticine. *Synthesis* **2018**, 50, 2181-2190.

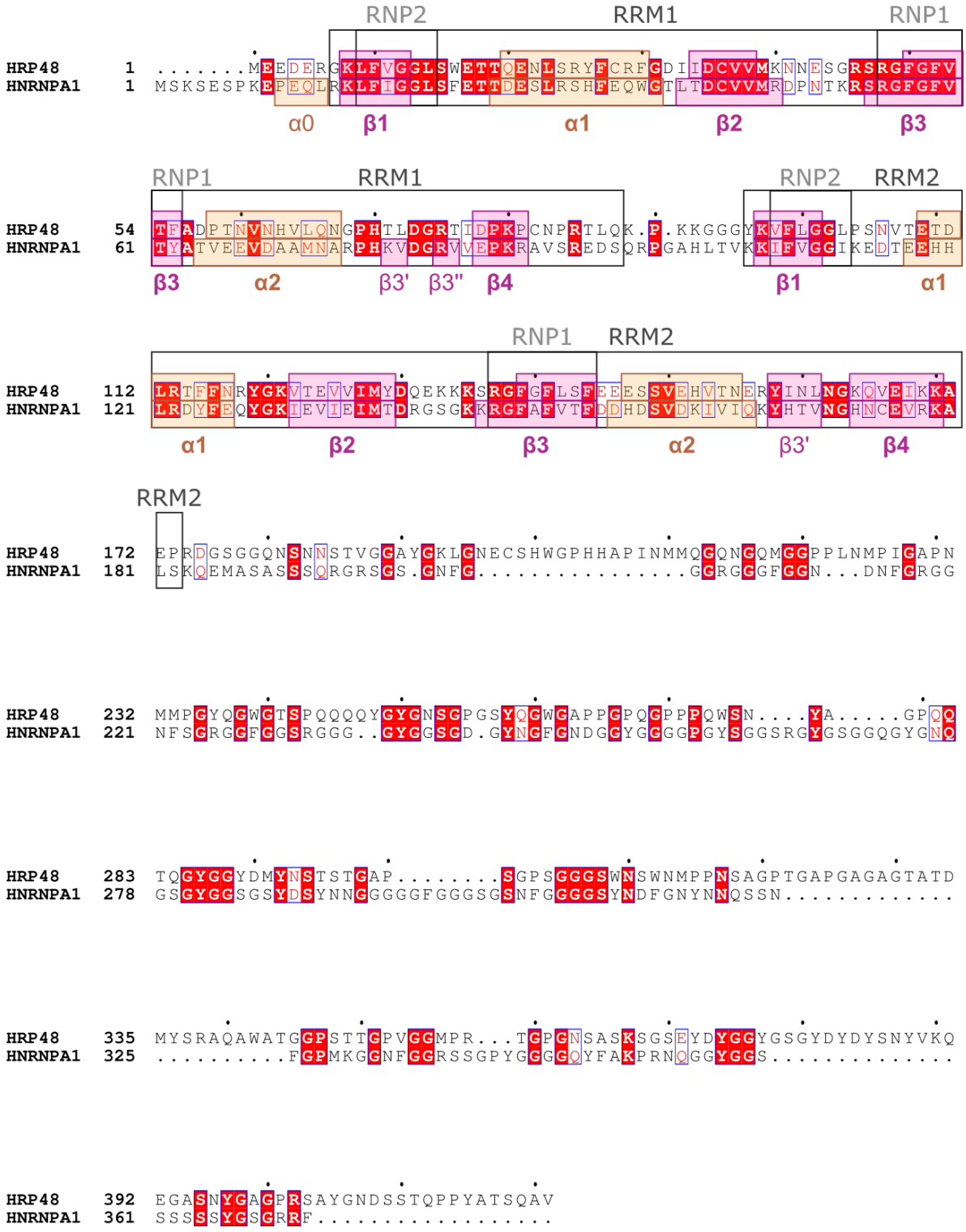
10 Appendix

11 Supplementary Figures



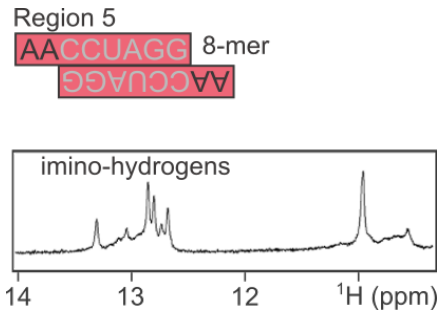
Supplementary Figure 1 Sequence alignment of Hrp48 and hnRNP A1

The identical residues between the sequences are highlighted in red, the similar amino acids are indicated with red fonts. The RRM domains are marked by purple (RRM1) and green (RRM2) boxes, and the conserved RNP-sequences with darker purple and green respectively. The sequence similarity and identity was calculated with Emboss Needle [218]. The figure was prepared using Clustal Omega [218] multiple sequence alignment tool and ESPript [219] for visualization.



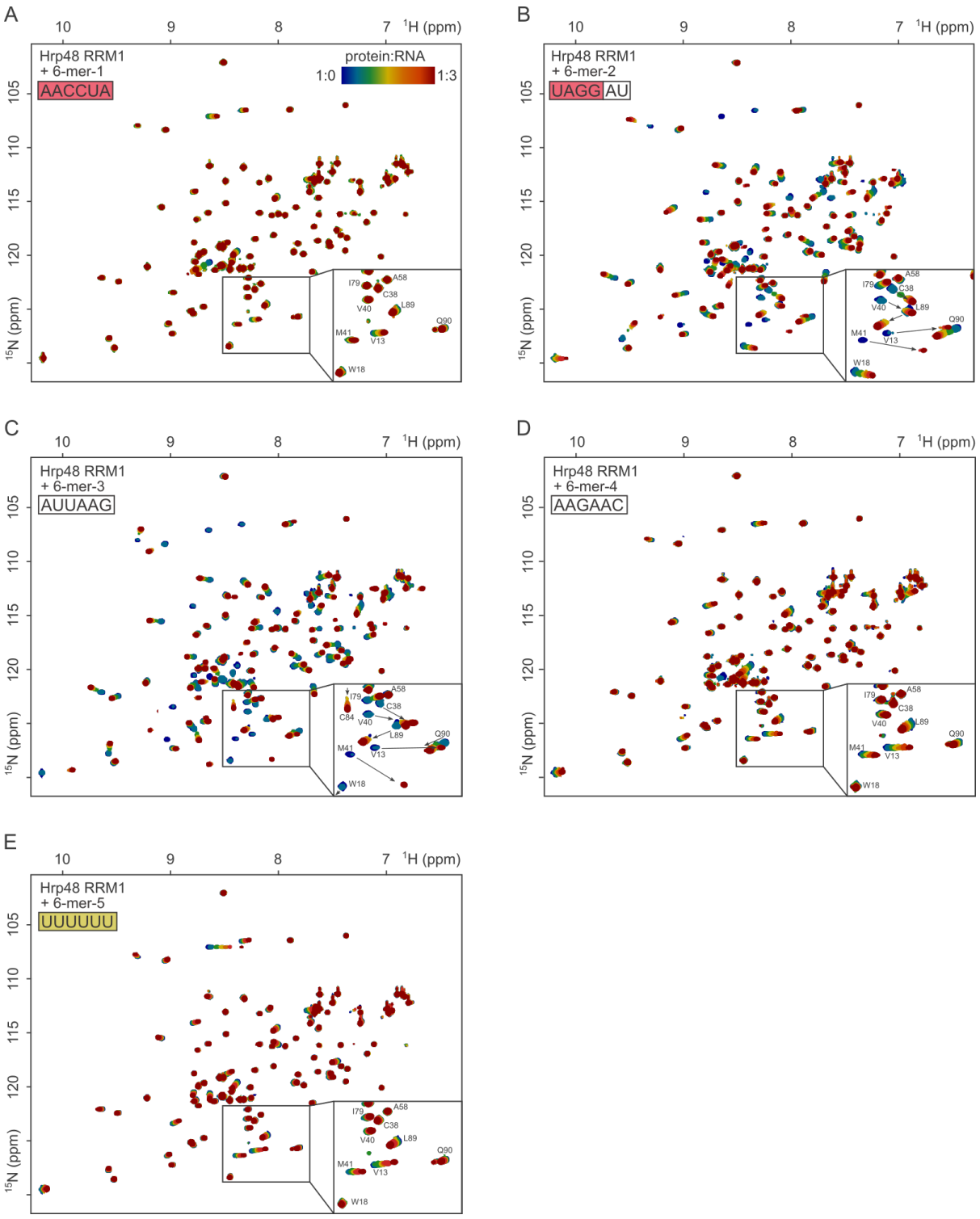
Supplementary Figure 2 Sequence alignment of Hrp48 and hnRNP A1 with secondary structures

The identical residues between the sequences are highlighted in red, the similar amino acids are indicated with red fonts. The secondary structure elements are marked as follows: α-helix, orange β-strand, pink. The figure was prepared using Clustal Omega [218] multiple sequence alignment tool and ESPript [219] for visualization.



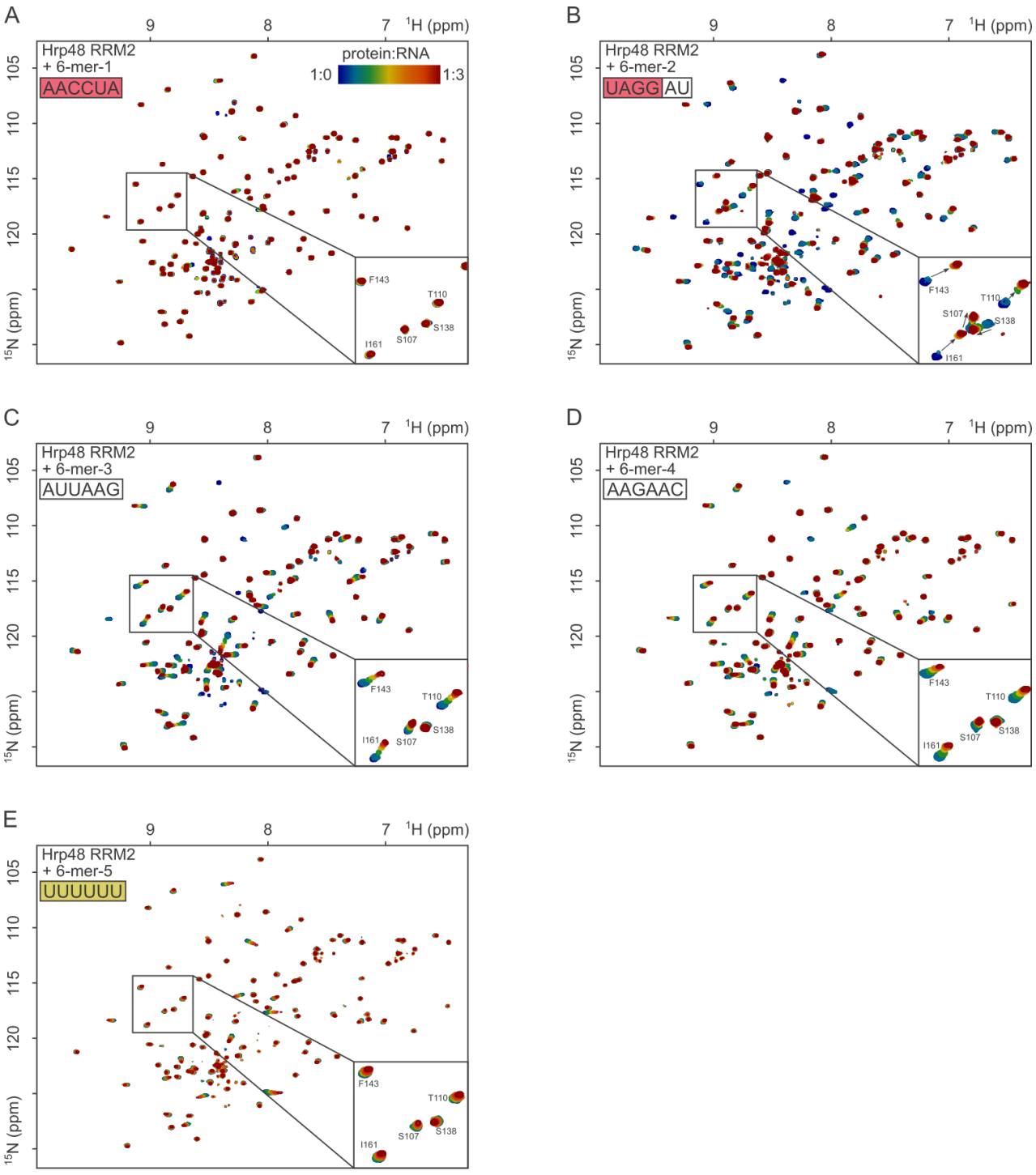
Supplementary Figure 3 Self-association of the palindromic sequence

Imino-hydrogen region of the one-dimensional ^{1}H-NMR spectrum of the 8-mer RNA oligomer (Region 5).



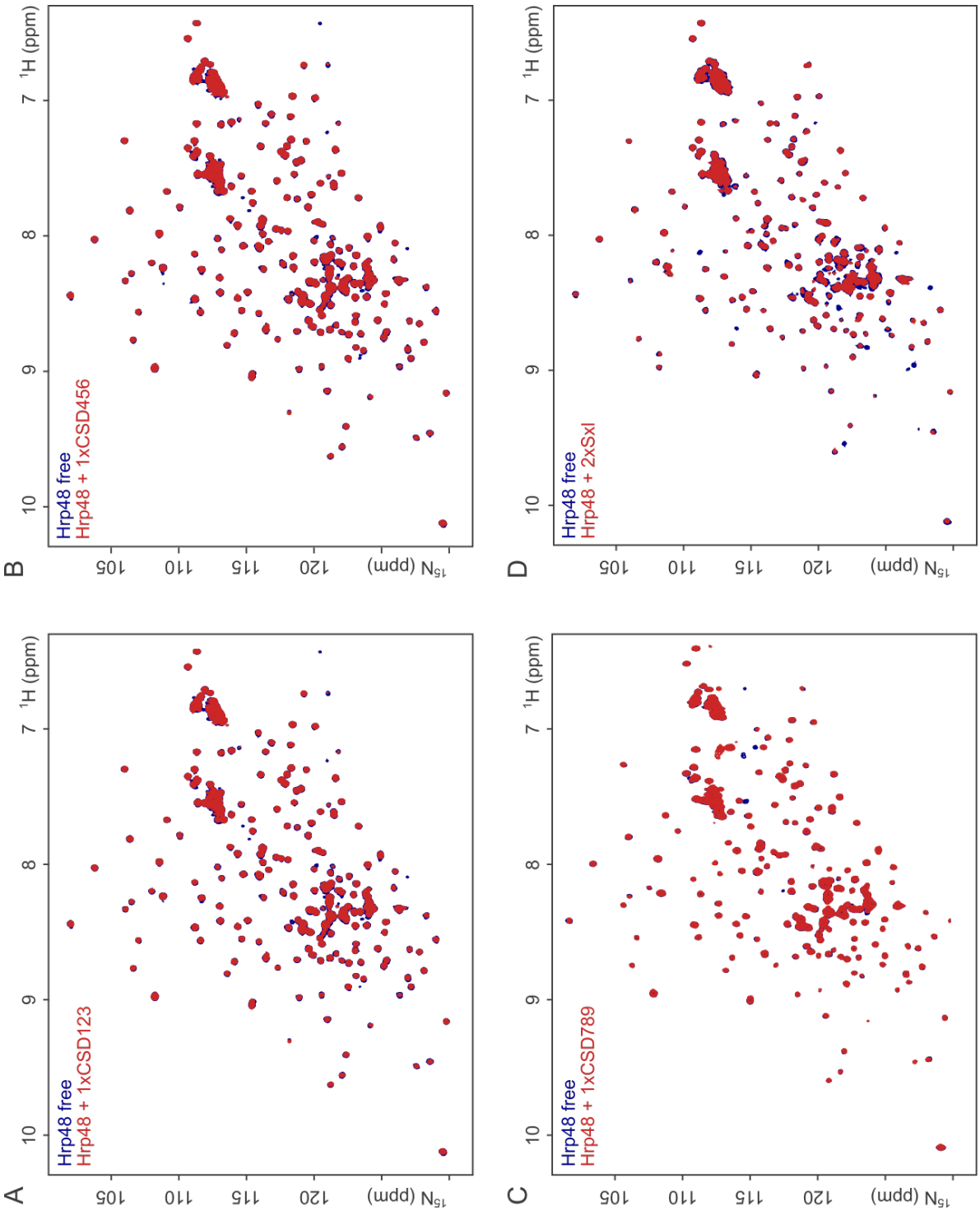
Supplementary Figure 4 Complete $^{1}H,^{15}N$-HSQC titration spectra for Hrp48-RRM1 interaction experiments with 6-mers.

Zoomed in regions are the same as on Figure 14. The coloring scheme indicated on A is the same for each titration. E 6-mer-5 is a control of which the sequence is not related to the other 6-mers. RRM1binds this motif, albeit with weak affinity.



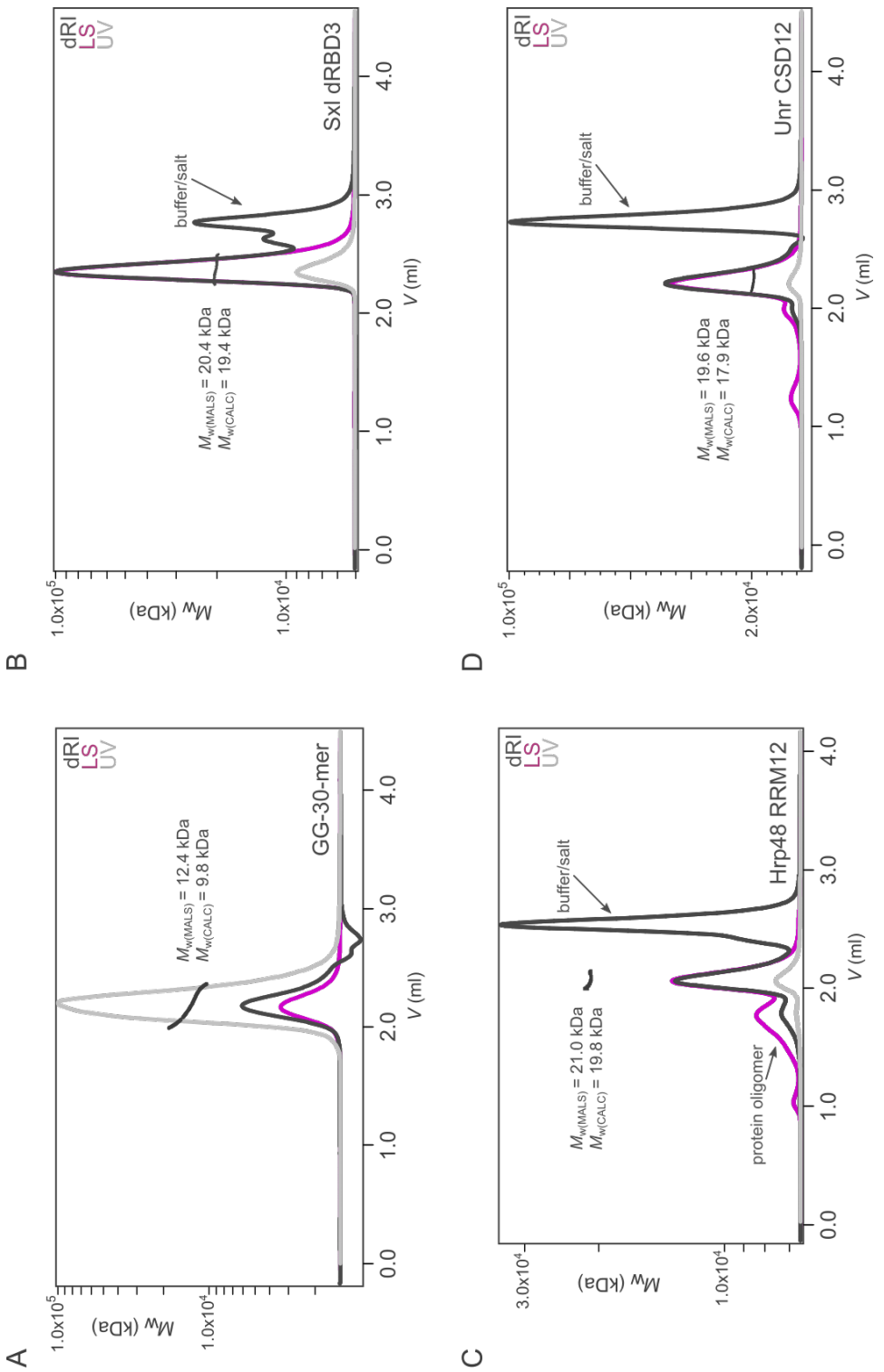
Supplementary Figure 5 Complete ^{1}H,^{15}N-HSQC titration spectra for Hrp48-RRM2 interaction experiments with 6-mers.

Zoomed in regions are the same as on Figure 17. The coloring scheme indicated on A is the same for each titration.

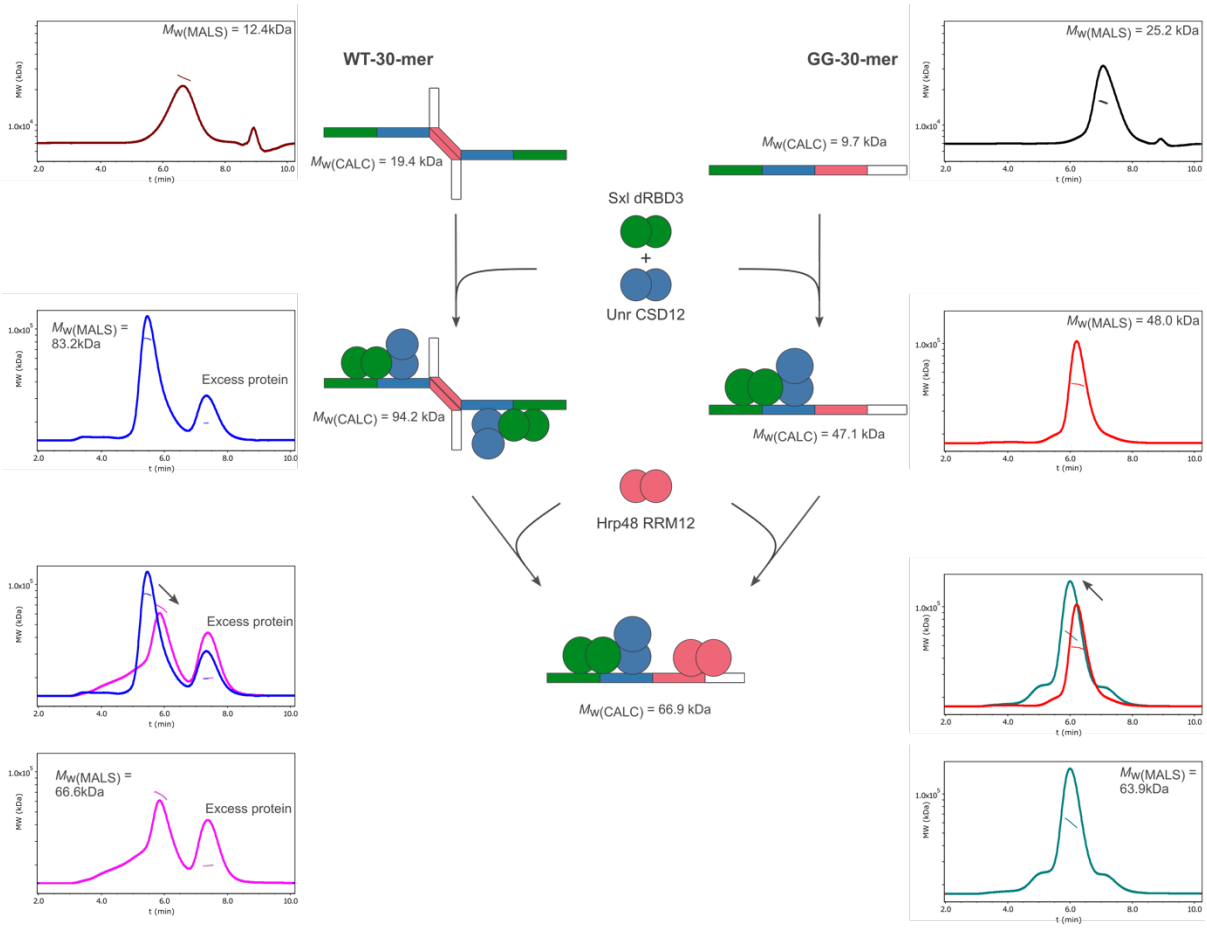


Supplementary Figure 6 Interaction tests by NMR without RNA

A-D ^{1}H,^{15}N-HSQC NMR spectra of ^{15}N labelled Hrp48-RRM12 free protein (blue) and Hrp48-RRM12 mixed with an unlabeled Sxl or Unr construct (red).

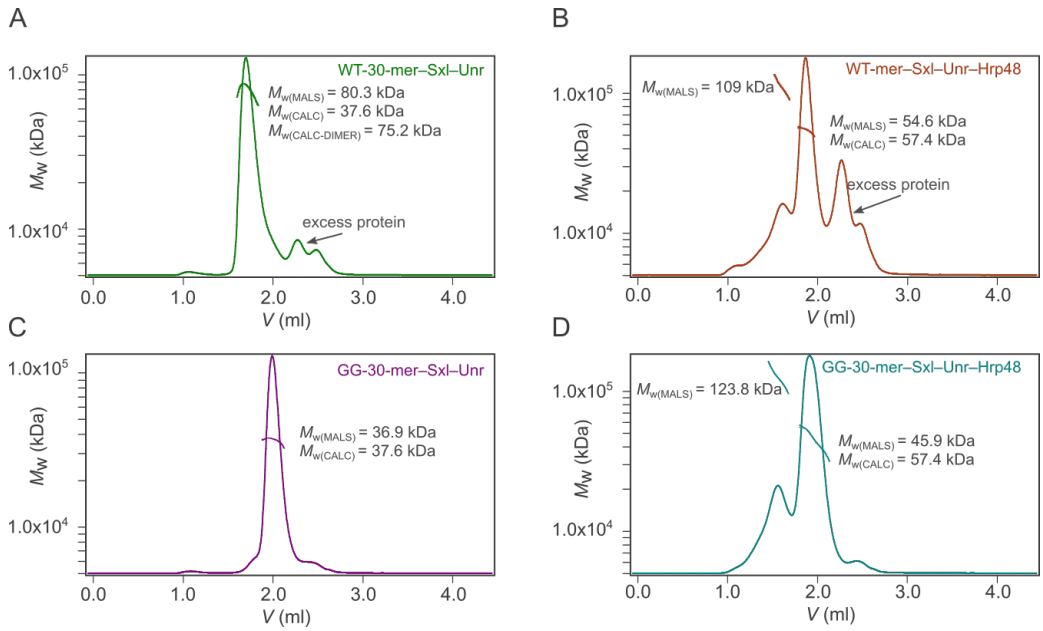


Supplementary Figure 7 SEC-MALLS chromatograms of proteins used for complex formation.



Supplementary Figure 8 Steps of complex formation with the WT and GG 30-mer

Light scattering SEC-MALLS chromatograms of the steps of complex formation and the schematic representation of the complexes in between the steps.



Supplementary Figure 9 Complex formation with CSD1

A-D: SEC-MALLS light-scattering chromatograms of the different complexes with Unr-CSD1, WT-30-mer and GG-30-mer, Sxl-dRBD3 and Hrp48-RRM12.

12 Supplementary Tables

Supplementary Table 1 Data collection and refinement statistics of Hrp48-RRM1

	Hrp48_RRM1
Wavelength	
Resolution range	29.33 - 1.14 (1.181 - 1.14)
Space group	P 21 21 21
Unit cell	38.376 45.483 57.04 90 90 90
Total reflections	455730 (43181)
Unique reflections	36899 (644)
Multiplicity	12.4 (12.1)
Completeness (%)	88.07 (17.81)
Mean I/sigma(I)	13.70 (1.56)
Wilson B-factor	13.97
R-merge	0.07746 (1.516)
R-meas	0.08083 (1.583)
R-pim	0.02273 (0.4484)
CC1/2	0.999 (0.554)
CC*	1 (0.845)
Reflections used in refinement	32677 (644)
Reflections used for R-free	1630 (33)
R-work	0.1555 (0.2058)
R-free	0.1810 (0.2588)
CC(work)	0.969 (0.822)
CC(free)	0.953 (0.934)
Number of non-hydrogen atoms	774
macromolecules	644
ligands	0
solvent	130
Protein residues	81
RMS(bonds)	0.010
RMS(angles)	1.10
Ramachandran favored (%)	98.73
Ramachandran allowed (%)	1.27
Ramachandran outliers (%)	0.00
Rotamer outliers (%)	0.00

Clashscore	3.96
Average B-factor	21.39
macromolecules	18.90
solvent	33.67

13 References

[1] F. H. Crick, *Symp Soc Exp Biol* **1958**, *12*, 138-163.

[2] M. Samata, A. Akhtar, *Annu Rev Biochem* **2018**, *87*, 323-350.

[3] A. Graindorge, C. Militti, F. Gebauer, *Wiley Interdiscip Rev RNA* **2011**, *2*, 534-545.

[4] R. F. Hoekstra, *Nature* **2005**, *434*, 571-573.

[5] B. Charlesworth, *Curr Biol* **1996**, *6*, 149-162.

[6] J. J. Bull, *The Quarterly Review of Biology* **1980**, *55*, 3-21.

[7] T. Ezaz, R. Stiglec, F. Veyrunes, J. A. Marshall Graves, *Curr Biol* **2006**, *16*, R736-743.

[8] J. A. Graves, *Nat Rev Genet* **2016**, *17*, 33-46.

[9] J. C. Lucchesi, M. I. Kuroda, *Cold Spring Harb Perspect Biol* **2015**, *7*.

[10] T. H. Morgan, in *Nobel Lectures, Physiology or Medicine 1922-1941*, Elsevier Publishing Company, Amsterdam, **1965**.

[11] E. C. R. Reeve, *Encyclopedia of Genetics*, CRC Press, United Kingdom, **2014**.

[12] C. W. Metz, *Journal of Experimental Zoology* **1914**, 14.

[13] M. D. Adams, S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y. H. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson, G. L. Gabor, J. F. Abril, A. Agbayani, H. J. An, C. Andrews-Pfannkoch, D. Baldwin, R. M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E. M. Beasley, K. Y. Beeson, P. V. Benos, B. P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. C. Burtis, D. A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. M. Cherry, S. Cawley, C. Dahlke, L. B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A. D. Mays, I. Dew, S. M. Dietz, K. Dodson, L. E. Doup, M. Downes, S. Dugan-Rocha, B. C. Dunkov, P. Dunn, K. J. Durbin, C. C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A. E. Gabrielian, N. S. Garg, W. M. Gelbart, K. Glasser, A. Glodek, F. Gong, J. H. Gorrell, Z. Gu, P. Guan, M. Harris, N. L. Harris, D. Harvey, T. J. Heiman, J. R. Hernandez, J. Houck, D. Hostin, K. A. Houston, T. J. Howland, M. H. Wei, C. Ibegwam, et al., *Science* **2000**, *287*, 2185-2195.

[14] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H. H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, J. C. Venter, *Science* **2000**, *287*, 2196-2204.

[15] F. Cunningham, J. E. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, O. Austine-Orimoloye, A. G. Azov, I. Barnes, R. Bennett, A. Berry, J. Bhai, A. Bignell, K. Billis, S. Boddu, L. Brooks, M. Charkhchi, C. Cummins, L. Da Rin Fioretto, C. Davidson, K. Dodiya, S. Donaldson, B. El Houdaigui, T. El Naboulsi, R. Fatima, C. G. Giron, T. Genez, J. G. Martinez, C. Guijarro-Clarke, A. Gymer, M. Hardy, Z. Hollis, T. Hourlier, T. Hunt, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos,

J. C. Marugan, S. Mohanan, A. Mushtaq, M. Naven, D. N. Ogeh, A. Parker, A. Parton, M. Perry, I. Pilizota, I. Prosovetskaia, M. P. Sakthivel, A. I. A. Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, J. G. Perez-Silva, W. Stark, E. Steed, K. Sutinen, R. Sukumaran, D. Sumathipala, M. M. Suner, M. Szpak, A. Thormann, F. F. Tricomi, D. Urbina-Gomez, A. Veidenberg, T. A. Walsh, B. Walts, N. Willhoft, A. Winterbottom, E. Wass, M. Chakiachvili, B. Flint, A. Frankish, S. Giorgetti, L. Haggerty, S. E. Hunt, I. I. GR, J. E. Loveland, F. J. Martin, B. Moore, J. M. Mudge, M. Muffato, E. Perry, M. Ruffier, J. Tate, D. Thybert, S. J. Trevanion, S. Dyer, P. W. Harrison, K. L. Howe, A. D. Yates, D. R. Zerbino, P. Flicek, *Nucleic Acids Res* **2022**, *50*, D988-D995.

[16] B. S. Baker, *Nature* **1989**, *340*, 521-524.

[17] R. Piergentili, *ScientificWorldJournal* **2010**, *10*, 1749-1767.

[18] L. O. F. Penalva, L. Sanchez, *Microbiology and Molecular Biology Reviews* **2003**, *67*, 343-359.

[19] M. Bernstein, T. W. Cline, *Genetics* **1994**, *136*, 1051-1061.

[20] L. R. Bell, J. I. Horabin, P. Schedl, T. W. Cline, *Cell* **1991**, *65*, 229-239.

[21] H. K. Salz, J. W. Erickson, *Fly (Austin)* **2010**, *4*, 60-70.

[22] D. Bopp, L. R. Bell, T. W. Cline, P. Schedl, *Genes Dev* **1991**, *5*, 403-415.

[23] M. E. Samuels, P. Schedl, T. W. Cline, *Mol Cell Biol* **1991**, *11*, 3584-3602.

[24] R. Moschall, M. Gaik, J. Medenbach, *FEBS Lett* **2017**, *591*, 1471-1488.

[25] N. Handa, O. Nureki, K. Kurimoto, I. Kim, H. Sakamoto, Y. Shimura, Y. Muto, S. Yokoyama, *Nature* **1999**, *398*, 579-585.

[26] A. J. Lopez, *Annu Rev Genet* **1998**, *32*, 279-305.

[27] R. L. Kelley, I. Solovyeva, L. M. Lyman, R. Richman, V. Solovyev, M. I. Kuroda, *Cell* **1995**, *81*, 867-877.

[28] M. J. Scott, L. L. Pan, S. B. Cleland, A. L. Knox, J. Heinrich, *EMBO J* **2000**, *19*, 144-155.

[29] S. Maenner, M. Muller, J. Frohlich, D. Langer, P. B. Becker, *Mol Cell* **2013**, *51*, 174-184.

[30] P. K. A. Jagtap, M. Müller, A. E. Kiss, A. W. Thomae, K. Lapouge, M. Beck, P. B. Becker, J. Hennig, *BioRxiv* **2022**.

[31] T. Straub, C. Grimaud, G. D. Gilfillan, A. Mitterweger, P. B. Becker, *PLoS Genet* **2008**, *4*, e1000302.

[32] R. Villa, T. Schauer, P. Smialowski, T. Straub, P. B. Becker, *Nature* **2016**, *537*, 244-248.

[33] Y. Y. Shevelyov, S. V. Ulianov, M. S. Gelfand, S. N. Belyakin, S. V. Razin, *Int J Mol Sci* **2022**, *23*.

[34] T. Conrad, A. Akhtar, *Nat Rev Genet* **2012**, *13*, 123-134.

[35] A. Akhtar, P. B. Becker, *Mol Cell* **2000**, *5*, 367-375.

[36] D. Dunlap, R. Yokoyama, H. Ling, H. Y. Sun, K. McGill, S. Cugusi, J. C. Lucchesi, *Nucleic Acids Res* **2012**, *40*, 11281-11291.

[37] E. R. Smith, C. Cayrou, R. Huang, W. S. Lane, J. Cote, J. C. Lucchesi, *Mol Cell Biol* **2005**, *25*, 9175-9188.

[38] M. Taipale, S. Rea, K. Richter, A. Vilar, P. Lichter, A. Imhof, A. Akhtar, *Mol Cell Biol* **2005**, *25*, 6798-6810.

[39] S. Chlamydas, H. Holz, M. Samata, T. Chelmicki, P. Georgiev, V. Pelechano, F. Dundar, P. Dasmeh, G. Mittler, F. T. Cadete, F. Ramirez, T. Conrad, W. Wei, S. Raja, T. Manke, N. M. Luscombe, L. M. Steinmetz, A. Akhtar, *Nat Struct Mol Biol* **2016**, *23*, 580-589.

[40] A. Graindorge, C. Carre, F. Gebauer, *Genes & Development* **2013**, *27*, 1421-1433.

[41] G. J. Bashaw, B. S. Baker, *Development* **1995**, *121*, 3245-3258.

[42] S. Zhou, Y. Yang, M. J. Scott, A. Pannuti, K. C. Fehr, A. Eisen, E. V. Koonin, D. L. Fouts, R. Wrightsman, J. E. Manning, et al., *EMBO J* **1995**, *14*, 2884-2895.

[43] V. H. Meller, K. H. Wu, G. Roman, M. I. Kuroda, R. L. Davis, *Cell* **1997**, *88*, 445-457.

[44] A. Franke, B. S. Baker, *Mol Cell* **1999**, *4*, 117-122.

[45] V. H. Meller, *Mech Dev* **2003**, *120*, 759-767.

[46] M. J. Palmer, R. Richman, L. Richter, M. I. Kuroda, *Genes Dev* **1994**, *8*, 698-706.

[47] G. J. Bashaw, B. S. Baker, *Cell* **1997**, *89*, 789-798.

[48] F. Gebauer, L. Merendino, M. W. Hentze, J. Valcarcel, *RNA* **1998**, *4*, 142-150.

[49] L. Merendino, S. Guth, D. Bilbao, C. Martinez, J. Valcarcel, *Nature* **1999**, *402*, 838-841.

[50] P. Forch, L. Merendino, C. Martinez, J. Valcarcel, *RNA* **2001**, *7*, 1185-1191.

[51] K. Beckmann, M. Grskovic, F. Gebauer, M. W. Hentze, *Cell* **2005**, *122*, 529-540.

[52] R. L. Kelley, J. Wang, L. Bell, M. I. Kuroda, *Nature* **1997**, *387*, 195-199.

[53] F. Gebauer, M. W. Hentze, *Nat Rev Mol Cell Biol* **2004**, *5*, 827-835.

[54] R. J. Jackson, C. U. Hellen, T. V. Pestova, *Nat Rev Mol Cell Biol* **2010**, *11*, 113-127.

[55] C. E. Aitken, J. R. Lorsch, *Nat Struct Mol Biol* **2012**, *19*, 568-576.

[56] S. Z. Tarun, Jr., S. E. Wells, J. A. Deardorff, A. B. Sachs, *Proc Natl Acad Sci U S A* **1997**, *94*, 9046-9051.

[57] S. E. Wells, P. E. Hillner, R. D. Vale, A. B. Sachs, *Mol Cell* **1998**, *2*, 135-140.

[58] I. N. Shatsky, I. M. Terenin, V. V. Smirnova, D. E. Andreev, *Trends Biochem Sci* **2018**, *43*, 882-895.

[59] Y. Yang, Z. Wang, *J Mol Cell Biol* **2019**, *11*, 911-919.

[60] V. Truniger, M. Miras, M. A. Aranda, *Front Plant Sci* **2017**, *8*, 2047.

[61] M. E. Samuels, D. Bopp, R. A. Colvin, R. F. Roscigno, M. A. Garcia-Blanco, P. Schedl, *Mol Cell Biol* **1994**, *14*, 4975-4990.

[62] F. Gebauer, D. F. Corona, T. Preiss, P. B. Becker, M. W. Hentze, *EMBO J* **1999**, *18*, 6146-6154.

[63] M. Grskovic, M. W. Hentze, F. Gebauer, *Embo Journal* **2003**, *22*, 5571-5581.

[64] F. Gebauer, M. Grskovic, M. W. Hentze, *Molecular Cell* **2003**, *11*, 1397-1404.

[65] F. Gebauer, M. Grskovic, M. W. Hentze, *Mol Cell* **2003**, *11*, 1397-1404.

[66] K. Duncan, M. Grskovic, C. Strein, K. Beckmann, R. Niggeweg, I. Abaza, F. Gebauer, M. Wilm, M. W. Hentze, *Genes Dev* **2006**, *20*, 368-379.

[67] I. Abaza, O. Coll, S. Patalano, F. Gebauer, *Genes Dev* **2006**, *20*, 380-389.

[68] J. Hennig, C. Militti, G. M. Popowicz, I. Wang, M. Sonntag, A. Geerlof, F. Gabel, F. Gebauer, M. Sattler, *Nature* **2014**, *515*, 287-290.

[69] N. M. Hollmann, P. K. A. Jagtap, J. B. Linse, P. Ullmann, M. Payr, B. Murciano, B. Simon, J. S. Hub, J. Hennig, *Nucleic Acids Res* **2023**.

[70] K. E. Duncan, C. Strein, M. W. Hentze, *Mol Cell* **2009**, *36*, 571-582.

[71] N. Amrani, S. Ghosh, D. A. Mangus, A. Jacobson, *Nature* **2008**, *453*, 1276-1280.

[72] E. Szostak, M. Garcia-Beyaert, T. Guitart, A. Graindorge, O. Coll, F. Gebauer, *Nucleic Acids Res* **2018**, *46*, 4099-4113.

[73] A. G. Hinnebusch, *Trends Biochem Sci* **2006**, *31*, 553-562.

[74] L. S. Valasek, J. Zeman, S. Wagner, P. Beznoskova, Z. Pavlikova, M. P. Mohammad, V. Hronova, A. Herrmannova, Y. Hashem, S. Gunisova, *Nucleic Acids Res* **2017**, *45*, 10948-10968.

[75] D. Leopoldus, University of Heidelberg **2022**.

[76] J. Medenbach, M. Seiler, M. W. Hentze, *Cell* **2011**, *145*, 902-913.

[77] D. Dominguez, P. Freese, M. S. Alexis, A. Su, M. Hochman, T. Palden, C. Bazile, N. J. Lambert, E. L. Van Nostrand, G. A. Pratt, G. W. Yeo, B. R. Graveley, C. B. Burge, *Mol Cell* **2018**, *70*, 854-867 e859.

[78] G. Dreyfuss, V. N. Kim, N. Kataoka, *Nat Rev Mol Cell Biol* **2002**, *3*, 195-205.

[79] C. G. Burd, G. Dreyfuss, *Science* **1994**, *265*, 615-621.

[80] C. Maris, C. Dominguez, F. H. Allain, *FEBS J* **2005**, *272*, 2118-2131.

[81] A. Clery, M. Blatter, F. H. Allain, *Curr Opin Struct Biol* **2008**, *18*, 290-298.

[82] R. Valverde, L. Edwards, L. Regan, *FEBS J* **2008**, *275*, 2712-2726.

[83] G. Masliah, P. Barraud, F. H. Allain, *Cell Mol Life Sci* **2013**, *70*, 1875-1895.

[84] M. W. Hentze, A. Castello, T. Schwarzl, T. Preiss, *Nat Rev Mol Cell Biol* **2018**, *19*, 327-341.

[85] B. M. Lunde, C. Moore, G. Varani, *Nat Rev Mol Cell Biol* **2007**, *8*, 479-490.

[86] Y. Shamoo, N. Abdul-Manan, K. R. Williams, *Nucleic Acids Res* **1995**, *23*, 725-728.

[87] G. Dreyfuss, M. S. Swanson, S. Piñol-Roma, *Trends in Biochemical Sciences* **1988**, *13*, 86-91.

[88] S. A. Adam, T. Nakagawa, M. S. Swanson, T. K. Woodruff, G. Dreyfuss, *Mol Cell Biol* **1986**, *6*, 2932-2943.

[89] A. B. Sachs, M. W. Bond, R. D. Kornberg, *Cell* **1986**, *45*, 827-835.

[90] A. Kumar, K. R. Williams, W. Szer, *Journal of Biological Chemistry* **1986**, *261*, 1266-1273.

[91] S. R. Price, P. R. Evans, K. Nagai, *Nature* **1998**, *394*, 645-650.

[92] R. C. Deo, J. B. Bonanno, N. Sonenberg, S. K. Burley, *Cell* **1999**, *98*, 835-845.

[93] H. Kooshapur, N. R. Choudhury, B. Simon, M. Muhlbauer, A. Jussupow, N. Fernandez, A. N. Jones, A. Dallmann, F. Gabel, C. Camilloni, G. Michlewski, J. F. Caceres, M. Sattler, *Nat Commun* **2018**, *9*, 2479.

[94] T. Afroz, Z. Cienikova, A. Clery, F. H. T. Allain, *Methods Enzymol* **2015**, *558*, 235-278.

[95] C. Dominguez, F. H. Allain, *Nucleic Acids Res* **2006**, *34*, 3634-3645.

[96] S. D. Auweter, R. Fasan, L. Reymond, J. G. Underwood, D. L. Black, S. Pitsch, F. H. Allain, *EMBO J* **2006**, *25*, 163-173.

[97] J. M. Perez-Canadillas, *EMBO J* **2006**, *25*, 3167-3178.

[98] M. R. Conte, T. Grune, J. Ghuman, G. Kelly, A. Ladas, S. Matthews, S. Curry, *EMBO J* **2000**, *19*, 3132-3141.

[99] F. C. Oberstrass, S. D. Auweter, M. Erat, Y. Hargous, A. Henning, P. Wenter, L. Reymond, B. Amir-Ahmady, S. Pitsch, D. L. Black, F. H. Allain, *Science* **2005**, *309*, 2054-2057.

[100] I. Wang, J. Hennig, P. K. Jagtap, M. Sonntag, J. Valcarcel, M. Sattler, *Nucleic Acids Res* **2014**, *42*, 5949-5966.

[101] J. Ding, M. K. Hayashi, Y. Zhang, L. Manche, A. R. Krainer, R. M. Xu, *Genes Dev* **1999**, *13*, 1102-1115.

[102] R. M. Xu, L. Jokhan, X. Cheng, A. Mayeda, A. R. Krainer, *Structure* **1997**, *5*, 559-570.

[103] S. Fribourg, D. Gatfield, E. Izaurralde, E. Conti, *Nat Struct Biol* **2003**, *10*, 433-439.

[104] L. ElAntak, A. G. Tzakos, N. Locker, P. J. Lukavsky, *J Biol Chem* **2007**, *282*, 8165-8174.

[105] N. Ripin, J. Boudet, M. M. Duszczyk, A. Hinniger, M. Faller, M. Krepl, A. Gadi, R. J. Schneider, J. Sponer, N. C. Meisner-Kober, F. H. Allain, *Proc Natl Acad Sci U S A* **2019**, *116*, 2935-2944.

[106] C. Schelhorn, J. M. Gordon, L. Ruiz, J. Alguacil, E. Pedroso, M. J. Macias, *Nucleic Acids Res* **2014**, *42*, 10185-10195.

[107] R. Kanaar, A. L. Lee, D. Z. Rudner, D. E. Wemmer, D. C. Rio, *EMBO J* **1995**, *14*, 4530-4539.

[108] M. Samuels, G. Deshpande, P. Schedl, *Nucleic Acids Res* **1998**, *26*, 2625-2637.

[109] E. Sakashita, H. Sakamoto, *J Biochem* **1996**, *120*, 1028-1033.

[110] S. Patalano, M. Mihailovich, Y. Belacortu, N. Paricio, F. Gebauer, *Development* **2009**, *136*, 689-698.

[111] C. Militti, S. Maenner, P. B. Becker, F. Gebauer, *Nat Commun* **2014**, *5*, 4762.

[112] I. Abaza, F. Gebauer, *RNA* **2008**, *14*, 482-490.

[113] H. Jacquemin-Sablon, G. Triqueneaux, S. Deschamps, M. le Maire, J. Doniger, F. Dautry, *Nucleic Acids Res* **1994**, *22*, 2643-2650.

[114] N. M. Hollmann, P. K. A. Jagtap, P. Masiewicz, T. Guitart, B. Simon, J. Provaznik, F. Stein, P. Haberkant, L. J. Sweetapple, L. Villacorta, D. Mooijman, V. Benes, M. M. Savitski, F. Gebauer, J. Hennig, *Cell Rep* **2020**, *32*, 107930.

[115] M. J. Matunis, E. L. Matunis, G. Dreyfuss, *J Cell Biol* **1992**, *116*, 245-255.

[116] E. L. Matunis, M. J. Matunis, G. Dreyfuss, *J Cell Biol* **1992**, *116*, 257-269.

[117] T. Yano, S. Lopez de Quinto, Y. Matsui, A. Shevchenko, A. Shevchenko, A. Ephrussi, *Dev Cell* **2004**, *6*, 637-648.

[118] L. E. Hammond, D. Z. Rudner, R. Kanaar, D. C. Rio, *Mol Cell Biol* **1997**, *17*, 7260-7267.

[119] D. S. Finger, A. E. Williams, V. V. Holt, E. T. Ables, *Dev Dyn* **2022**.

[120] C. W. Siebel, A. Admon, D. C. Rio, *Genes Dev* **1995**, *9*, 269-283.

[121] M. Bose, M. Lampe, J. Mahamid, A. Ephrussi, *Cell* **2022**, *185*, 1308-1324 e1323.

[122] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583-589.

[123] C. W. Siebel, L. D. Fresco, D. C. Rio, *Genes Dev* **1992**, *6*, 1386-1401.

[124] S. Majumdar, D. C. Rio, *Microbiol Spectr* **2015**, *3*, MDNA3-0004-2014.

[125] C. W. Siebel, R. Kanaar, D. C. Rio, *Genes Dev* **1994**, *8*, 1713-1725.

[126] M. Blanchette, R. E. Green, S. E. Brenner, D. C. Rio, *Genes Dev* **2005**, *19*, 1306-1314.

[127] F. E. Baralle, J. Giudice, *Nat Rev Mol Cell Biol* **2017**, *18*, 437-451.

[128] A. J. Lopez, R. D. Artero, M. Perez-Alonso, *Rouxs Arch Dev Biol* **1996**, *205*, 450-459.

[129] J. M. Burnette, A. R. Hatton, A. J. Lopez, *Genetics* **1999**, *151*, 1517-1529.

[130] R. P. Jansen, *Nat Rev Mol Cell Biol* **2001**, *2*, 247-256.

[131] V. Riechmann, A. Ephrussi, *Curr Opin Genet Dev* **2001**, *11*, 374-383.

[132] N. Gunkel, T. Yano, F. H. Markussen, L. C. Olsen, A. Ephrussi, *Genes Dev* **1998**, *12*, 1652-1664.

[133] J.-R. Huynh, T. P. Munro, K. Smith-Litière, J.-A. Lepesant, D. S. Johnston, *Developmental Cell* **2004**, *6*, 625-635.

[134] K. N. Clouse, S. B. Ferguson, T. Schupbach, *Dev Biol* **2008**, *313*, 713-724.

[135] J. S. Goodrich, K. N. Clouse, T. Schupbach, *Development* **2004**, *131*, 1949-1958.

[136] A. Norvell, R. L. Kelley, K. Wehr, T. Schupbach, *Genes Dev* **1999**, *13*, 864-876.

[137] Y. Kalifa, S. T. Armenti, E. R. Gavis, *Dev Biol* **2009**, *326*, 68-74.

[138] C. Geng, P. M. Macdonald, *Mol Cell Biol* **2006**, *26*, 9508-9516.

[139] M. R. Nelson, H. Luo, H. K. Vari, B. J. Cox, A. J. Simmonds, H. M. Krause, H. D. Lipshitz, C. A. Smibert, *J Biol Chem* **2007**, *282*, 34031-34038.

[140] Y. Suissa, Y. Kalifa, T. Dinur, P. Graham, G. Deshpande, P. Schedl, O. Gerlitz, *Proceedings of the National Academy of Sciences* **2010**, *107*, 6930-6935.

[141] J. K. Penn, P. Schedl, *Dev Cell* **2007**, *12*, 275-286.

[142] D. Dutta, M. S. Paul, A. Singh, M. Mutsuddi, A. Mukherjee, *Genetics* **2017**, *206*, 905-918.

[143] D. Dutta, M. Mutsuddi, A. Mukherjee, *Cell Biol Int* **2019**, *43*, 350-357.

[144] E. T. Ables, G. H. Hwang, D. S. Finger, T. D. Hinnant, D. Drummond-Barbosa, *G3 (Bethesda)* **2016**, *6*, 2629-2642.

[145] J. K. Bhogal, J. M. Kanaskie, J. R. DiAngelo, *Biochem Biophys Res Commun* **2020**, *524*, 178-183.

[146] J. M. Mainak Bose, Anne Ephrussi, *bioRxiv* **2021**, 2021.2003.2031.437848.

[147] T. P. Munro, R. J. Magee, G. J. Kidd, J. H. Carson, E. Barbarese, L. M. Smith, R. Smith, *J Biol Chem* **1999**, *274*, 34389-34395.

[148] J. Shan, T. P. Munro, E. Barbarese, J. H. Carson, R. Smith, *J Neurosci* **2003**, *23*, 8859-8866.

[149] H. Bruckert, G. Marchetti, M. Ramialison, F. Besse, *PLoS One* **2015**, *10*, e0136610.

[150] D. Hodgkin, **1937**.

[151] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, D. C. Phillips, *Nature* **1958**, *181*, 662-666.

[152] S. K. Burley, C. Bhikadiya, C. Bi, S. Bittrich, L. Chen, G. V. Crichlow, C. H. Christie, K. Dalenberg, L. Di Costanzo, J. M. Duarte, S. Dutta, Z. Feng, S. Ganesan, D. S. Goodsell, S. Ghosh, R. K. Green, V. Guranovic, D. Guzenko, B. P. Hudson, C. L. Lawson, Y. Liang, R. Lowe, H. Namkoong, E. Peisach, I. Persikova, C. Randle, A. Rose, Y. Rose, A. Sali, J. Segura, M. Sekharan, C. Shao, Y. P. Tao, M. Voigt, J. D. Westbrook, J. Y. Young, C. Zardecki, M. Zhuravleva, *Nucleic Acids Res* **2021**, *49*, D437-D451.

[153] R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, *bioRxiv* **2022**, 2021.2010.2004.463034.

[154] M. P. Rout, A. Sali, *Cell* **2019**, *177*, 1384-1403.

[155] I. R. Kleckner, M. P. Foster, *Biochim Biophys Acta* **2011**, *1814*, 942-968.

[156] M. Kovermann, P. Rogne, M. Wolf-Watz, *Q Rev Biophys* **2016**, *49*, e6.

[157] K. Henzler-Wildman, D. Kern, *Nature* **2007**, *450*, 964-972.

[158] P. J. Hore, *Nuclear magnetic resonance*, Oxford University Press, **2015**.

[159] J. Cavanagh, in *Protein NMR Spectroscopy (Second Edition)* (Eds.: J. Cavanagh, W. J. Fairbrother, A. G. Palmer, M. Rance, N. J. Skelton), Academic Press, Burlington, **2007**.

[160] J. Keeler, *Understanding NMR Spectroscopy*, 2nd ed., Wiley, **2010**.

[161] O. Zerbe, S. Jurt, *Applied NMR Spectroscopy for Chemists and Life Scientists*, Wiley, **2014**.

[162] I. Solomon, *Physical Review* **1955**, *99*, 559-565.

[163] D. S. Wishart, C. G. Bigam, J. Yao, F. Abildgaard, H. J. Dyson, E. Oldfield, J. L. Markley, B. D. Sykes, *J Biomol NMR* **1995**, *6*, 135-140.

[164] H. Kessler, M. Gehrke, C. Griesinger, *Angewandte Chemie International Edition in English* **1988**, *27*, 490-536.

[165] G. A. Morris, R. Freeman, *Journal of the American Chemical Society* **1979**, *101*, 760-762.

[166] M. Sattler, J. Schleucher, C. Griesinger, *Progress in Nuclear Magnetic Resonance Spectroscopy* **1999**, *34*, 93-158.

[167] L. E. Kay, M. Ikura, R. Tschudin, A. Bax, *J Magn Reson* **1990**, *89*, 496-514.

[168] M. P. Williamson, *Prog Nucl Magn Reson Spectrosc* **2013**, *73*, 1-16.

[169] A. W. Overhauser, *Physical Review* **1953**, *92*, 411-415.

[170] Y. Shen, F. Delaglio, G. Cornilescu, A. Bax, *J Biomol NMR* **2009**, *44*, 213-223.

[171] M. Karplus, *Journal of the American Chemical Society* **1963**, *85*, 2870-2871.

[172] J. L. Battiste, G. Wagner, *Biochemistry* **2000**, *39*, 5355-5365.

[173] N. Tjandra, A. Szabo, A. Bax, *Journal of the American Chemical Society* **1996**, *118*, 6986-6991.

[174] G. Lipari, A. Szabo, *Journal of the American Chemical Society* **1982**, *104*, 4546-4559.

[175] L. E. Kay, D. A. Torchia, A. Bax, *Biochemistry* **1989**, *28*, 8972-8979.

[176] F. van den Ent, J. Lowe, *J Biochem Biophys Methods* **2006**, *67*, 67-74.

[177] B. Simon, H. Kostler, *J Biomol NMR* **2019**, *73*, 155-165.

[178] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, A. Bax, *J Biomol NMR* **1995**, *6*, 277-293.

[179] W. F. Vranken, W. Boucher, T. J. Stevens, R. H. Fogh, A. Pajon, M. Llinas, E. L. Ulrich, J. L. Markley, J. Ionides, E. D. Laue, *Proteins* **2005**, *59*, 687-696.

[180] S. P. Skinner, R. H. Fogh, W. Boucher, T. J. Ragan, L. G. Mureddu, G. W. Vuister, *J Biomol NMR* **2016**, *66*, 111-124.

[181] W. Lee, M. Tonelli, J. L. Markley, *Bioinformatics* **2015**, *31*, 1325-1327.

[182] V. r. p. The PyMOL Molecular Graphics System, Schrödinger, LLC.

[183] M. Niklasson, R. Otten, A. Ahlner, C. Andresen, J. Schlagnitweit, K. Petzold, P. Lundstrom, *J Biomol NMR* **2017**, *69*, 93-99.

[184] A. Ahlner, M. Carlsson, B. H. Jonsson, P. Lundstrom, *J Biomol NMR* **2013**, *56*, 191-202.

[185] D. Fushman, R. Weisemann, H. Thuring, H. Ruterjans, *J Biomol NMR* **1994**, *4*, 61-78.

[186] F. Cordier, A. J. Dingley, S. Grzesiek, *J Biomol NMR* **1999**, *13*, 175-180.

[187] P. Dosset, J. C. Hus, D. Marion, M. Blackledge, *J Biomol NMR* **2001**, *20*, 223-231.

[188] B. Simon, T. Madl, C. D. Mackereth, M. Nilges, M. Sattler, *Angew Chem Int Ed Engl* **2010**, *49*, 1967-1970.

[189] A. T. Brunger, *Nat Protoc* **2007**, *2*, 2728-2733.

[190] A. T. Brunger, P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J. S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson, G. L. Warren, *Acta Crystallogr D Biol Crystallogr* **1998**, *54*, 905-921.

[191] P. D. Adams, P. V. Afonine, G. Bunkoczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, P. H. Zwart, *Acta Crystallogr D Biol Crystallogr* **2010**, *66*, 213-221.

[192] T. C. Terwilliger, R. W. Grosse-Kunstleve, P. V. Afonine, N. W. Moriarty, P. H. Zwart, L. W. Hung, R. J. Read, P. D. Adams, *Acta Crystallogr D Biol Crystallogr* **2008**, *64*, 61-69.

[193] P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, *Acta Crystallogr D Biol Crystallogr* **2010**, *66*, 486-501.

[194] P. V. Afonine, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty, M. Mustyakimov, T. C. Terwilliger, A. Urzhumtsev, P. H. Zwart, P. D. Adams, *Acta Crystallogr D Biol Crystallogr* **2012**, *68*, 352-367.

[195] D. A. Case, H. M. Aktulga, K. Belfon, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, I. T.E. Cheatham, G. A. Cisneros, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, G. Giambasu, M. K. Gilson, H. Gohlke, A. W. Goetz, R. Harris, S. Izadi, S. A. Izmailov, C. Jin, K. Kasavajhala, M. C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T. S. Lee, S. LeGrand, P. A. Kollman, University of California, San Francisco, **2021**.

[196] M. Zgarbova, M. Otyepka, J. Sponer, A. Mladek, P. Banas, T. E. Cheatham, 3rd, P. Jurecka, *J Chem Theory Comput* **2011**, *7*, 2886-2902.

[197] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, *J Chem Theory Comput* **2015**, *11*, 3696-3713.

[198] M. Krepl, P. Pokorna, V. Mlynsky, P. Stadlbauer, J. Sponer, *Nucleic Acids Res* **2022**, *50*, 12480-12496.

[199] H. J. C. Berendsen, J. R. Grigera, T. P. Straatsma, *The Journal of Physical Chemistry* **1987**, *91*, 6269-6271.

[200] I. S. Joung, T. E. Cheatham, 3rd, *J Phys Chem B* **2008**, *112*, 9020-9041.

[201] M. Krepl, J. Vogele, H. Kruse, E. Duchardt-Ferner, J. Wohnert, J. Sponer, *Nucleic Acids Res* **2018**, *46*, 6528-6543.

[202] M. M. Garner, A. Revzin, *Nucleic Acids Res* **1981**, *9*, 3047-3060.

[203] Y. Shamoo, U. Krueger, L. M. Rice, K. R. Williams, T. A. Steitz, *Nat Struct Biol* **1997**, *4*, 215-222.

[204] M. Krepl, P. Pokorna, V. Mlynsky, P. Stadlbauer, J. Sponer, *bioRxiv* **2022**, *2022.07.22.501120*.

[205] S. K. Upadhyay, C. D. Mackereth, *Nucleic Acids Res* **2020**, *48*, 4538-4550.

[206] M. Sattler, S. W. Fesik, *Structure* **1996**, *4*, 1245-1249.

[207] Y. Shamoo, N. Abdul-Manan, A. M. Patten, J. K. Crawford, M. C. Pellegrini, K. R. Williams, *Biochemistry* **1994**, *33*, 8272-8281.

[208] M. Pabis, G. M. Popowicz, R. Stehle, D. Fernandez-Ramos, S. Asami, L. Warner, S. M. Garcia-Maurino, A. Schlundt, M. L. Martinez-Chantar, I. Diaz-Moreno, M. Sattler, *Nucleic Acids Res* **2019**, *47*, 1011-1029.

[209] X. P. Wang, N. G. Cooper, *Bioinform Biol Insights* **2010**, *4*, 61-83.

[210] C. L. Lin, Y. T. Huang, J. D. Richter, *RNA* **2012**, *18*, 1050-1061.

[211] J. D. Chodera, D. L. Mobley, *Annu Rev Biophys* **2013**, *42*, 121-142.

[212] V. Lafont, A. A. Armstrong, H. Ohtaka, Y. Kiso, L. Mario Amzel, E. Freire, *Chem Biol Drug Des* **2007**, *69*, 413-422.

[213] N. M. King, M. Prabu-Jeyabalan, R. M. Bandaranayake, M. N. Nalam, E. A. Nalivaika, A. Ozen, T. Haliloglu, N. K. Yilmaz, C. A. Schiffer, *ACS Chem Biol* **2012**, *7*, 1536-1546.

[214] A. I. Dragan, C. M. Read, C. Crane-Robinson, *Eur Biophys J* **2017**, *46*, 301-308.

[215] M. L. Verteramo, O. Stenstrom, M. M. Ignjatovic, O. Caldararu, M. A. Olsson, F. Manzoni, H. Leffler, E. Oksanen, D. T. Logan, U. J. Nilsson, U. Ryde, M. Akke, *J Am Chem Soc* **2019**, *141*, 2012-2026.

[216] J. E. Shattuck, A. C. Waechter, E. D. Ross, *Prion* **2017**, *11*, 249-264.

[217] M. Sharon, C. V. Robinson, *Annu Rev Biochem* **2007**, *76*, 167-193.

[218] F. Madeira, M. Pearce, A. R. N. Tivey, P. Basutkar, J. Lee, O. Edbali, N. Madhusoodanan, A. Kolesnikov, R. Lopez, *Nucleic Acids Res* **2022**, *50*, W276-279.

[219] X. Robert, P. Gouet, *Nucleic Acids Res* **2014**, *42*, W320-324.