# NMR investigations of 15.5K associated protein-RNA complexes

Dissertation

zur Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultäten

der Georg-August-Universität zu Göttingen

Vorgelegt von

Ping Li

aus Beijing, China

Göttingen 2007

D7

Referent: Prof. Dr. Ulf Diederichsen

Korreferent: Prof. Dr. Christian Griesinger

Tag der mündlichen Prüfung:

# Zusammenfassung

Das Spleissosom besteht aus unterschiedlichen snRNPs (small nuclear ribonucleoproteins), die eine Spleissreaktion katalysieren. Um das Spleissosom zu aktivieren und den ersten Schritt der Spleissreaktion zu ermöglichen, dissoziieren die basengepaarten U4 und U6 snRNAs des U4/U6 snRNP Komplexes. Dieser Komplex ist mit den spleissosomalen gemeinsamen Sm- und LSm-Proteinen, sowie mit den spezifischen Proteinen 15.5K, hPrp31 und dem CypH/hPrp4/hPrp3 Protein-Trikomplex assoziiert. Um die strukturellen Umlagerungen, innerhalb des U4/U6 snRNA-Duplexes, die der Spleissreaktion vorrangehen, zu verstehen, ist das Wissen über diese Proteine und ihre Wechselwirkungen von immenser Wichtigkeit. Insbesondere deshalb da konformationelle änderungen der snRNAs von den gebundenen Proteinen stabilisiert und unterstützt werden. Das Ziel der vorliegenden Arbeit war es, die Interaktionen von zwei der spezifisch bindenden Proteine, dem 15.5K und hPrp31, des U4 snRNP Komplexes zu untersuchen.

Biochemische Untersuchungen identifizierten das humane 15.5K Protein als Bindungspartner der U4 snRNA, wobei die Bindung an der 5' Stammschleife (5'-SL) erfolgt und notwendig für die Zusammenführung weiterer U4/U6 spezifischen Proteine ist. Es zeigte sich, dass die Länge des Stamms II der U4 5'-SL und nicht die Sequenz ausschlaggebend für die Bindung des hPrp31 ist. Das 15.5K Protein wurde ferner als Kernprotein von box C/D snoRNAs im Nucleoli beschrieben. Die Bindung von 15.5K ist auch dort ver-

antwortlich für die Rekrutierung weiterer Kernproteine, wie Nop56 und Nop 58, deren Nop-Domäne homolog zum hPrp31 Protein ist. Bei der Bindung von Nop56 und Nop58 wurde festgestellt, dass die Sequenz des Stammes II essentiell für die Bindung ist. Weiterhin wurde 15.5K als Bindeprotein der U3 box C/D snoRNA erkannt. Die Assoziation an das B/C Motiv der U3 box C/D snoRNA rekrutiert das hU3-55K Protein, welches statt einer Nop- eine WD40 repeat- Domäne besitzt. In all diesen RNPs erkennt das 15.5K Protein ein charakteristisches Kink-turn (K-turn) Motiv der RNAs. Die rekrutierten sekundären Bindeproteine weisen direkte Kontakte zu den RNAs auf. Detaillierte Protein-RNA und Protein-Protein Interaktionen sind in diesen RNPs jedoch nicht bekannt. Wie diese sehr ähnlichen primären RNPs zwischen den sekundären Bindeproteinen unterscheiden, ist auch von grossem Interesse für uns.

In dieser Arbeit wurde unter Verwendung zweidimensionaler NMR-Titrationsexperimente (HSQCs) sowie Kreuz-Sättigungsexperimenten an den Komplexen hPrp31-15.5K-U4 5'-SL und hPrp$^{78-333}$-15.5K-U4 5'-SL die Bindungsoberfläche des 15.5K zum hPrp31 untersucht. Durch Kombination von Homologie-Modulierung und dem HADDOCK2.0 docking wurde mit Hilfe von NMR und biochemischen Daten ein 3D Model des ternären Komplexes entwickelt. Aus diesen Ergebnissen charakterisieren wir die Nop-Domäne als bona fide RNP Bindedomäne. Die Rolle des 15.5K Proteins ist daher nicht auf das Induzieren und Stabilisieren einer hPrp31 Bindungsstelle in der RNA beschränkt, sondern es stellt selber etwa die Hälfte der Kontaktoberfläche der Nop Domäne dar. Mit unserem Andockmodel waren wir in der Lage aufzuzeigen, dass eine Verlängerung des Stamms II der U4 snRNA aufgrund einer physikalischen Barriere, welche die Nop Domäne bereitstellt, benachteiligt wird. Da in box C/D snoRNAs der Stamm II die ideale Länge überschreitet, kann hPrp31 nicht in box C/D snoRNP Komplexen gebunden werden. Bei einer Oberflächenmutante von 15.5K, die Mutatio-

nen ausserhalb der Bindungsoberfläche zwischen 15.5K und hPrp31 besitzt, wurde zuvor gezeigt, dass die Bindung von hPrp31 stark reduziert wird. Unter Verwendung von HSQC-Titrationsexperimenten und Verfeinerungen der Struktur mit Hilfe residualer dipolarer Kopplungen konnten wir demonstrieren, dass die Struktur der Mutante nicht wesentlich vom Wildtyp abweicht und die niedrigere Bindungsaffinität daher von kleinen änderungen im Ladungsverhalten der Bindeoberfläche hervorgerufen werden könnte. Weiterhin wurden strukturelle änderungen des 15.5K Proteins im Komplex mit verschiedenen box C/D snoRNA Konstrukten mittels Titrationsexperimenten untersucht, um die Frage zu klären, ob die Struktur von 15.5K auch zur Selektivität der primären RNPs beiträgt. Die Daten der chemischen Verschiebungsabweichungen deuten daraufhin, dass die Struktur von 15.5K sich nicht wesentlich in den primären RNP Komplexen unterscheidet. Die Selektivität wird daher auf die Unterschiede der RNAs und der sekundären Bindeproteine zurückzuführen sein.

iv

# Abstract

The spliceosome, consisting of different snRNPs (small nuclear ribonucleoproteins) and numerous non-U snRNP factors, catalyzes the splicing reaction. To activate the spliceosome and enable the first step of splicing to begin, the extensively base-paired U4, U6 snRNAs in the U4/U6 snRNP dissociate from each other. The U4, U6 snRNAs are associated to the common core proteins of the spliceosome namely the Sm and LSm proteins, as well as the specific proteins including 15.5K, hPrp31, and the CypH/hPrp4/hPrp3 protein tricomplex. The knowledge on these proteins is of paramount importance for the understanding of the dramatic structural rearrangement of the U4/U6 snRNA duplex prior to splicing, as changes in the structures of these snRNAs are likely to be stabilized and assisted by the bound proteins. In this work, effort was made to understand the interactions between two of these specific proteins namely 15.5K and hPrp31 in the context of the U4 snRNP.

Biochemical studies demonstrated that 15.5K binds to the 5' stem-loop of the U4 snRNA (U4 5'-SL) and nucleates the binding of the other U4/U6 specific proteins. It was revealed that the length but not the sequence of the stem II of U4 5'-SL is crucial for the binding of hPrp31. 15.5K was also characterized in the nucleoli as one of the core proteins of the box C/D snoRNAs. The binding of 15.5K is again required prior to the association of other core proteins, including Nop56 and Nop58 proteins, which share the Nop homology domain with hPrp31. It was found in this case that for the

binding of Nop56 and Nop58, the sequence of stem II of the snoRNA cannot be altered. Moreover, 15.5K associates with the box B/C motif of the U3 box C/D snoRNA and recruits the hU3-55K protein, which contains the WD 40 repeat domain. In all of these RNPs, 15.5K recognizes the characteristic K-turn motif of the RNAs and the secondary proteins were shown to form direct contacts with the RNAs. However, detailed protein-RNA and protein-protein interactions in these RNPs are not known. How do these very similar primary RNPs discriminate among their secondary binding proteins is also of great interest to us.

In this work, using HSQC titrations and cross-saturation experiments on the hPrp31-15.5K-U4 5'-SL and hPrp31$^{78-333}$-15.5K-U4 5'-SL complexes, we defined the interaction surface on 15.5K in complex with hPrp31. Combining the NMR and biochemical data, we successfully generated a 3D model of the ternary complex using comparative modeling and the HADDOCK2.0 docking program. From these results, we characterized the Nop domain as a *bona fide* RNP binding domain. The role of the assembly-initiating 15.5K is, therefore, not restricted to inducing or stabilizing a hPrp31 binding site in the RNA; rather 15.5K itself provides approximately half of the contact surface for the Nop domain of hPrp31. In the docking model, the interaction surfaces on 15.5K and hPrp31 showed charge complementarity. From our docking model we could also demonstrate that the elongation of stem II in U4 snRNA is highly unfavoured due to the physical barrier provided by the Nop domain. As in box C/D snoRNAs the length of the stem II naturally exceeds the required length, hPrp31 is not recruited into the box C/D snoRNPs. A surface mutant of 15.5K, which contains mutations outside the interaction surface between 15.5K and hPrp31, was previously shown to strongly reduce the binding of hPrp31. Using HSQC titration and RDC refinement, we could demonstrate that the structure of this mutant does not significantly vary from the wild type and therefore, the effect of this 15.5K mutant on

hPrp31 binding could arise from subtle changes in the charge property of the interaction surface.

Furthermore, the structural changes of 15.5K in complexes with different box C/D snoRNA constructs were studied using HSQC titrations to address the question whether the structure of 15.5K also contributes to the selectivity of these primary RNPs. The chemical shift perturbation data showed that the structure of 15.5K does not differ dramatically in this primary RNPs. Therefore, the selectivity arises primarily from the differences in the RNAs and the secondary binding proteins.

# Acknowledgments

I would like take the chance to sincerely thank all the colleagues in the departments of Prof. Christian Griesinger and Prof. Reinhard Lührmann, who have helped and encouraged me enormously during these years of my PhD study. Due to their presence, I have greatly enjoyed working in these two departments and will keep this fond memory for the rest of my life.

Firstly, I would like to express my deep gratitude to my supervisor Dr. Teresa Carlomagno for her high standards, intellectual stimulation, tremendous help, patience and guidance throughout my thesis. It has been a great pleasure to work with her and to learn from her. I wish her all the best with her future career in EMBL, Heidelberg.

I am deeply grateful to Prof. Reinhard Lührmann and Prof. Christian Griesinger for accepting me in this world-renown institute and for this very interesting project as well as for making this excellent collaboration possible. I would also like to thank them both for providing the state-of-art facilities in these departments, which made this challenging work possible and for supporting me throughout my PhD.

I would also like to thank Prof. Ulf Diederichsen for accepting me as an external PhD student and Prof. Lutz Ackermann, Prof. Jörg Magull, Prof. Michael Buback for agreeing to be members of my external PhD thesis committee. Thanks also to Prof. Ralf Ficner for agreeing to be my examiner outside the chemistry department.

I am indebted to Dr. Markus Wahl for the excellent collaboration, helpful discussion and his tremendous input in our joined publication. At the same time, I would like to sincerely thank Dr. Sunbin Liu for his patience in helping me in the lab and for sharing his expertise in protein expression and purification.

Many thanks to Dr. Stephanie Nottrott, who introduced me to the biochemical aspects of this work. Her encouragement and friendship helped me greatly especially in the first one and half years of my study.

Many special thanks to Dr. Devanathan Raghunathan for his immense help on the structural calculation part of my thesis. I would never forget that he willingly provided me with help even when he was in hospital.

Many thanks to Claudia Schwiegk for her help on RNA sample preparation.

I deeply appreciate the support from Marcel Reese, who constantly helped me with questions related to structural calculations, computing at gwdg, Unix commands and 'awk' scripting.

I owe thanks to Dr. Reinhard Rauhut for his help on sequence alignment and comparative modeling.

I would also like to thank Siegrid Silberer, the secretary of Prof. Griesinger and Juliane Moses, the secretary of Prof. Lührmann for their tremendous help on administrative issues and for organizing trips to conferences.

I feel very lucky to be in the department of Prof. Griesinger, which consists of friendly and helpful members from all nations. I would like to thank Dr. Victor Sánchez Pedregal, who is now at University of Santiago de Compostela in Spain, for his help on NMR experiments at the beginning of my study. Many thanks to Dr. Christophe Farès, who is now at the Ontario Institute for Cancer Research in Toronto, for helping me greatly on IPAP experiments and providing very useful Felix macros as well as laughters in the office. Additional thanks to Nils-Alexander Lakomek and Dr. Vinesh Vijayan for helpful advices on NMR related problems. I would like to sincerely thank

# Publication

This thesis lead to the publication below:

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1  NMR spectroscopy and its role in structural biology

NMR stands for nuclear magnetic resonance. The word 'nuclear' comes about, because the phenomena is observed for nuclear spins. 'Magnetic' comes from the external magnetic field, which splits the spins into two unequally populated energy levels. The transitions between the energy levels can be induced by the irradiation with a frequency equal to the energy difference between the levels, which accounts for the word 'resonance'. The phenomena of NMR was discovered through the detection of weak radiofrequency signals generated by the nuclei of atoms in wax and water in 1946 (1; 2). Since its discovery, NMR has become a powerful physical tool for investigation of matters varying from small chemicals, biomolecules, cells to human brains in order to obtain information on molecular identity, structure, dynamics and spatial distribution (magnetic resonance imaging) (3). Modern NMR spectroscopy is based on pulse Fourier transform NMR spec-

troscopy by Ernst and Anderson and on multidimensional NMR spectroscopy by Ernst and Jeener (4; 5; 6). In the past, the application of NMR spectroscopy to structural investigation of large bio-molecules was largely limited due to its intrinsic low sensitivity and the complexity of the information contained in NMR spectra. In the past decades, great amount of effort has been made in order to increase the signal-to-noise ratio and to disentangling the complex information content in the spectra. The increase in magnetic field strength brings great improvement in both sensitivity (3/2 power of the ratio of the magnetic fields strengths being compared) and resolution. The common usage of cryo-probes results in a 3-4 fold increase in signal sensitivity. Furthermore, the rapid advancing in NMR methodology as well as various techniques for sample preparation have greatly increased the importance of NMR in structural biology. NMR spectroscopy and X-ray crystallography currently are the two techniques for determining three-dimensional structures of macromolecules at atomic resolution. Furthermore, the cryo-electron microscopy (EM) method can also provide structural information at atomic resolution when the sample is highly ordered. However, for complexes such as the spliceosome, which is dynamic in nature, the maximum resolution with EM is $\sim$ 8-10 Å(7).

In comparison with X-ray crystallography and cryo-EM techniques, NMR spectroscopy has the prevailing ability of studying intra-molecular dynamics, kinetics and interactions between molecules. It enables structural investigations on unfolded proteins and intrinsically flexible RNAs, which usually can not be crystallized, and provides the possibility to study time-resolved phenomena. In high throughput structural genomics approaches, NMR is used routinely for rapid screening of good structural targets for both crystallography and NMR studies and structure determination of proteins that contain floppy regions and multiple domains. In functional ge-

nomics, NMR is employed in rapid screening for ligand-substrate binding, protein-protein interactions as well as in characterization of dynamic regions in proteins (8). With the development of transverse relaxation optimized spectroscopy (TROSY) (9; 10), the advances in deuteration and selective labeling (11; 12; 13; 14; 15; 16) as well as new computational methods (17), the size of the macromolecule which can be studied by NMR has been increased from 10-15 kDa to well above 50 kDa. The development on NMR spectroscopy has led to several Nobel prizes (18). However, large amount of effort is still on-going to battle low sensitivity and to minimize the limitation on the size of the macromolecules which can be studied.

The majority of the structures present in the protein data bank (PDB) are from single molecules, whereas structures from biomolecular complexes are still scarce. As biomolecules largely function in complexes, it is conspicuous that the knowledge of complex structures is of crucial importance for the understanding of processes in the cell. The obstacles in complex structure solution come from the difficulties usually present in crystallization attempts and from the limitation on molecular size in NMR. In present days NMR spectroscopy in combination with X-ray crystallography as well as computational methods work together to tackle this task (19). In this thesis such an example of using combined structural biology methods to study a ternary complex in the spliceosome will be presented.

## 1.2 Pre-mRNA splicing and the spliceosome

Genes are segments of DNA that encode functional biopolymers, either RNA or protein (20). Gene expression is a multistep process involving chromatin remodeling, transcription, RNA processing, RNA export and finally translation in the cytoplasm. Eukaryotic genes show mosaic structure, in which

the coding sequences of genes, the exons, are interrupted by non-coding sequences, the introns. The transcription product of RNA polymerase II in eukaryotes is called precursor-messenger RNA (pre-mRNA), which is modified through two processing steps: 5'-end capping with 7-methylguanylate ($m^7$G-Cap) and 3'-end polyadenylation, the addition of a poly (A) tail by poly (A) polymerase. The modified pre-mRNA contains both introns, which are removed and exons, which are joined to form mature mRNA. The mature mRNA will then be transported from the nucleus into the cytoplasm, where translation occurs. In higher eukaryotes, a gene sequence contains a much larger fraction of intron sequences, hundreds of thousands of nucleotides, than of exon sequences (50-300nt for one exon). The range of intron densities in annotated eukaryotic genomes spans more than three orders of magnitude, from 140,000 introns in the human genome (8.4 introns per gene) to 13 introns in a lower eukaryotic species *Encephalitozoon cuniculi* (0.0065 introns per gene). In contrast, most prokaryotes lack introns, therefore, transcription and translation occur concurrently.

### 1.2.1   Pre-mRNA splicing

In eukaryotes, the process of removing introns and simultaneously joining the adjacent exons is called splicing. The process is carried out by a complex and dynamic molecular machine called the spliceosome. Introns need to be removed with single nucleotide precision, to avoid non-functional proteins as the result of faults in this process. The precise removal of introns involves the recognition of a set of specific sequences within the pre-mRNA. Two types of spliceosomal introns have been described namely U2-type introns and U12-type introns. U2-type introns are canonical introns, which account for the vast majority (>99.5%) of spliceosomal introns and are removed by the U2-dependent spliceosome, also called the major spliceosome. For U2-
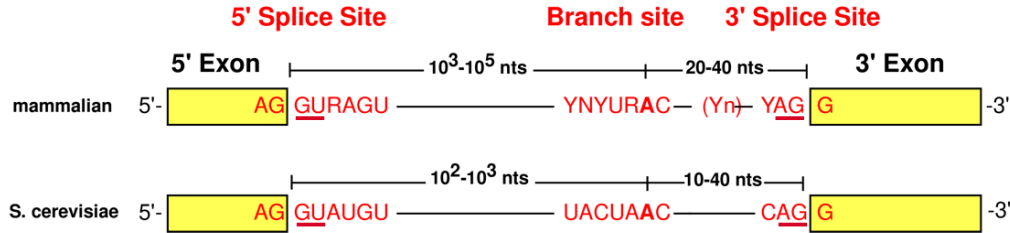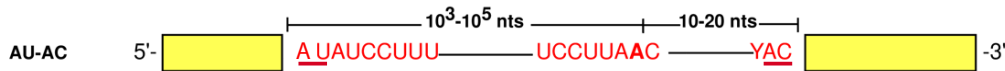
**Figure 1.1:** Introns are defined by conserved sequences in the pre-mRNA. The spliceosome recognizes the four conserved elements namely the 5'- splice site, the branch point, the polypyrimidine tract ($Y_n$) and the 3'-splice site. In this figure Y stands for pyrimidines, R stands for purines and N represents any nucleotide. Most introns are characterized by GU-AG boundaries (underlined). The U12 dependent introns use AU-AC instead (underlined).

type introns in higher eukaryotes, the 5' splice site is defined by an eight nucleotide sequence AG/GURAGU, in which GU is the invariable start of an intron, R represents purines and '/' indicates the exon/intron junction (21). The 3' splice site is indicated by the YAG/G sequence, in which Y stands for a pyrimidine and AG is invariable the end of an intron. The AG dinucleotide is preceded by 10-15 pyrimidines of the so-called polypyrimidine tract ($Y_n$) and the branch site, characterized by the YNYURAC sequence, in which the adenosine is highly conserved and is located 20-40 nt upstream from the 3' splice site (22). The tiny population of non-canonical U12-dependent introns is removed by the U12-dependent spliceosome, also named the minor spliceosome. Instead of using GU-AG termini, these introns usually have AU-AC at their termini (Figure 1.1).

Pre-mRNA splicing is catalyzed by a two-step stereospecific transesterification reaction (23). In the first step, the 2'-OH group of the conserved

branchpoint adenosine attacks the phosphate at the 5' exon/intron junction (the 5' splice site) resulting in a free 5' exon containing a 3' terminal OH group and a branched lariat intermediate which contains the intron and 3' exon. In the second step, the 3' OH of the 5' exon attacks the phosphate at the intron/3' exon boundary (the 3' splice site), thereby ligating the 5' and 3' exons and releasing the intron as a lariat molecule (24; 25; 26) (Figure 1.2).

### 1.2.2   The spliceosome

**The U snRNPs**

The spliceosome is a multi-megadalton ribonucleoprotein (RNP) enzyme, which consists of U snRNPs (<u>u</u>ridine-rich <u>s</u>mall <u>n</u>uclear <u>r</u>ibonucleo<u>p</u>roteins). These snRNPs contain at least one U snRNA (<u>u</u>ridine-rich <u>s</u>mall <u>n</u>uclear RNA) packaged with a number of common and specific proteins. A U snRNP is named after the U snRNA it contains. The major spliceosome consists of U1, U2, U5 snRNPs and the U4/U6 di-snRNP (27; 28), while their counterparts in the minor spliceosome are U11, U12, U5 snRNPs and U4atac/U6atac di-snRNP (29).

**The U snRNAs**   The RNA components of the spliceosome are phylogenetically conserved small uridine-rich RNAs (snRNAs), with conserved primary and secondary structures (28). For the U2 dependent spliceosome there are five snRNAs: U1, U2, U4, U5 and U6 snRNA, while in U12 dependent spliceosome the counterparts are U11, U12, U4atac, U5 (as in the U2 dependent spliceosome) and U6atac snRNA. With the exception of U6 and U6atac snRNAs, all U snRNAs are transcribed by RNA-polymerase II and capped at the 5'-end with a 2,2,7- trimethylguanosine cap ($m_3G$-cap) (30). The 5'-end $m_3G$-cap was found to serve as a signal for reimporting the U snRNPs into the nucleus (? ? ).U6 and U6atac snRNA, on the other hand, are tran-
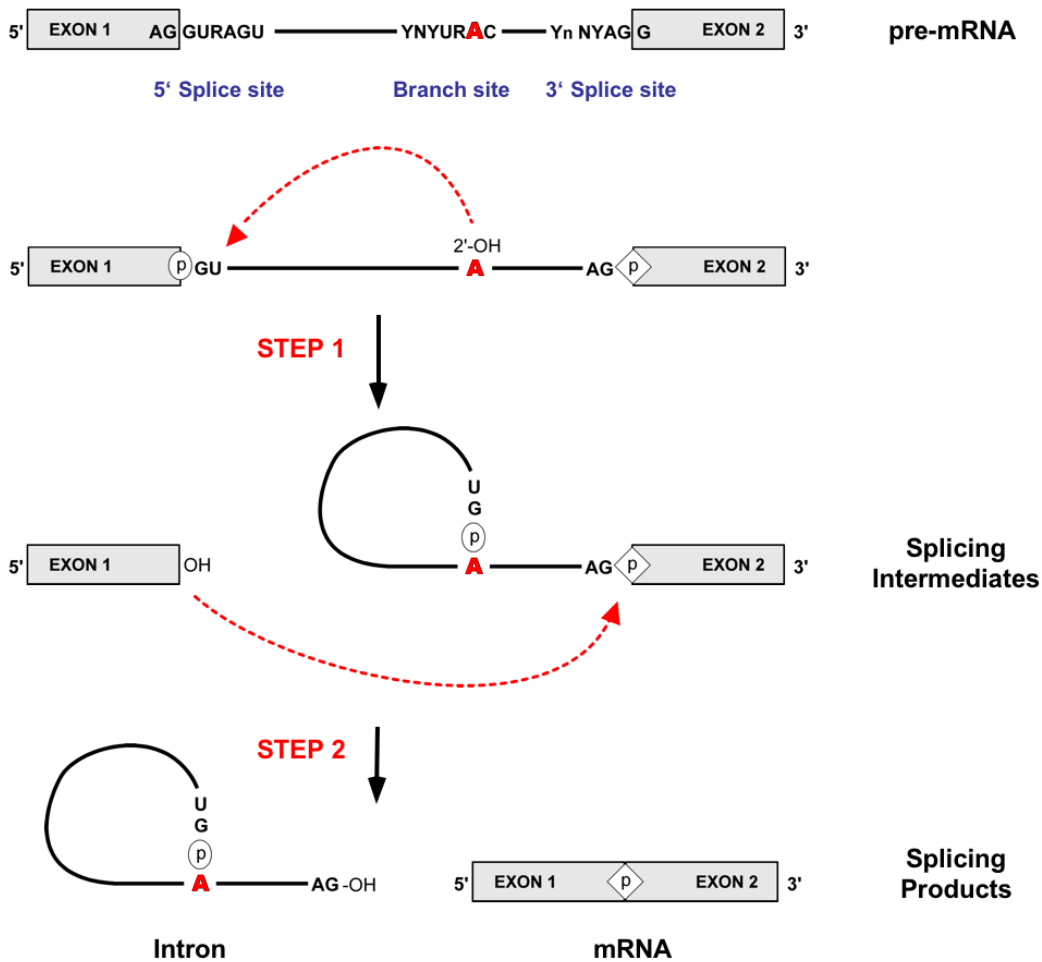
**Figure 1.2: Schematic representation of the splicing reactions.** In the first step, the 2'-OH group of the branch point adenosine attacks the phospho-ate of the conserved guanine at the 5'-splice site forming a 2'-5' phosphodiester bond. The junction between exon I and the intron is, thus, cleaved. This results a 2'-5' phosphodiester bond in the lariat RNA and an upsteam exon with a free 3'-OH group. In the second step, the 3'-OH group of exon 1 attacks the phosphodiester bond of the guanine at the 3'-splice site. This reaction results a free intron lariat and spliced exons (picture modified after a diagram provided by Dr. Berthold Kastner).

scribed by RNA-polymerase III and carry 5'-end r-Monomethylphosphates (29) as well as many other post-translational modifications like pseudouridine, 6-methyladenosine and 2' O-ribosemethyl (31). RNA-polymerase II transcribed snRNAs possess the so-called Sm-binding sites, which is characterized by a consensus sequence of $RAU3_6GR$ (R=purine) and are flanked by two hairpin loops. Sm-binding sites bind Sm-proteins. snRNAs have complex secondary structures with single and double stranded regions and multiple stem loop structures. The U4 and U6 snRNA, as well as the U4atac and U6atac snRNA form Y-shaped duplexes(27; 32; 33). Of the snRNA components of the spliceosome, U6 and U2 have the highest degree of sequence conservation, with U6 having the most highly conserved sequence. This reflects its extensive base-pairing with U4 and with U2 during the splicing reaction and with the intron itself (34; 35; 27) (Figure 1.3 and Figure 1.4).

**The protein composition of the U snRNPs**

**The common proteins.**   The snRNAs are packaged with a large number of proteins, several of them common to all snRNPs. Apart from U6 and U6atac snRNPs, all U snRNPs contain seven small polypeptides: B/B', D1, D2, E, F, and G. They form the group of the Sm proteins (37; 38; 39; 40). Sm proteins form a doughnut shaped structure and associate with the snRNAs at the Sm sites. They are essential for the biogenesis of U snRNPs (30). U6 and U6atac snRNAs lack Sm sites, but they associate with another group of proteins, called the Lsm-proteins (Like Sm) (41; 42). Similar to the Sm proteins, Lsm proteins Lsm2-Lsm8 form stable heteromeric complex and associate to the oligo-U-sequence at the 3'-end of the U6 and the U6atac snRNA.
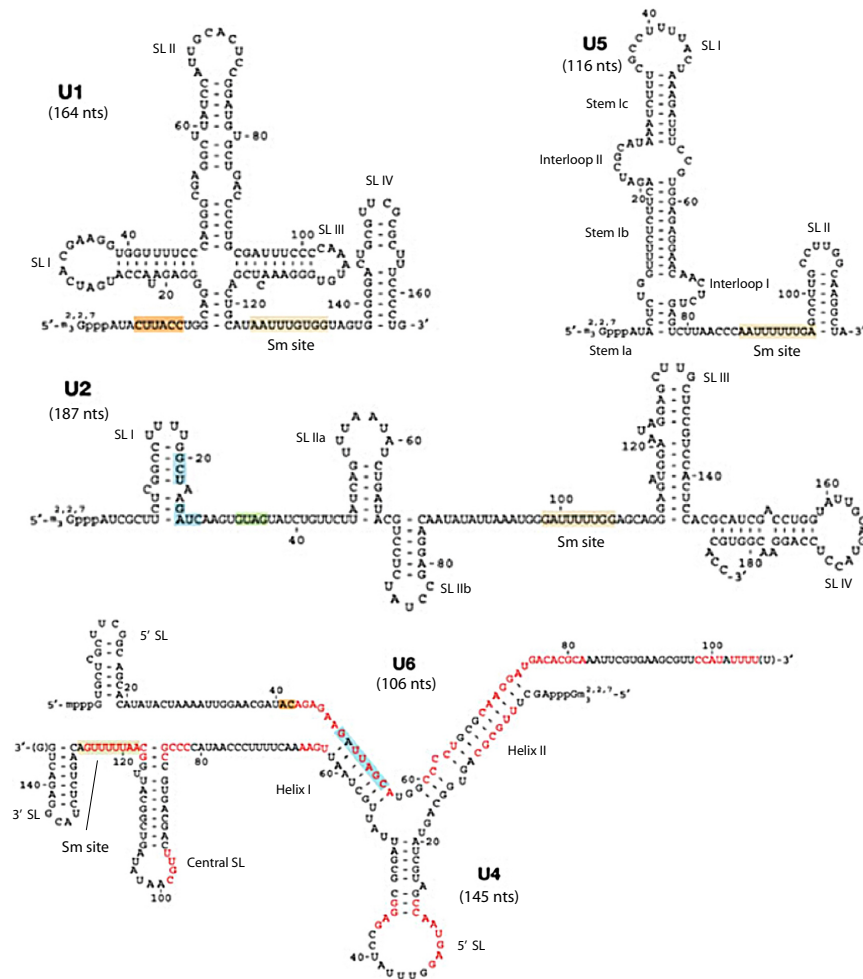
**Figure 1.3:** For the U2 dependent spliceosome there are five snRNAs: U1, U2, U4, U5 and U6 snRNA, in which U4 and U6 form duplex. The Sm-binding sites are shaded in light yellow. Coloured boxes indicate sequences which are predicted to be involved in intermolecular RNA-RNA interactions: orange for interactions with the 5'-splice site, green for interactions with the branch site and blue for U2/U6 helix I interactions. Sequences in red represent stretches of four or more identical nucleotides between U4-U6. Picture modified from figure 4 in the review by Patel and Steitz 2003 (36).
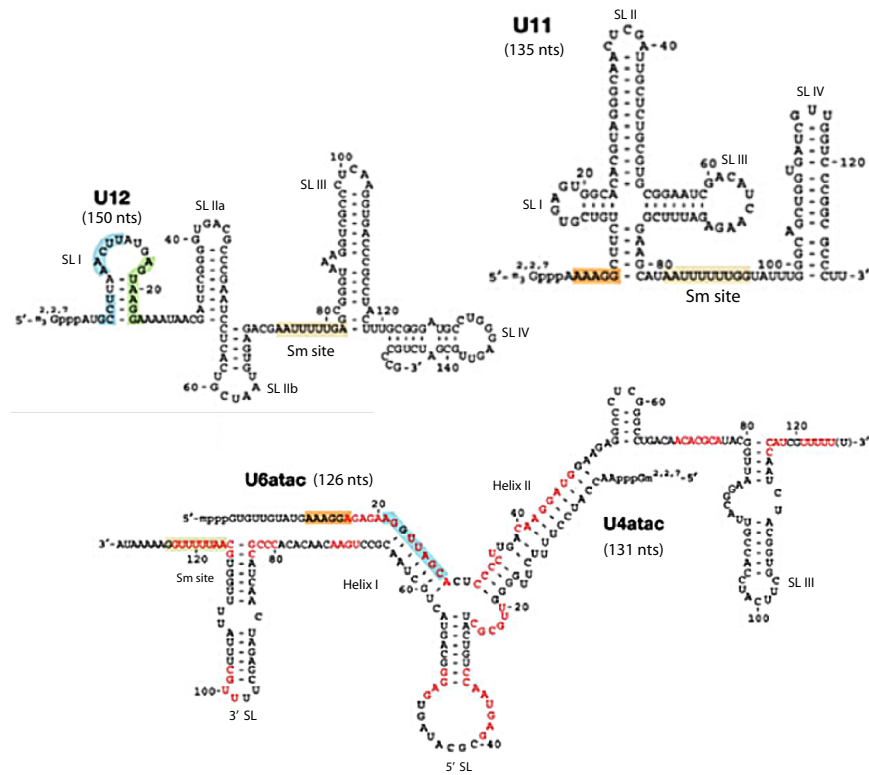
**Figure 1.4:** In U12 dependent spliceosome the counterparts are U11, U12, U4atac, U5 (as in the U2 dependent spliceosome) and U6atac snRNA. The Sm-binding sites are shaded in light yellow. Coloured boxes indicate sequences which are predicted to be involved in intermolecular RNA-RNA interactions: orange for interactions with the 5'-splice site, green for interactions with the branch site and blue for U12-U6atac helix I interactions. Sequences in red represent stretches of four or more identical nucleotides between U4atac-U6atac. Picture modified from figure 4 in the review by Patel and Steitz 2003 (36).

**The specific proteins.** Apart from the common proteins, snRNPs contain numerous specific proteins, which are unique for different snRNPs. These proteins are required to enable specific functions of the individual snRNPs during the splicing event (43). They are involved in splice site recognition and enable structural rearrangement in the snRNPs (44). The protein composition of the snRNPs has been investigated by biochemical purification and mass spectrometry and is shown in Figure 1.5 (45).

The 12S U1 snRNP specific proteins include 70K, U1A and U1C (46; 47). U1-70K and U1-C proteins stabilize the U1 snRNA base-pairing with the 5'-splice site. The 17S U2 snRNP comprises a much larger number of specific proteins, which are U2 A', U2 B" and the heteromeric protein complexes SF3a and SF3b (48; 49; 50). SF3a and SF3b proteins are found to directly associate with the branch site, which facilitate the binding of U2 snRNP and therefore the formation of the spliceosome (A-complex) (51). The 20S U5 snRNP specific proteins are U5-15K, U5-40K, U5-52K, hPrp28 (100K), hPrp6 (102K), hSnu114 (116K), U5-200K and hPrp8 (220 K) proteins(52; 53). They contain striking domain features and are found to be crucial for the splicing process and the structural rearrangement of the spliceosome during the course of splicing (54; 55; 43). For example, hPrp28 (U5-100K) and U5-200K(Brr2p in yeast), which belong to the DEXH/D-box family of RNA unwindases/RNPases, facilitate the integration of the tri-snRNP into the spliceosome. hSnu114 (U5-116K) protein, a homolog of the ribosomal elongation factor EF-2 and hPrp8 (U5-220K) are found to be involved in unwinding the U4/U6 duplex (56; 57). U5-200K is also found to be important for the disruption of the U4/U6 duplex. (58). In the 13S U4/U6 di-snRNP, apart from the U6 specific Lsm proteins, there are other specific proteins namely 15.5K (Snu13 in yeast), hPrp31 (61K), CypH (20K), hPrp4 (60K) and hPrp3 (90K). They are discussed in more detail later on in the thesis. The tri-snRNP contains three additional specific proteins, including 27K,

65K and 110K. The 110K (Snu66 in yeast) protein has been found to facilitate the re-association of the U4 and U6 snRNPs after splicing(59).

Apart from the U snRNPs there are many more non-snRNP proteins which contribute to form a functional spliceosome (44).

**The assembly of the spliceosome and the dynamics of the snRNAs**

The splicing reaction is a coordinated, dynamic process and involves the step-wise assembly of the spliceosome on the pre-mRNA, the two catalytic steps of the reaction and the dissociation as well as the recycling of splicing factors (25; 26). The catalytic center of the spliceosome is formed through the coordinated binding of the U snRNP particles and numerous non-U snRNP-splicing factors (60) (Figure 1.6). The first catalytic step involves the formation of the so-called E-complex (early complex), in which the U1 snRNP contacts the 5' splice site through base-pairing with the intronic sequence of the pre-mRNA in an ATP-independent manner. The recruitment of the U1 snRNP is mediated by the ASF/SF2 splicing factor and its interaction with the 5'-splice site is stabilized by the U1C protein (61). The binding of U1 snRNP is required for the association of other snRNPs. In a subsequent ATP-dependent step, the 17S U2 snRNP recognizes and associates with the branch point forming the so-called A-complex (also called the pre-spliceosome). The base-pairing of the U2 snRNP with the branch point sequence forms a short RNA duplex and leaves the branch point adenosine unpaired and bulged out. Thus, the branch point adenosine is favorably positioned for the subsequent nucleophilic attack (60; 62). The splice factors U2AF as well as SF3 and SF3b are essential for the successful binding of the U2 snRNP(63; 64; 65; 66; 67). In the next step, the pre-formed 25S [ U4/U6.U5 ] tri-snRNP is integrated to form the B-complex through the mediation from tri-snRNP specific proteins 65K and 110K (68). Dramatic structural changes then result in the release of U4 snRNA from the U4/U6 snRNA duplex and the dissociation of

| U snRNP proteins | | app. Mr kDa | 12 S U1 | 17 S U2 | 20 S U5 | 13 S U4/U6 | 25 S U4/U6.U5 | sequence motifs | yeast homologue |
|---|---|---|---|---|---|---|---|---|---|
| Common proteins | B/B' | 28/29 | ● | ● | ● | ● | ● | | SMB1 |
| | D3 | 18 | ● | ● | ● | ● | ● | | SMD3 |
| | D2 | 16.5 | ● | ● | ● | ● | ● | Sm-motif | SMD2 |
| | D1 | 16 | ● | ● | ● | ● | ● | | SMD1 |
| | E | 12 | ● | ● | ● | ● | ● | | SME1 |
| | F | 11 | ● | ● | ● | ● | ● | | SMX3 |
| | G | 9 | ● | ● | ● | ● | ● | | SMX2 |
| U1 snRNP | 70K | 70 | ● | | | | | RRM | SNP1 |
| | A | 34 | ● | | | | | 2 RRMs | MUD1 |
| | C | 22 | ● | | | | | Zn-finger | YHC1 |
| U2 snRNP | A' | 31 | | ● | | | | Leu-rich | LEA1 |
| | B" | 28.5 | | ● | | | | 2 RRMs | MSL1 |
| | SF3a120 | 120 | | ● | | | | 2 SURP, UBQ | PRP21 |
| | SF3a66 | 66 | | ● | | | | Zn-finger | PRP11 |
| | SF3a60 | 60 | | ● | | | | Zn-finger | PRP9 |
| | SF3b155 | 160 | | ● | | | | HEAT repeats | HSH155 |
| | SF3b145 | 150 | | ● | | | | SAP, Pro-rich | CUS1 |
| | SF3b130 | 120 | | ● | | | | CPSF A | RSE1 |
| | SF3b49 | 53 | | ● | | | | 2 RRMs | HSH49 |
| | SF3b14a | 15 | | ● | | | | RRM | SNU17? |
| | SF3b14b | 15 | | ● | | | | Cys-rich | RDS3 |
| | SF3b10 | 9 | | ● | | | | - | YSF3 |
| | hPrp5 | 140 | | ● | | | | DEAD box | PRP5 |
| | SR140 | 140 | | ● | | | | RRM, SWAP | - |
| | CHERP | 130 | | ● | | | | SWAP, G-patch | - |
| | hPrp43 | 90 | | ● | | | | DEAH box | PRP43 |
| | SPF45 | 50 | | ● | | | | G-patch, RRM | - |
| | SPF31 | 33 | | ● | | | | DnaJ domain | HLJ1? |
| | SPF30 | 31 | | ● | | | | Tudor | - |
| U5 snRNP | hPrp8 | 220 | | | ● | | ● | - | PRP8 |
| | U5-200K | 200 | | | ● | | ● | 2 DExH | BRR2 |
| | U5-116K | 116 | | | ● | | ● | G-domain | SNU114 |
| | hPrp6 | 102 | | | ● | | ● | TPR repeats | PRP6 |
| | hPrp28 | 100 | | | ● | | ● | DEAD +RS | PRP28 |
| | U5-52K | 52 | | | ● | | ● | GYF domain | LIN1 |
| | U5-40K | 40 | | | ● | | ● | WD40 | - |
| | U5-15K | 15 | | | ● | | ● | thioredoxin fold | DIB1 |
| U4/U6 snRNP | hPrp3 | 90 | | | | ● | ● | PWI, +dsRNA bdg. | PRP3 |
| | hPrp31 | 61 | | | | ● | ● | Nop homol. domain | PRP31 |
| | hPrp4 | 60 | | | | ● | ● | WD40 | PRP4 |
| | CypH | 20 | | | | ● | ● | cyclophilin | - |
| | U4/U6-15.5K | 15.5 | | | | ● | ● | RNA bdg. motif | SNU13 |
| | Lsm2 | 10 | | | | ● | ● | | LSM2 |
| | Lsm3 | 15 | | | | ● | ● | | LSM3 |
| | Lsm4 | 15 | | | | ● | ● | | LSM4 |
| | Lsm5 | 10 | | | | ● | ● | Sm-motif | LSM5 |
| | Lsm6 | 8 | | | | ● | ● | | LSM6 |
| | Lsm7 | 13 | | | | ● | ● | | LSM7 |
| | Lsm8 | 13 | | | | ● | ● | | LSM8 |
| tri-snRNP | 110K | 110 | | | | | ● | RS-domain | SNU66 |
| | 65K | 65 | | | | | ● | RS-domain | SAD1 |
| | 27K | 27 | | | | | ● | RS-domain | - |

**Figure 1.5:** The snRNPs contain a set of common and specific proteins which belong to various protein families. They contain domain features. A yeast protein marked with '?' indicates that it is currently not clear whether this is the true ortholog to the human protein. (picture kindly provided by Dr. Reinhard Rauhut, adapted from (44)).

the U1 snRNP from the 5'-splice site (69; 70). The active spliceosome (B*-complex) thus formed has U2 and U6 snRNAs duplexed, which presumably contributes to the formation of the catalytic center (71; 72; 35; 73). After the first transesterification reaction has occurred, the spliceosome arrives at the C-complex status (74). In the second catalytic step, the two exons are linked together through a second transesterification reaction and the intron is released as a lariat. The intron, unlike the mature mRNA is retained in the nucleus and is subjected to full degradation. The freed U snRNPs after dissociation of the spliceosome are recycled for a new cycle of splicing (75).

Despite the dependence of spliceosome on a large number of proteins for its catalytic activation, it very likely functions essentially as a ribozyme (RNA enzyme), with its active site composed of RNA. The snRNAs form a highly dynamic RNA network, which is extensively restructured during maturation of the spliceosome. The RNA-RNA interactions during the splicing cycle are crucial for splice site recognition, juxtaposing the reactive groups of the pre-mRNA and the two transesterification steps (76; 54; 25; 77). The most dramatic structural rearrangement happens to the U6 snRNA. It is first delivered to the pre-mRNA in a repressed state in the context of the U4/U6-U5 tri-snRNP, in which catalytically important regions of the U6 snRNA are base-paired to the U4 snRNA (33; 78; 32; 28). During spliceosome activation this U4-U6 interaction is disrupted and the U4 snRNA is released from the complex. Afterwards, the U6 snRNA forms short duplexes with the U2 snRNA and the pre-mRNA substrate, which subsequently support splicing (Figure 1.7).

Evidently, knowledge of the structures of spliceosomal components is of great importance for a better understanding of how the spliceosome functions. Long term on-going efforts in X-ray crystallography, NMR and electron microscopy have provided some insights. Among the snRNPs, U1 is the structurally best characterized of all, with its three-dimensinoal struc-
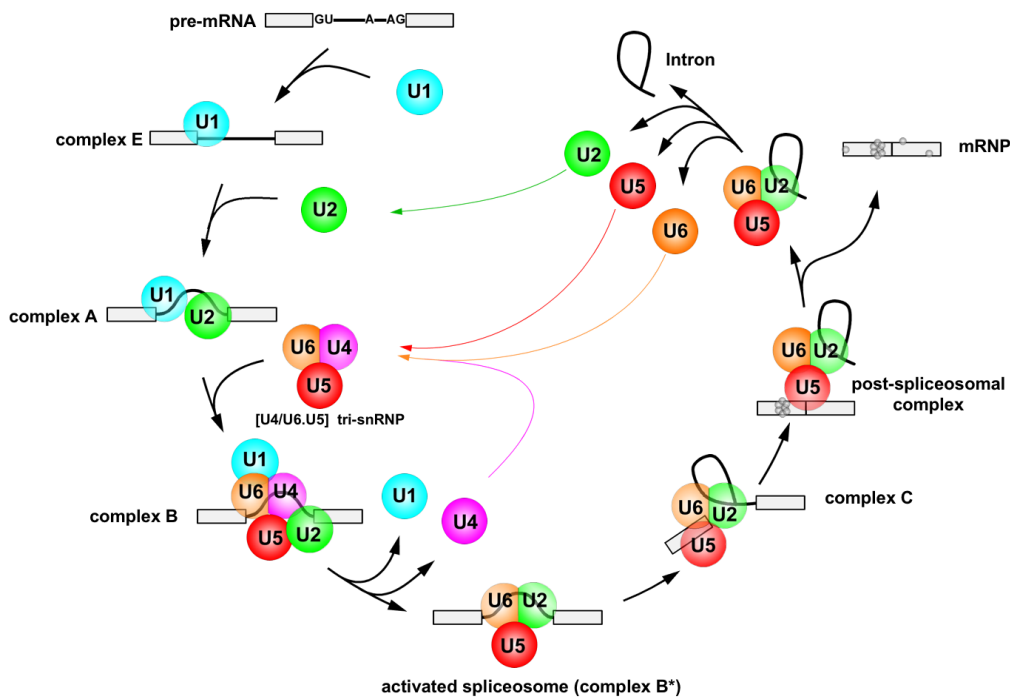
**Figure 1.6: Schematic depiction of the spliceosome assembly pathway.** The spliceosome carries out the splicing event in a stepwise, coordinated manner through complex protein-protein, protein-RNA and RNA-RNA interactions. In the first step, the U1 snRNP recognizes and base-pairs with the 5'-splice site forming the E-complex. Then the U2 snRNP forms a stable duplex with the branch point sequence (A-complex). After the integration of the [U4/U6.U5] tri-snRNP, the B-complex is assembled. The subsequent dissociation of U1 and U4 snRNP allows the formation of the activated spliceosome (B*-complex), in which U2 and U6 snRNAs are base-paired and presumably form the active catalytic center. After the first transesterification reaction, the C-complex hosts the second splicing reaction, which splices the two exons together and excises the intron in a lariat form. The matured mRNA is then exported into the cytoplasm and to the ribosome as the template for protein biosynthesis. The free U snRNPs are recycled for the next splicing cycle while the intron is retained in the nucleus and are subjected to subsequent degradation (picture kindly provided by Dr. Berthold Kastner).
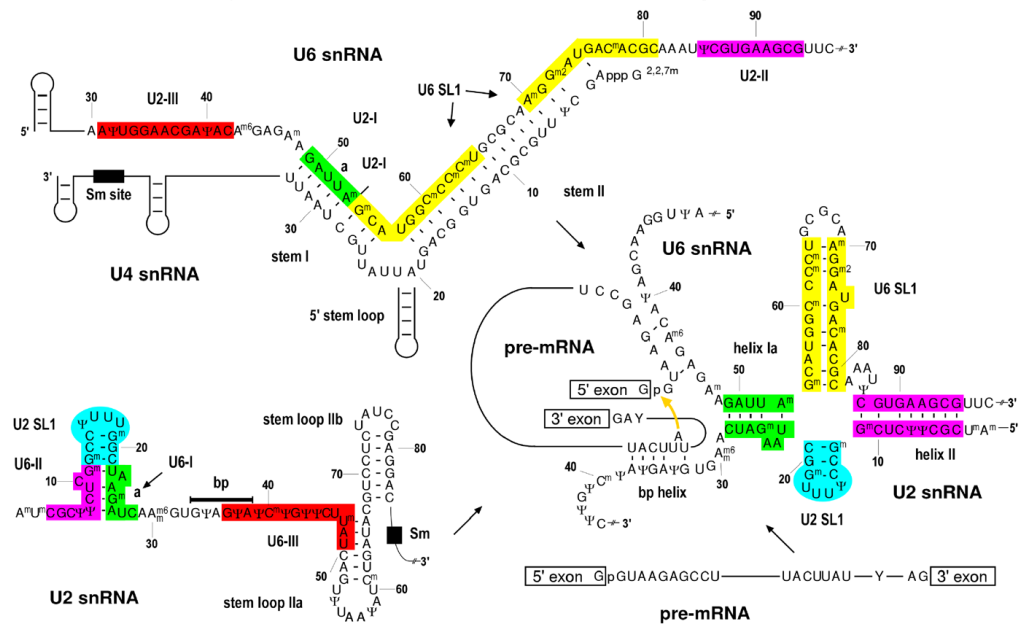
**Figure 1.7:** Splicing is achieved through dynamic interactions among the spliceosomal snRNAs. Before the first step of the splicing reaction, U4 and U6 snRNAs form extensive base-pairs with each other. However, U4/U6 duplex dissociate completely in order to active the splicesome. In the activated spliceosome, U6 forms duplex with U2 snRNA instead as well as the intron region of the pre-mRNA substrate (picture kindly provided by Dr. Berthold Kastner).

ture revealed at 10 Å resolution (48; 39; 79). Much information has been obtained about the general architecture of the U2 snRNP, but a higher-order structure is still lacking (44; 80; 81). Currently, the higher-order structures and subunit structures of [U4/U6.U5] tri-snRNP and U4/U6 di-snRNP are the least characterized of all snRNPs. Therefore, little is known about how the U4/U6 di-snRNP or the [U4/U6.U5] tri-snRNP are designed to sustain spliceosome activation. In the following section, structural features of the U4 snRNA and of two U4/U6 di-snRNP specific proteins, will be discussed in more detail. In the scope of this work, investigation of the intermolecular interactions in the ternary complex constituted by these two proteins and the RNA was carried out.

## 1.2.3 The U4/U6 snRNP

U4 snRNA is not found as a free molecule species. Instead it always forms extensive base-pairing with U6 snRNA before it enters the splicing machinery. U4 and U6 form two intermolecular helices (helix I and helix II), which are separated by an intra-molecular 5'-terminal stem loop (U4 5'-SL). The branching region of this Y-shaped duplex contains the three-way juction (27). The U4 5'-SL has been demonstrated by mutational studies in Xenopus oocytes and in HeLa in vitro splicing system to be essential for spliceosome assembly (82; 83; 84) . It serves as a binding site for the highly conserved U4/U6-15.5K protein (85). The U4 5'-SL contains a long stem I (seven canonical Watson-Crick base pairs, one U-G wobble base pair and one bulged-out adenosine), an asymmetric (5+2) internal loop, which will be discussed later in detail, a short stem II (two G-C Watson-Crick base pairs) and a penta-loop (Figure 1.8).
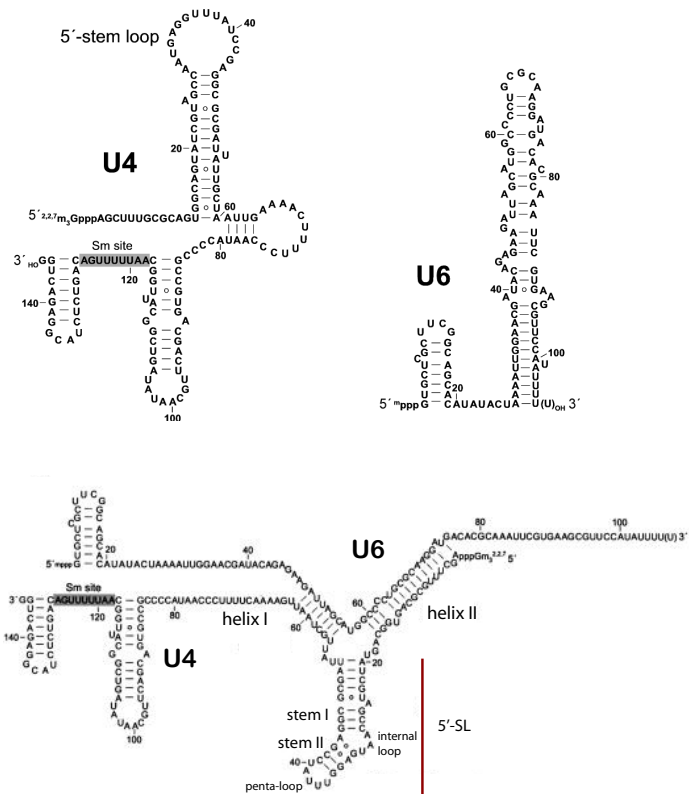
**Figure 1.8:** The postulated secondary structures of the U4 and U6 snRNA not in duplex are shown. These two snRNAs are found to form a Y-shaped duplex, with helix I and helix II interrupted by the U4 5'-SL. The (5+2) internal loop of the U4 5'-SL is indicated in black boxes. The Sm site indicated, exists in U4 snRNA and is absent in U6 snRNA.

```
15.5kD human         1 M T E A D V N P K A Y P L A D A H L T K K L L D L V Q Q S C N Y K Q L R K G A N E A T K
15.5kD C.elegans     1 M A D D G V N P K A F P L A D T N L S Q K L M D L V Q Q A M N Y K Q L K K G A N E A T K
15.5kD S.pombe       1     M S V N P K A F P L A D S G L T Q Q I L D L V Q Q A S H Y K Q L R K G A N E A T K
Snu13p S.cerevisiae  1   M S A P N P K A F P L A D A A L T Q Q I L D V V Q Q A A N L R Q L K K G A N E A T K
15.5kD D.melanogaster 1   M T E E V N P K A F P L A D A Q L T A K I M N L L Q Q A L N Y N Q L R K G A N E A T K
15.5kD A.thaliana    1 M T V E G V N P K A Y P L A D S Q L S I T I L D L V Q Q A T N Y K Q L K K G A N E A T K

15.5kD human        45 T L N R G I S E F I V M A A D A E P L E I I L H L P L L C E D K N V P Y V F V R S K Q A
15.5kD C.elegans    46 T L N R G I S E I I V M A A D A E P L E I L L H L P L L C E D K N V P Y V F V R S K A A
15.5kD S.pombe      43 T L N R G I S E F I V M A A D T E P I E I L L H L P L L C E D K N V P Y V F V P S K A A
Snu13p S.cerevisiae 44 T L N R G I S E F I I M A A D C E P I E I L L H L P L L C E D K N V P Y V F V P S R V A
15.5kD D.melanogaster 45 T L N R G L A D I V V L A G D A E P I E I L L H L P L L C E D K N V P Y V F V R S K Q A
15.5kD A.thaliana   46 T L N R G I S E F I V M A A D T E P L E I L L H L P L L A E D K N V P Y V F V P S K Q A

15.5kD human        89 L G R A C G V S R P V I A C S V T I K E G S Q L K Q Q I Q S I Q Q S I E R L L V
15.5kD C.elegans    90 L G R A C G V T R P V I A A S I T Q N E G S Q L K S Q I Q K I K E D V E K L L I
15.5kD S.pombe      87 L G R A C G V S R P V I S A S I T T N E A S D L L P Q I Q A L K L A I E K L L I
Snu13p S.cerevisiae 88 L G R A C G V S R P V I A A S I T T N D A S A I K T Q I Y A V K D K I E T L L I
15.5kD D.melanogaster 89 L G R A C G V S R P I V A C S V T T N E G S Q L K S Q I T S I Q Q E I E R L L V
15.5kD A.thaliana   90 L G R A C D V T R P V I A C S V T S N E A S Q L K S Q I Q H L K D A I E K L L I
```

**Figure 1.9:** The sequence of human 15.5K protein is aligned with its or-
thologues from *Caenorhabditis elegans*, *Schizosaccharomyces pombe*, *Saccha-
romyces cerevisiae*, *Drosophila melanogaster* and *Arabidopsis thaliana*. iden-
tical residues and conserved residues are boxed in black and grey respectively
(picture taken from Nottrott *et al* 1999 (85)).

### The 15.5K protein

The 128 amino acid long human 15.5K protein does not contain a canonical
RNA-binding domain like the RNP domain. Instead, it uses a novel RNA-
binding domain (85). 15.5K is evolutionarily highly conserved. In a wide
range of species from yeast *Saccharomyces cerevisiae* (called Snu13p) to hu-
man, it shows 71-77% sequence identity and up to 83-89% sequence similarity
(Figure 1.9). This is a strong indication for the functional importance of the
15.5K protein. After the first characterization of this protein, it was found
to be required for the first step of splicing (85).

The crystal structure of 15.5K associated with a 22nt RNA construct of
the U4 5'-SL was solved to a resolution of 2.9 Å(86), whereas 15.5K alone did
not lead to crystallization. In this crystal structure, 15.5K folds into a single
globular domain consisting of alternating $\alpha$-helices and $\beta$-sheets to form an
$\alpha$-$\beta$-$\alpha$ sandwich structure (Figure 1.10). Three of the four $\beta$ stands, $\beta 1$, $\beta 2$,
$\beta 4$, run anti-parallel to each other and $\beta 3$ is parallel to $\beta 2$. Helics $\alpha 1$, $\alpha 4$

and $\alpha 5$ are found on one side of the $\beta$ strands, while $\alpha 2$ and $\alpha 3$ are located on the other side. The $\alpha$-$\beta$-$\alpha$ sandwich structure is a very common protein fold. Indeed 15.5K has a number of structural homologs, which include the ribosomal S12, L7ae and L30 (87). Ribosomal L7ae (Nhp2 in eukaryotes) exhibits a dual function and is also implicated in the H/ACA snoRNP (small nucleolar ribonucleoproteins) assembly (57; 88; 89). The RNA binding site of 15.5K consists of amino acid residues located in two $\alpha$ helices ($\alpha 2$ and $\alpha 4$), one $\beta$ strand ($\beta 1$), and three different loops ($\beta 1$-$\alpha 2$, $\beta 2$-$\alpha 3$, and $\alpha 4$-$\beta 4$). The predominant contacts of these amino acids locate within the RNA on the (5+2) internal loop and the stem II phosphate backbone. The bulged out U31 forms the most extensive contacts with the 15.5K RNA bind site. The structure of the U4 5'-SL revealed a characteristic sharp kink in the phosphate backbone.

With the determination of atomic structures of the *Haloarcula maris-mortui* large ribosomal subunit (89) and the *Thermus thermophilus* small ribosomal subunit (90; 91), the size of the structure database of RNAs has been greatly increased (92). The sharp kink motif revealed in U4 5'-SL has also been found in eight ribosomal RNAs and has been named the K-turn motif. It is characterized by a two-stranded, helix-internal loop-helix motif. The first helical stem (stem I in the case of U4 5'-SL) ending at the internal loop with two Watson-Crick base pairs, typically C-Gs, is called the 'canonical stem' or the 'C-stem', while the second helical stem (stem II in the case of U4 5'-SL), which follows the internal loop, starts usually with two non-Watson-Crick base pairs (typically sheared G-A base pairs) and is therefore termed the 'non-canonical stem' or the 'NC-stem'. The internal loop between the helical stems is always asymmetrical and usually has three unpaired nucleotides on one strand while none on the other(Figure 1.11). The 5'-most nucleotide of three unpaired nucleotides stacks on the C-stem (A29 in the case of U4 5'-SL), the second nucleotide extends to stack on the NC-stem
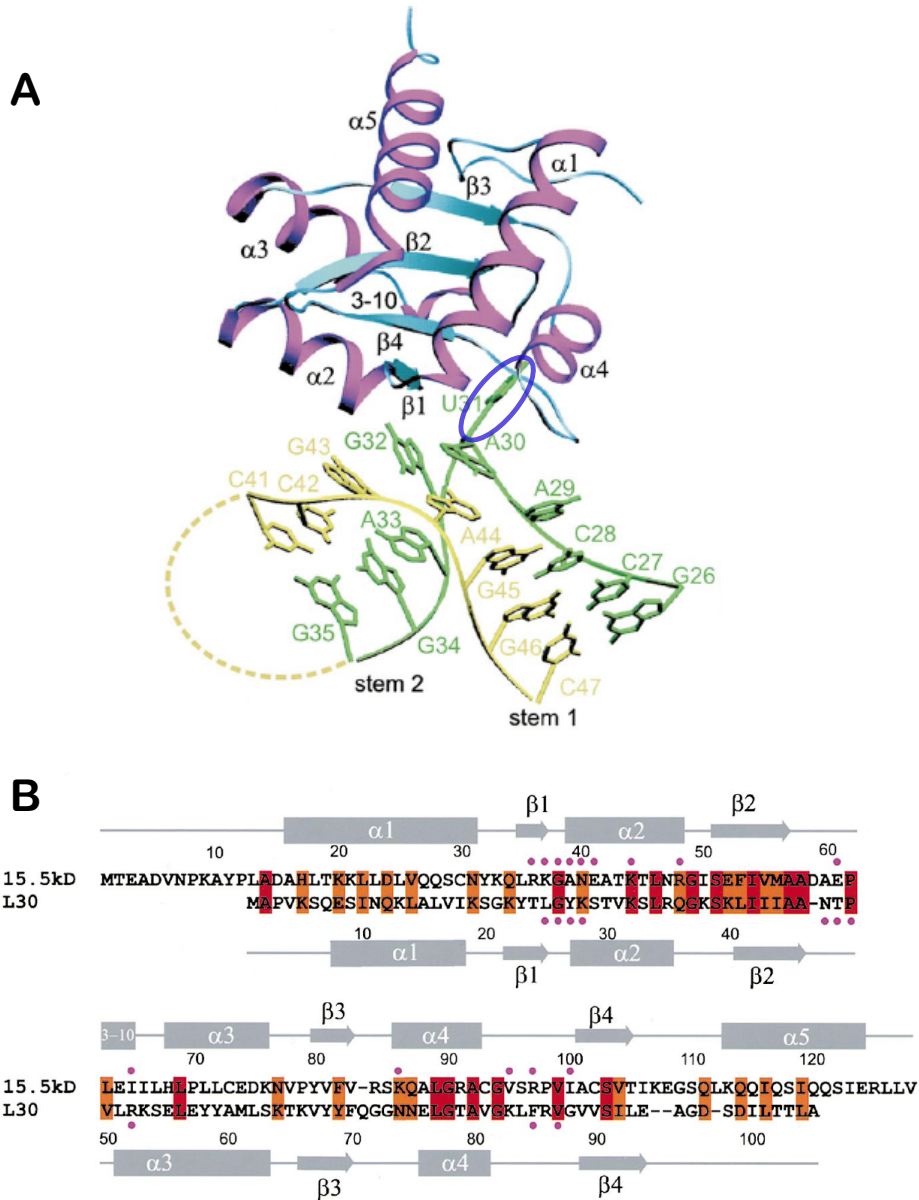
**Figure 1.10:** A. Crystal structure of 15.5K bound to a 22nt construct of the U4 5'-SL was solved by Vidovic *et al* to a 2.9 Å resolution. 15.5K folds into a single globular domain consisting of alternating $\alpha$-helices and $\beta$-sheets to form an $\alpha$-$\beta$-$\alpha$ sandwich structure. The RNA showed a characteristic sharp kink in the phosphate backbone with U31 (indicated in blue ellipse) bulging out and forming extensive contacts with the protien. B. Sequences of 15.5K and its homolog L30 in the ribosome are aligned. Secondary structures are indicated in grey. Amino acids, which form direct RNA contacts are indicated with a red dot (picture taken from Vidovic *et al* 2000 ([86])).

(A30 in U4 5'-SL) and the third nucleotide protrudes into solution (U31 in U4 5'-SL), which allows extensive contact with the protein. The U4 5'-SL, thus, exhibits all hallmarks of K-turns. Moreover, this motif is conserved in U4atac snRNA and also appears in a number of snoRNAs (small nucleolar RNAs). The examples, which will be discussed in detail in later chapters, include the U3-specific box B/C, box C'/D motif and the U14 box C/D motif. Interestingly 15.5K also associates with these snoRNAs and functions as an important player in the assembly of the corresponding snoRNPs. Recent studies using FRET measurement and molecular dynamics simulation methods suggested in the case of U4 5'-SL that the K-turn is flexible and 15.5K binds to the U4 5'-SL in an induced-fit manner and restructures the U4 5'-SL into the sharp kink (93; 94). Whether or not there are structural changes also present in 15.5K protein upon binding is not known. The results from our NMR studies concerning this question will be shown in this thesis.

The important role of 15.5K in U4/U6 di-snRNP assembly came to light with findings from Nottrott *et al* (95). 15.5K was identified as a nucleation factor for the di-snRNP assembly. The prior binding of 15.5K to the U4 5'-SL is a precondition for the further association of other U4/U6 specific proteins, namely hPrp31 and the CypH/hPrp4/hPrp3 protein complex (Figure 1.12). hPrp31 and CypH/hPrp4/hPrp3 can, however, associate independently of each other to the primary RNP formed by 15.5K and U4 snRNA in a non-cooperative manner. From combined mutational studies, pull-downs and UV cross-linking investigations, it was shown that hPrp31 forms direct contacts with U4 5'-SL, while CypH/hPrp4/hPrp3 seem to require the U4 5'-SL as well as the helix II of the U4/U6 duplex for efficient binding. hPrp3 was identified as forming direct contact with the U4 5'-SL and with part of the helix II on the U6 snRNA. Both hPrp31 and hPrp3 contain novel RNA-binding domains. hPrp31 belongs to the Nop domain family of RNA binding proteins, which will be discussed in detail in the following section, while hPrp3 is found
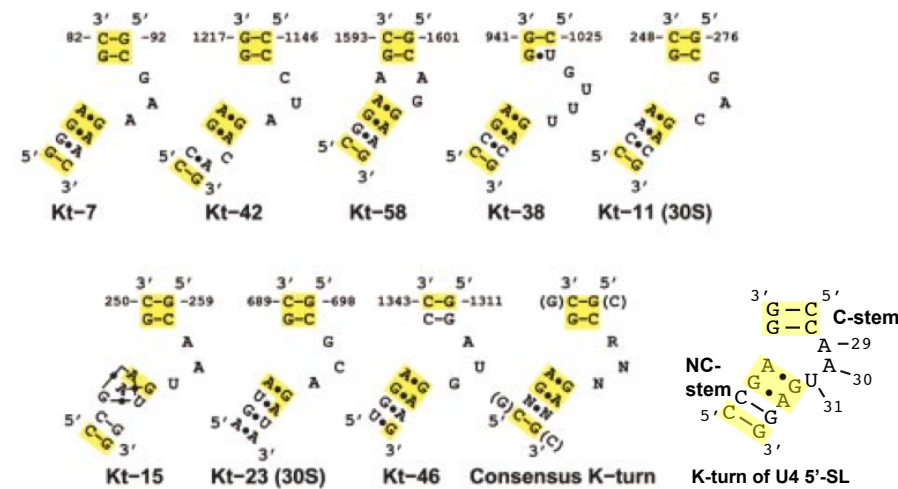
**Figure 1.11:** Secondary structures of the eight K-turns found in the *H. marismortui* 50S and *T. thermophilus* 30S subunit structures as well as the derived consensus sequence. Also shown is the K-turn of the U4 snRNA. Dots between nucleotides indicate miss match base-pairing. The yellow boxes indicate nucleotides which match the consensus sequence (picture modified from figure 1 of Klein *et al* 2001 (92)).
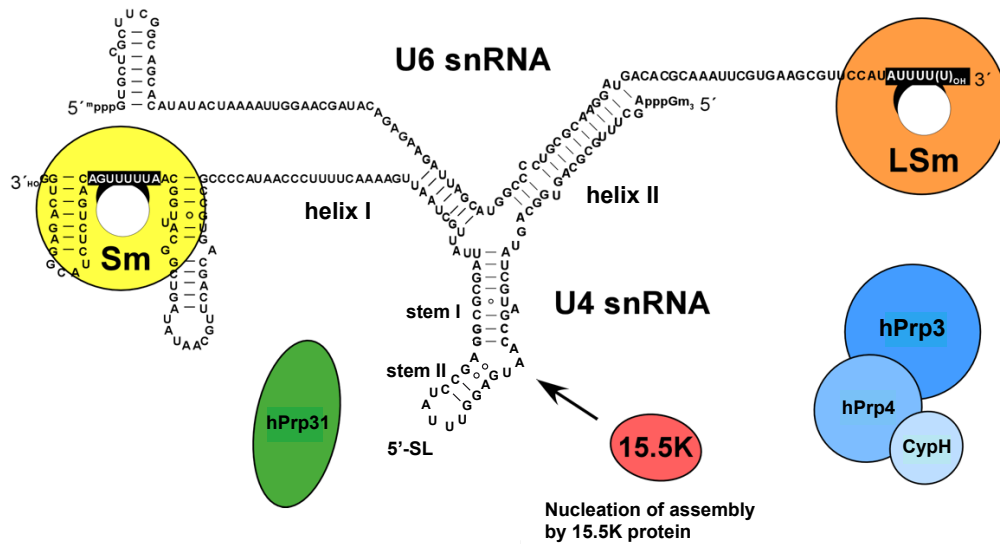
**Figure 1.12:** The U4/U6 specific proteins assemble onto the U4/U6 duplex following a hierarchical pathway. 15.5K is found to be the nucleation factor. Its prior binding to the U4 5'-SL is required for the association of other U4/U6 specific proteins namely hPrp31, CypH/hPrp4/hPrp3 protein complex. The Sm and LSm protein are indicated as yellow and orange rings respectively. Sm protein associate with the Sm-site on U4 snRNA and LSm protein binds specifically to the U6 snRNA (picture modified from the diagram by Dr. Berthold Kastner).

to contain a putative ds-RNA binding motif (96). However, in the absence of 15.5K, hPrp31 and CypH/hPrp4/hPrp3 are not able to associate with the U4/U6 snRNA. This hierarchical assembly pathway of U4/U6 specific proteins has also been observed for the U4 atac/U6atac snRNP. Intriguingly 15.5K is also implicated in snoRNP assembly as a nucleation factor for the binding of other specific proteins of that particular snoRNP. Since snoRNPs are involved in pre-rRNA (ribosomal RNA) processing, 15.5K forms a fascinating link between pre-mRNA and pre-rRNA processing.
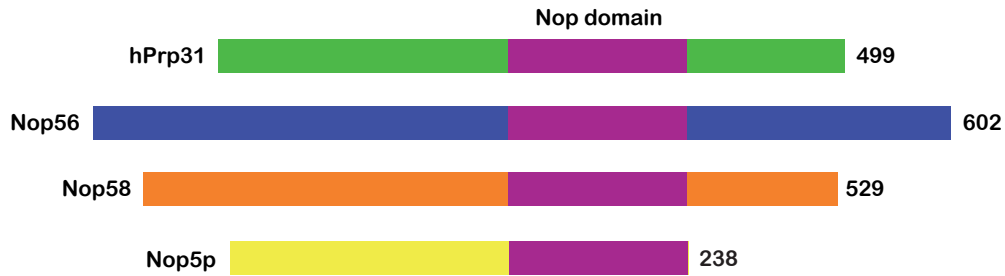
**Figure 1.13:** hPrp31, Nop56 and Nop58 (Nop5p in archaea)belong to the Nop family of proteins. hPrp31 is a specific protein of the U4/U6 di-snRNP, whereas Nop56 and Nop58 (Nop5p) are associated with box C/D snoRNPs. They share the homologous Nop domain, which is indicated in purple.

### The hPrp31 protein

The 499 amino-acid-long hPrp31 protein (human pre-mRNA processing factor 31) shows clear homology to two box C/D snoRNP-core proteins namely Nop56 and Nop58 (Nop5p in archaea), where Nop stands for nucleolar protein (97; 98; 99; 100). Nop56, Nop58 proteins were found to play crucial roles in pre-rRNA processing, which will be discussed in more detail in the following section. Although recruited to different primary RNPs, the hPrp31 and Nop56/Nop58 proteins exhibit significant sequence similarity, in particular over a  120 residue region termed the Nop domain (hPrp31$^{215-333}$) (Figure 1.13).

It was found that hPrp31 tightly associates with U4/U6 snRNP and is required for pre-mRNA splicing *in vitro* (99). hPrp31 interacts with the hPrp6 U5 protein, thus froming an important stabilizing bridge between the U4/U6 di-snRNP and the U5 snRNP during tri-snRNP formation. In yeast Prp31 is also suspected to anchor the tri-snRNP to the U2 snRNP (Figure 1.14 A) (101; 99; 102). The apparent discrepancy between the sequence similarity and functional diversity of the Nop domains in hPrp31, Nop56 and Nop58 poses fundamental questions about the function and binding mode(s) of the

Nop domain. hPrp31 is also of medical interest. Two missense mutations (A194E, A216P) in the gene of hPrp31 are found to be linked to the autosomal dominant retinitis pigmentosa (adRP), a progressive disease of the retina, which leads to degeneration of the rod photoreceptors of the eye and affects about one in 3500 individuals (103; 98; 104; 105; 106) (Figure 1.14 B). This finding further raises the importance of understanding the structure, function and binding mode of hPrp31. The highly similar Nop domains in hPrp31, Nop56 and Nop58 proteins and the involvement of 15.5K in both snRNP and snoRNP formation link pre-mRNA and pre-rRNA processing together.

## 1.3    Pre-rRNA processing and the snoRNPs

### 1.3.1    pre-rRNA processing

The biosynthesis of ribosomes is one of the major metabolic pathways in all cells (107). Ribosomes are large RNP particles, consisting of two unequal subunits (60S and 40S), that carry out protein synthesis. Eukaryotic ribosomes contain one copy of each of the four rRNAs, namely the 5S, 5.8S, 18S and 25-28S rRNAs, and about 75 different ribosomal proteins. Ribosome synthesis, rRNA transcription and maturation take place in the nucleolus, which is a specialized nuclear compartment. The nascent rRNA transcript from a rDNA unit by RNA polymerase I is a single 47S poycistronic precursor rRNA (pre-rRNA), which contains the sequences for the mature ribosomal RNAs (18, 5.8S, 25  28S rRNA), two external transcribed spacers (ETS) and two internal transcribed spacers (ITS) (108; 109; 110). The 47S pre-rRNA is processed into the 18S rRNA of the 40S (small) ribosomal subunit and the 5.8S, 28S rRNAs of the 60S (large) ribosomal subunit (Figure 1.15A), while the 5S rRNA, which also belongs to the large ribosomal subunit, is
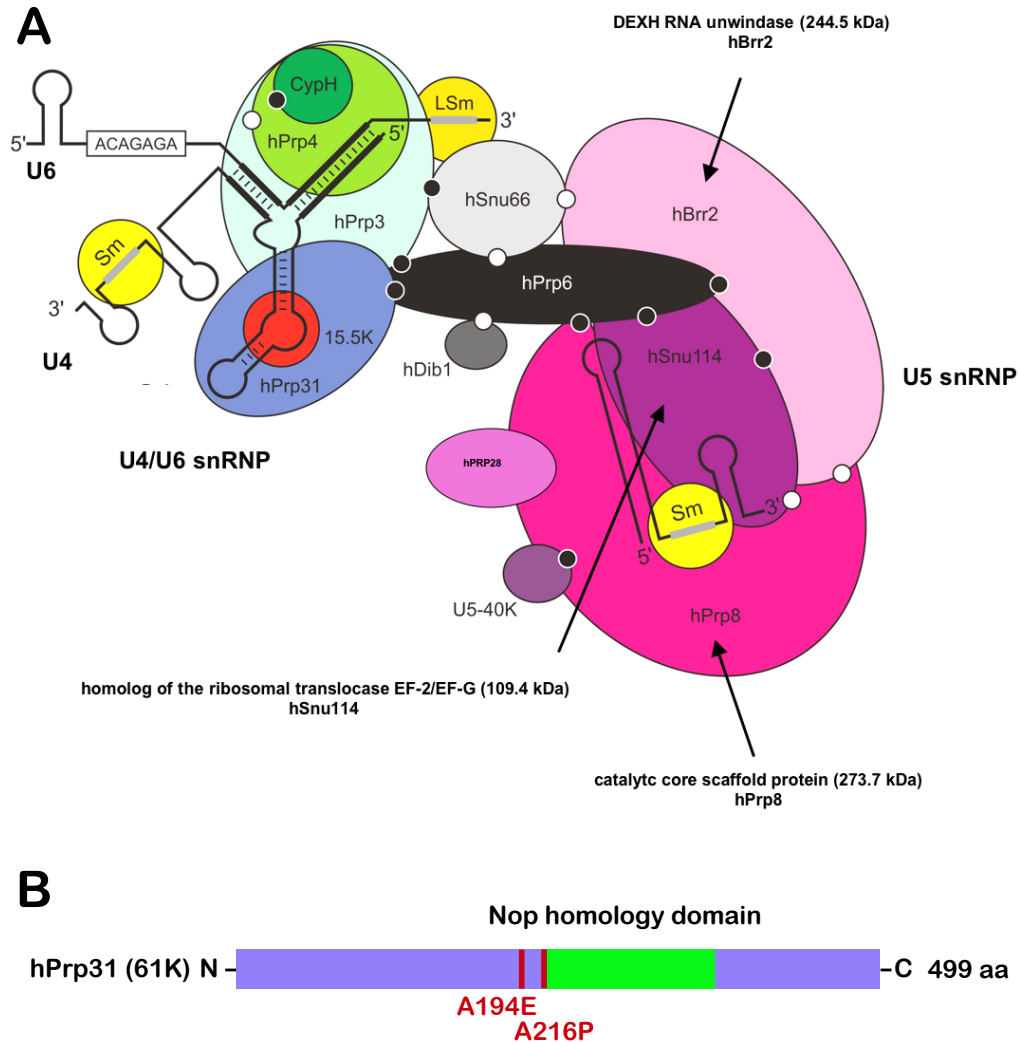
**Figure 1.14:** A. This is a schematic representation of the postulated arrangement of [U4/U6.U5] tri-snRNP protein and RNA components. hPrp31, a U4/U6 specific protein, is indicated in deep blue. It interacts with hPrp6 protein (indicated in black) of the U5 snRNP to form a stabilizing bridge between the U4/U6 di-snRNP and the U5 snRNP during tri-snRNP formation (picture kindly provided by Dr. Berthold Karstner). B. This is a schematic representation of the hPrp31, in which the Nop domain is indicated in green and the two known mutations, A194E and A216P, which cause the autosomal dominant retinitis pigmentosa (adRP) are indicated in red.

independently transcribed as a precursor by RNA polymerase III outside the nucleolus. The first step of 47S pre-rRNA processing in humans involves either the cleavage within 5'-ETS or ITS1, which produce the 21S rRNA, the precursor to the 18S rRNA and the 32S rRNA containing the 5.8S and 28S sections. The subsequent cleavages at ITS2 within the 32S precursor lead to the production of 5.8S and 28S rRNAs. Furthermore, the cleavage products need to undergo a large number of covalent base modifications, including about 100 2' O-methylations ($N_m$) and about 100 pseudouridylations ($\Psi$) (Figure 1.15B), before maturation (111; 112). These modifications are found to be confined to highly conserved core secondary structure of the rRNA and are implicated in ribosome maturation and activity. These modifications occur rapidly on nascent transcripts of pre-rRNAs and are mainly carried out by involving modification guided snoRNAs.

### 1.3.2   snoRNAs and snoRNPs

**snoRNAs**

snoRNAs are abundant and evolutionarily ancient RNAs, which locate to the nucleolus (113). Several hundreds of snoRNAs have been reported in different organisms with various length of 60 to 300 nts. They are mostly encoded within the introns of pre-mRNA genes and belong to the group of modification guided, non-coding RNAs. snoRNAs fall into two major classes containing conserved sequence motifs. One class carries the conserved box C (RUGAUGA) and D (CUGA) motif and are termed boxC/D snoRNAs, while the other class features box H (ANANNA) and box ACA elements and therefore, are named box H/ACA snoRNAs (Figure 1.16). Box C/D snoR-NAs usually also contain related box C'/D' regions. Box C/D snoRNAs are mainly implicated in 2' O-methylation while Box H/ACA is usually responsible for the conversion of uridines to pseudouridines. Both RNAs base pair
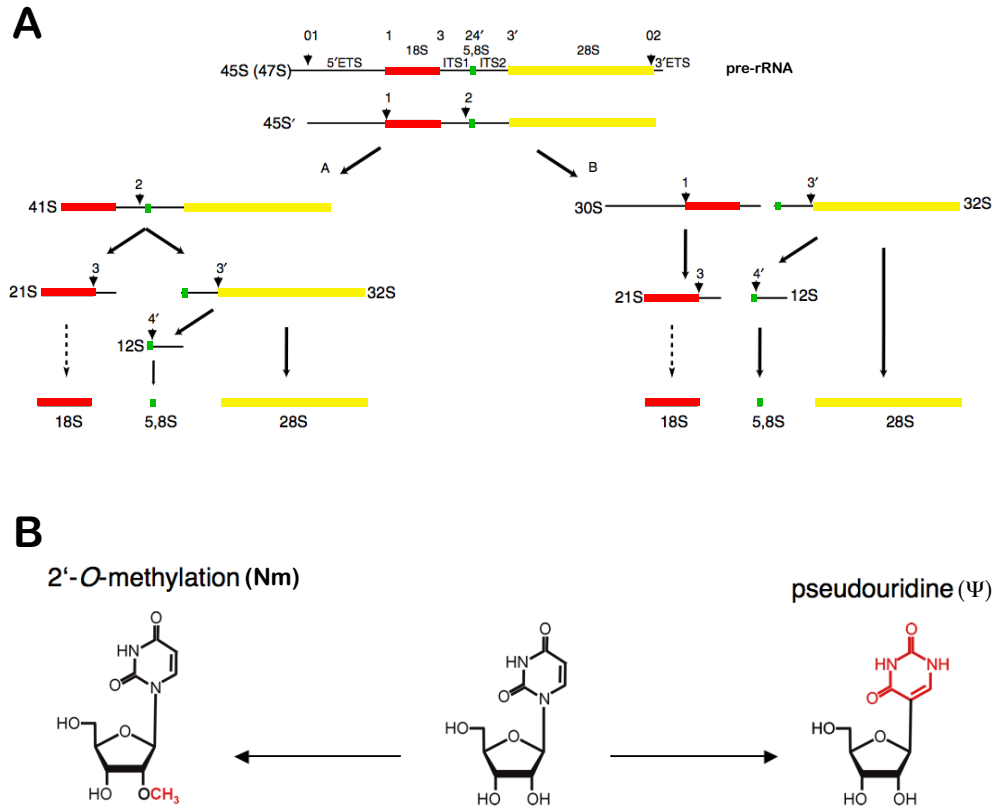
**Figure 1.15:** A. This is a schematic representation of the two alternative pathways for human rRNA processing described by Rouquette *et al*, in which 18S is indicated in red, 5,8S in green and 28S in yellow (picture modified from figure 1 in Rouquette *et al* 2005 (109). B. A uridine is modified to a pseudouridine or is methylated at the 2'-OH position. These are two very common modifications found in nascent cleavage products of the pre-rRNA.

directly with the target rRNAs. In the case of box C/D snoRNAs, a long
(10-21bp) helix with the substrate is formed, whereas box H/ACA snoRNAs
form two short (3-10bp) duplexes with the target sequence. 2' O-methylation
occurs 5 nucleotides upstream of the D or D' box and the pseudouridylation
is located about 15 nucleotides upstream of the H or ACA box. However, a
small number of box H/ACA and box C/D snoRNPs are found to function
in pre-rRNA cleavage rather than modification (114; 115; 116; 117). For
example, U3, U14, U22 box C/D and the snR30 box H/ACA snoRNAs are
essential for generating 18S rRNA (118), whereas the U8 box C/D snoRNA is
needed for the excision of the 5.8S and 28S rRNAs from the pre-rRNA. Apart
from being involved in rRNA processing and modification, snoRNAs are also
found to be involved in snRNA processing in the Cajal bodies. U6 snRNA,
which contains eight 2' O-methylation and three pseudouridines serves as an
example (31; 119).

**snoRNPs**

Like snRNAs, snoRNAs existing as ribonucleoprotein particles called the
snoRNPs. Here, the protein composition of box C/D snoRNPs will be men-
tioned and the U3 as well as U14 snoRNPs will be discussed in more detail
(Figure 1.17). Box C/D snoRNPs contain four core proteins, namely 15.5K
(snu13p in yeast, L7ae in archaea) (57), Nop56/58 (Nop5 in archaea) and
fibrillarin (120; 121; 122; 123; 124). Fibrillarin (Nop1p in yeast) contains the
S-adenosyl-L-methionine (SAM) binding motif and is the predicted methyl-
transferase, which catalyzes the actual methyl transfer reaction (125; 126).
Investigations of U14 snoRNP assembly showed that the 15.5K interacts di-
rectly with U14 box C/D snoRNAs and its binding solely depends on and
absolutely requires the integrity of the box C and D motives (124; 127). Stud-
ies in yeast showed that Nop58p and Nop1p can associate independently to
each other with the snoRNA, whereas Nop56p forms a stable complex with

**BoxC/D snoRNA**          **Box H/ACA snoRNA**
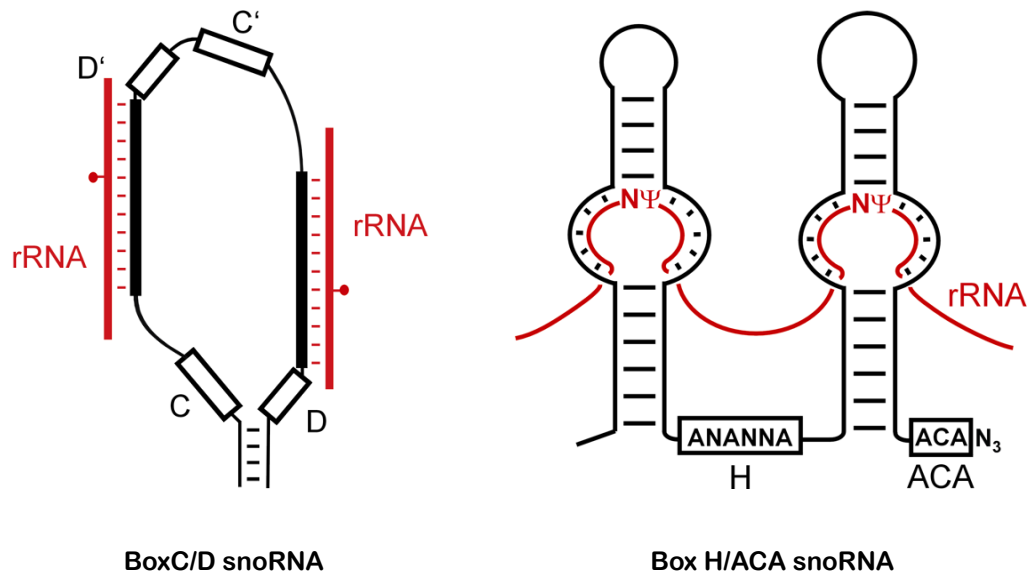
**Figure 1.16:** This is a schematic representation of the secondary structures of box C/D and box H/ACA snoRNAs bound to the pre-rRNA substrate. In box C/D snoRNA representation, 2' O-methylation sites, which locate at 5 nucleotides upsteam of the D or D' box, are indicated with red dots. In case of box H/ACA snoRNA, the pseudouridylation sites, indicated by NΨ, are located about 15 nucleotides upsteam of the H or ACA box. The rRNA substrates are indicated in red (picture kindly provided by Dr. Berthold Kastner).

Nop1p and requires the presence of Nop1p for its binding (97; 123). From UV cross-linking study, it has been shown that Nop58 and fibrillarin cross-linked to the U in box C and the U in box D respectively (128). Moreover, it was also found that the association of snoRNP proteins requires the prior binding of 15.5K to the box C/D motif. Therefore, also in this case a hierarchical assembly pathway similar to that shown in the U4/U6 and U4/U6atac snRNP assembly is followed.

In the most abundant U3 box C/D snoRNA, apart from the box C/D motif (named box C'/D in the case of U3) another sequence motif, the box B/C motif, is also found to directly interact with 15.5K and furthermore with protein U3-55K in vertebrates and Rrp9p in yeast (129; 130). U3-55K contains a WD 40 repeat, which is a protein-protein interaction domain usually folded in a propeller-like structure (130). U3-55K was identified through UV cross-linking studies as directly interacting with the box B/C motif of the U3 snoRNA and once again its binding strictly requires the prior association of the 15.5K protein (131).

### 1.3.3  Comparison of RNA structural requirements for 15.5K associated RNPs

It has been shown that box C and D motif fold to form a K-turn, which has very similar structure like that seen in U4 and U4atac snRNA and which is crucial for the binding of 15.5K (124; 132). The same structural motif is most likely assumed by Box B and C motifs in U3 snoRNA, and its integrity is required for 15.5K binding (131). As a recurring pattern 15.5K binds to K-turns in box C/D snoRNAs to initiate the assembly of the corresponding snoRNPs, therefore the box C/D motif is a pre-requisite for box C/D snoRNP formation. Despite structural similarity of the 15.5K-box C/D
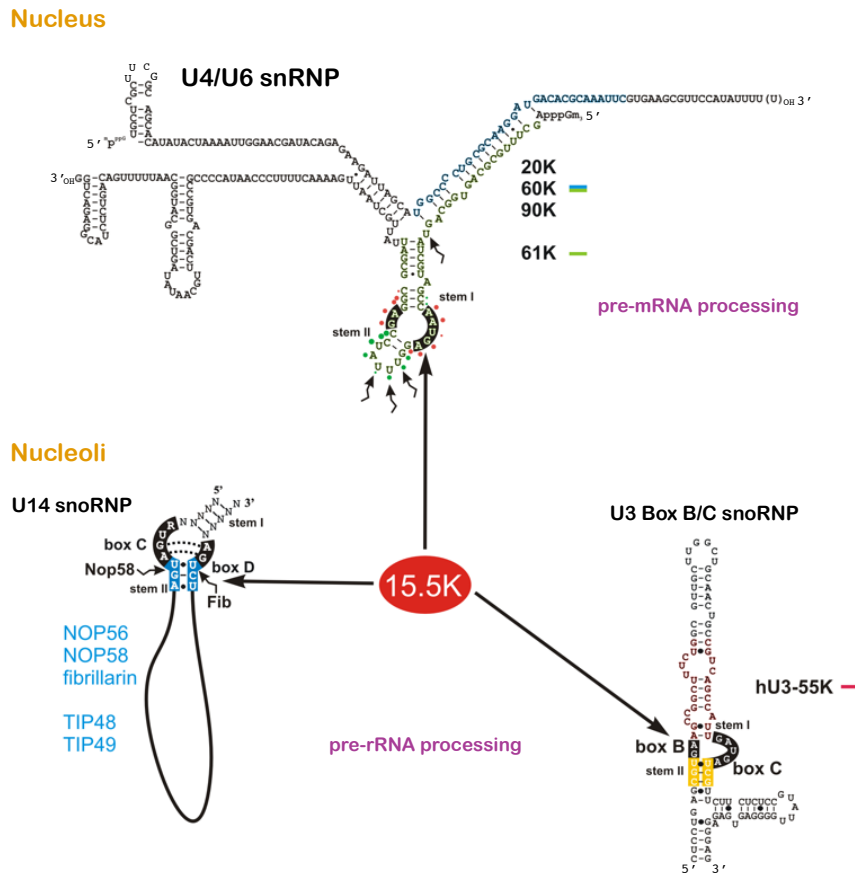
**Figure 1.17:** 15.5K forms an intriguing link between pre-mRNA and pre-rRNA splicing. In the cell nucleus, 15.5K nucleate the assembly of U4/U6 snRNP by the prior binding to U4 5'-SL. The recruited specific proteins include hPrp31 (61K) and the CYPH(20K)/hPrp4(60K)/hPrp3(90K) complex. In the U4/U6 snRNP representation, the red dots indicates the direct interactions between 15.5K and the K-turn region of the U4 5'-SL. The green dots show the contacts from hPrp31 to the U4 5'-SL and the arrows indicate the contacts from hPrp31 using UV-crosslinking. The RNA sequence coloured in blue represents the contacting region found for the CYPH/hPrp4/hPrp3 complex. In nucleoli, 15.5K plays the similar role as the nucleation factor to recruit box C/D snoRNA specific proteins. In the case of U14 snoRNP, these proteins are Nop56, Nop58, fibrillarin, TIP48 and TIP49. Nop56 and Nop58 (Nop5p in archaea) share the homology Nop domain with hPrp31. The cross-links between Nop58 as well as fibrillarin to the snoRNA are shown with arrows. The stem II of the box C/D motif in U14 snoRNA is boxed in blue. Furthermore, 15.5K was also found to enable the binding of hU3-55K to the box B/C motif in U3 box C/D snoRNA (picture made by Dr. Annemarie Schultz).

snoRNA complexes with the 15.5K-snRNA complexes, the 15.5K-box C/D
snoRNA complexes do not bind hPrp31, but recruit the Nop56 and Nop58
(Nop5p in archaea) (133). A great deal of biochemical studies have been car-
ried out in order to understand the structural requirements in the different
RNAs for recruiting the secondary binding partners. In the case of U4 and
U4atac snRNA, it has been shown by hydroxyl radical footprinting method
that hPrp31 forms direct interaction with the snRNA in the penta-loop re-
gion, on stem I (C-stem) as well as on stem II (NC-stem). For U4 snRNA,
although the exact sequence or the buldged-out A of stem I does not seem
to be required for hPrp31 association, complete intact stem I seems to be
required for efficient formation of the ternary complex of hPrp31-15.5K-U4
5'-SL (134). However, this requirement is much less stringent for U4atac
snRNA, where a shortened stem I still allows efficient ternary complex fro-
mation. It was also observed that the integrity, size and sequence of the
penta-loop is not required for establishing a direct interaction with hPrp31
in both U4 and U4atac. More important, however, is the requirement on the
length of stem II. It has been shown in both U4 and U4atac that a length
of 2 base pair in stem II can not be violated and that extension of stem II
by one C-G base pair already abolishes hPrp31 association. The importance
of the stem II is confirmed in the case of 15.5K associated snoRNAs. In
the case of U14 snoRNA, it was found that for efficient association of all
box C/D snoRNP proteins, the sequence of the stem II must not be altered
(127). Stem II of the different RNAs contains crucial identity elements for
secondary protein binding. It is an interesting paradox that in the case of
U4 or U4atac 5'-SLs the length, but not the sequence of stem II is criti-
cal, whereas for the box C/D snoRNAs stem II is naturally longer by one
base-pair and a sequence change is not tolerated. The molecular mechanisms
employed by the secondary binding proteins for measuring the length of stem
II are unknown. This question is clarified in this work. Apart from prob-

ing these requirements with mutated RNAs, surface mutants of 15.5K have been generated in an attempt to map the interaction surface on the 15.5K with the secondary binding partners (Figure 1.18) (135). Distinct regions of 15.5K have been shown to be specifically involved in the recruitment of the secondary binding proteins to the snRNPs/snoRNPs. In one of these surface mutants of 15.5K namely 15.5K-2, a stretch of 4 amino acids located on the $\alpha 3$ of 15.5K was changed to the corresponding amino acids in its ribosomal homolog namely L7a (E74R/D75K/K76M/N77T) (135). This mutant was shown to abolish the binding of hPrp31 in pull down assays.However, detailed structural information is lacking to confirm and define direct protein-protein interactions between 15.5K and the other proteins at the amino acid level. This question is also addressed in this thesis.

## 1.4 Aim and outline of this work

There are several interesting questions concerning 15.5K associated complexes, which are answered in this work:

1. Why is the prior binding of 15.5K an absolute requirement for all its associated complexes? In other words, what is the origin of the hierarchical assembly pathway? Structurally, it is not clear whether 15.5K only pre-structures the RNA into the favoured conformation or it also provides part of the interacting surface for the secondary proteins.

2. As hPrp31, Nop56 and Nop58 all require the preformed complexes of 15.5K-K-turns for their incorporation, they need to be able to distinguish between these very similar complexes. hPrp31, Nop56 and Nop58 are homologous proteins containing the Nop domain. These two points together lead to the interesting question: where does the specificity in the assemblies of these complexes come about? From previous stud-

| Mutant | Location of mutation |
|--------|----------------------|
| 15.5K-1 | Q27R/Q28K/C30K/N31E/ Y32S/K33G |
| 15.5K-2 | E74R/D75K/K76M/N77T |
| 15.5K-3 | A88D/R91E |
| 15.5K-4 | Q115A/Q117A/Q121A |
| 15.5K-5 | E124S/R125S |
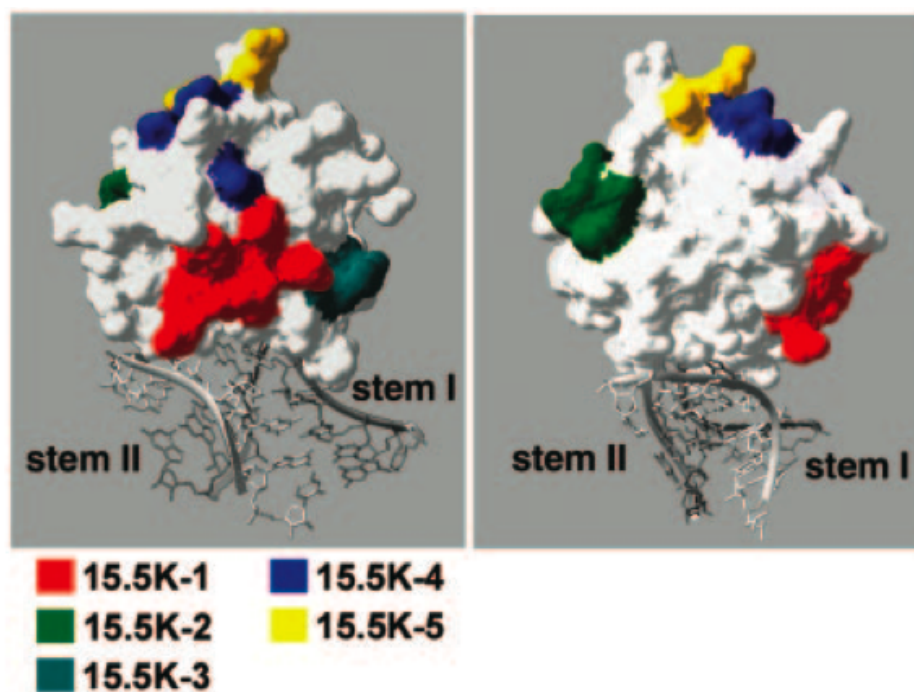


**Figure 1.18:** Surface mutants of 15.5K created by Dr. Annemarie Schultz (135). 15.5K-2, indicated in green, was found to strongly reduce the binding to hPrp31. Mutants 15.5K-2, 3, 4 and 5 were found to destabilize the complex formation with CYPH/hPrp4/hPrp3 in the presence of U4/U6 snoRNA duplex. These surface mutants were also found to have implication on snoRNP formation (picture taken from Figure.2 in (135).

ies, the role of the stem II of these RNAs in complex assembly has been investigated and it seems to make important contributions to the specificity. The length and not the sequence of the stem II of U4 or U4$_{atac}$ snRNA seems to play a crucial role in selecting hPrp31 as the secondary binding partner, whereas the sequence of the stem II in box C/D snoRNA cannot be varied. However, the molecular mechanism for these findings is not clear.

3. Detailed binding properties of Nop domain is also not revealed. It is not known whether it is only implicated in RNA binding or if it also contacts 15.5K directly. The hPrp31-15.5K-U4 5'-SL system is of particular interest, as in depth information about the RNP architecture of the U4/U6 interaction region is required to understand the mechanism of spliceosome activation and it is conceivable that protein interactions at the adjacent functionally important U4 5'-SL also contribute to this critical step of pre-mRNA splicing.

4. Does 15.5K adopt identical 3D structures in the snRNP and the snoRNPs or does it also provides selectivity through structural changes?

In order to tackle these questions, advanced NMR spectroscopy methods in combination with information from X-crystallography, biochemical studies and novel computational methods are used in this work. At the beginning of the work, backbone assignment of 15.5K protein was carried out as the basis for all further structural investigations. The main part of the work was carried out in order to answer the questions 1-3 listed above. With help of homology modeling and the crystal structure of Nop5 in archaea by Aittaleb *et al* (100), a 3D model of a hPrp31 fragment (amino acid 189-331) was generated. Using the UV cross-linking data confirmed by mass spectrometry in Dr. Henning Urlaub's group (136) and the novel docking method HADDOCK2.0, a 3D model of the hPrp31$^{189-331}$-15.5K-U4 5'-SL

complex was created. The crystal structure of the hPrp31$^{78-333}$-15.5K-U4 5'-SL complex was solved from the group of Dr. Markus Wahl. Interaction surface of hPrp31$^{78-333}$ is compared with that of the full length hPrp31 and the generated 3D model was compared to the corresponding part of the crystal structure. Furthermore, biochemical and NMR investigations were carried out on the mutant 15.5K-2. Single and double amino acid mutations of 15.5K at the same position were created and band shift assays of these mutants were carried out in order to pinpoint the responsible amino acid for the dysfunction of 15.5K-2. RDC refinement was performed on the 15.5K-2 mutant in an attempt to find out the structural difference between 15.5K-2 and the wild type protein. To address the question number 4, HSQC spectra of 15.5K bound to constructs of U14 and U3 box C/D snoRNAs were measured in order to observe changes in 15.5K upon association with these different RNAs.

# Chapter 2

# Materials and Methods

## 2.1 Materials

### 2.1.1 Chemicals and Media

Chemicals and ready-made media used in this work are listed in Table 2.1

**Table 2.1:** Chemicals and media

| Chemical/Media | Company |
|---|---|
| Agar, IPTG | AppliChem, Darmstadt, Germany |
| Pf1 NMR co-solvent | ASLA$^{biotech}$, Riga, Latvia |
| LB-Agar, LB-liquid media | BIO 101, USA |
| Chloramphenicol, tRNA *E. coli* | Boehringer, Mannheim |
| ammonium chloride (>98 % $^{15}$N) | Cambridge Isotope Laboratories, Andover, USA |
| $^{13}$C$_6$-D-glucose (>98 % $^{13}$C) | Spectra Stable Isotopes, Columbia, USA |

| Chemical/Media | Company |
| --- | --- |
| 99.9 % $D_2O$, 98 % $D_8$-glycerol | Eurisotop, Gif-sur-Yvette, Frankreich |
| ammonium molybdate tetrahydrate, ascorbic acid, Coomassie Brillant Blue R-250, copper chloride dihydrate, Xylene cyano iron (II) sulfate heptahydrate | Fluka, Neu-Ulm, Germany |
| DTT | Gerbu, Gaiberg, Germany |
| agarose, kanamycin sulphate | GibcoBRL, Karlsruhe, Germany |
| Agarose (low melting point), Bench Mark protein ladder | Invitrogen, Karlsruhe, Germany |
| acetic acid, $\alpha$-D(+)-glucose monohydrate, ammonium acetate, ammonium chloride, ammonium hydrocarbonate, Ammoniumperoxodisulfate (APS), boric acid, disodium hydrogen phosphate, Bromphenol blue ethanol, glycerol, glycine, hydrochloric acid, manganese chloride dihydrate, magnesium sulfate heptahydrate, sodium chloride, sodium dihydrogen phosphate, sodium hydroxide, TFA, thiaminechloride hydrochloride, Tris, urea | Merck, Darmstadt, Germany |
| BSA | New England Biolabs, Ipswich, USA |
| Ni-NTA Agarose, QIAGEN Plasmid Midi Kit, QIAprep Spin Miniprep Kit, QIAquick Gel Extraction Kit, QIAquick PCR Purification Kit | Qiagen, Hilden, Germany |
| Bradford dye | Bio-Rad, Munich |
| Cobalt chloride hexahydrate | Riedel-de Haën, Seelze, Germany |
| Complete$^{TM}$ protease inhibitors, Complete$^{TM}$ protease inhibitors EDTA-free, elastase, protease K, trypsin | Roche Diagnostics, Mannheim, Germany |
| ampicillin sodium salt, APS, dipotassium hydrogen phosphate, DTT, ethidium bromide, EDTA, HEPES, imidazole, magnesium chloride hexahydrate, potassium dihydrogen phosphate, Rotiphorese Gel 30, Rotiphorese Gel 40, sodium acetate, Tris, tryptone, yeast extract-B | Roth, Karlsruhe, Germany |
| SDS | Serva, Heidelberg, Germany |
| ATP, carbenicillin, chymotrypsin, Ipegal CA 630 (NP-40), N-Z-amine AS, Triton X-100, TEMED, Reduced glutathione | Sigma, Deisenhofen, Germany |

| Chemical/Media | Company |
| --- | --- |
| 99.9 % $D_2O$, 98 % $D_8$-glycerol | Eurisotop, Gif-sur-Yvette, Frankreich |
| dNTP | Stratagene, La Jolla, USA |

## 2.1.2   Enzymes and inhibitors

Enzymes and inhibitors used in this thesis can be found in Table 2.2.

**Table 2.2:** Enzymes and inhibitors

| Enzymes/Inhibitors | Company |
| --- | --- |
| *DPn*, T4 polynucleotide kinase | New England Biolabs, Ipswich, USA |
| Thrombin (1U/$\mu$L) | GE Healthcare life Sciences, Germany |
| RQ Dnase I (1U/$\mu$L) | Promega, USA |
| DNase I (1U/$\mu$L) | Roche Diagnostics, Mannheim, Germany |
| PEFABLOC SC (4-(2-Aminoethyl)benzenesulfonyfluoride.HCl) | Roth, Karlsruhe, Germany |
| Protease-inhibitor-cocktail Complete$^{TM}$ (EDTA-free) | Roche Diagnostics, Mannheim, Germany |
| PreScission protease | GE Healthcare life Sciences, Germany |
| Turbo *Pfu* DNA polymerase, dNTP | Stratagene, La Jolla, USA |

## 2.1.3   Bacterial strains and plasmids

### Bacterial strains

*E.coli* strains used in this work are listed in Table 2.3.

**Table 2.3:** Bacterial strains used in this work

| Strain | Company |
|--------|---------|
| BL21(DE3) pLysS | Novagen, USA |
| BL21(DE3)-RIl | Stratagene, USA |
| Strain HMS174 (DE3) | Novagen, USA |
| Strain Turner (DE3)pLacI | Novagen, USA |

## Plasmids

Plasmids used in this work are listed in Table 2.4.

**Table 2.4:** Plasmids used in this work

| Plasmid | Reference/Origin | Description |
|---------|------------------|-------------|
| pGEX-15.5K | Dr. Stephanie Nottrott [85] | Expression vector GST-15.5K; $Amp^R$ |
| pGEX-15.5K-2 | Dr. Annemarie Schultz [135] | Expression vector GST-15.5K-2 mutant; $Amp^R$ |
| pGEX-15.5K-E74R | this work | Expression vector GST-15.5K-E74R mutant; $Amp^R$ |
| pGEX-15.5K-D75K | this work | Expression vector GST-15.5K-D75K mutant; $Amp^R$ |
| pGEX-15.5K-K76M | this work | Expression vector GST-15.5K-K76M mutant; $Amp^R$ |
| pGEX-15.5K-E74R/D75K | this work | Expression vector GST-15.5K-E74R/D75K mutant; $Amp^R$ |
| pTriEx-hPrp31 | Dr. Sunbin Liu, MPIBPC Göttingen [137] | Expression vector hPrp31-His; Amp, $Cam^R$ |
| pGEX-6P-hPrp31$^{78-333}$ | Dr. Sunbin Liu, MPIBPC Göttingen [137] | Expression vector GST-hPrp31$^{78-333}$; Amp, $Car^R$ |
| pETM-40-hPrp31 | Dr. Sunbin Liu, MPIBPC Göttingen [137] | Expression vector MBP-hPrp31; $Kan^R$ |

## 2.1.4 Oligonucleotides

### RNA oligonucleotides

All RNA oligonucleotides used for NMR and electrophoresis gel mobility shift assay were ordered from IBA (Göttingen, Germany) and are presented in Table 2.6.

**Table 2.6:** The name and the RNA sequences

| Name | Sequence |
| --- | --- |
| U4 5'-SL-24nt | 5'-GGCCAAUGAGGUUUAUCCGAGGCC-3' |
| U4 5'-SL-33nt | 5'-GCUUUGCGCAGUGGCAGUAUCGUA GCCAAUGAGGUUUAUCCGAGGCGCGAU-3' |
| U14CD-26nt | 5'-UCGCUGUGAUGAUUUAUUCUGAGCGA-3' |
| U3CD-35nt duplex | 5'-CCACGAGGAAGAGAGGUAGCG-3' 5'-CGCGGUCUGAGUGG -3' |
| U3BC-62nt | 5'-AGCGUGAAGCCGGCUUUCUGGCGUUGCUUGGCU GCAACUGCCGUCAGCCAUUGAUGAUCGUU-3' |

**DNA oligonucleotides**

All DNA oligonucleotides (primers) used in mutagenesis of 15.5K mutants were ordered from IBA (Göttingen, Germany) and are presented in Table 2.7.

**Table 2.7:** The name and the DNA sequences

| Name | Sequence |
|------|----------|
| 15.5K-E74R | 5'-CTGCACCTGCCGCTGCTGTGT**AG**AGACAAGAATGTGCCCTACGTGTTTGTG-3' |
| | 5'-CACAAACACGTAGGGCACATTCTTGTCT**CT**ACACAGCAGCGGCAGGTGCAG-3' |
| 15.5K-D75K | 5'-CTGCACCTGCCGCTGCTGTGTGAA**AAG**AAGAATGTGCCCTACGTGTTTGTG-3' |
| | 5'-CACAAACACGTAGGGCACATTCTT**CTT**TTCACACAGCAGCGGCAGGTGCAG-3' |
| 15.5K-K76M | 5'-CCTGCCGCTGCTGTGTGAAGACA**T**GAATGTGCCCTACG -3' |
| | 5'-GCACAAACACGTAGGGCACATTC**A**TGTCTTCACACAGCAGC-3' |
| 15.5K-E74R/D75K | 5'-CTGCACCTGCCGCTGCTGTGT**AG**AA**AG**AAGAATGTGCCCTACGTGTTTGTG-3' |
| | 5'-CACAAACACGTAGGGCACATTCTT**CT**T**CT**ACACAGCAGCGGCAGGTGCAG-3' |

## 2.1.5   Culture media and stock solutions

All culture media, except the minimal medium for 100% deuteration, were prepared according to Table 2.8 followed by autoclave sterilisation. Before use, the appropriate antibiotics in the desired concentrations were added to every medium. For agar plates 15 g agar per 1 L medium were added before autoclaving.

**Table 2.8:** Culture media and antibiotics

| Name | Amount | Chemical | Remark |
|---|---|---|---|
| 50x5052 | 25% | glycerol | |
| | 2.5% | glucose | |
| | 25% | $\alpha$-lactose monohydrate | |
| | up to 1 L | $H_2O$ | (138) |
| 50xM | 1.25M | $Na_2HPO_4$ | |
| | 1.25M | $K_2HPO_4$ | |
| | 2.5 M | $NH_4Cl$ | |
| | 0.25 M | $Na_2SO_4$ | |
| | up to 1 L | $H_2O$ | (138) |
| LB-medium | 10 g | tryptone | |
| | 5 g | yeast extract | |
| | 10 g | NaCl | |
| | up to 1 L | $H_2O$ | |
| M9-minimal medium | 6.8 g | $Na_2HPO_4$ | |
| | 3 g | $KH_2PO_4$ | |
| | 0.5 g | NaCl | |
| | 1 g | $NH_4Cl$ or $^{15}NH_4Cl$ | |
| | 4 g | glucose, $^{13}C_6$-glucose | |
| | | or $D_8$-glycerol | |
| | 2 mL | 1 M $MgSO_4$ | |
| | 50 $\mu$L | 2 M $CaCl_2$ | |
| | 0.03 g | thiaminechloride.HCl | |
| | 10 mL | trace elements | |
| | up to 1 L | $H_2O$ or 99.9 % $D_2O$ | |
| trace elements | 0.6 g | $FeSO_4\cdot$ 7 $H_2O$ | |
| | 0.094 g | $MnCl_2\cdot$ 2 $H_2O$ | |
| | 0.08 g | $CoCl_2\cdot$ 6 $H_2O$ | |
| | 0.07 g | $ZnSO_4\cdot$ 7 $H_2O$ | |
| | 0.03 g | $CuCl_2\cdot$ 2 $H_2O$ | |
| | 0.002 g | $H_3BO_3$ | |
| | 0.025 g | $(NH_4)_6Mo_7O_{24}\cdot$ 4 $H_2O$ | |
| | up to 100 mL | $H_2O$ | stir 10 min |
| | 0.5 g | EDTA | stir over night |
| 1000x trace metals | 50 mM | $FeCl_3$ | |
| | 20 mM | $CaCl_2$ | |
| | 10 mM | $MnCl_2$ | |
| | 10 mM | $ZnSO_4$ | |
| | 2 mM | $CuCl_2$ | |
| | 2 mM | $CoCl_2$ | |
| | 2 mM | $NiCl_2$ | |
| | 2 mM | $NaMoO_4$ | |
| | 2 mM | $Na_2SeO_3$ | |
| | 2 mM | $H_3BO_3$ | |
| | dissolve in 60mM HCl | | |
| ZY-medium | 10 g | N-Z-amine AS | |
| | 5 g | yeast extract-B | |
| | up to 1 L | $H_2O$ | |
| ampicillin | 100 mg/mL | ampicillin sodium salt | stored at -20 °C |
| chloramphenicol | 30 mg/mL | chloramphenicol | in abs. EtOH, and stored at -20 °C |
| carbenicillin | 50 mg/mL | carbenicillin | stored at -20 °C |
| kanamycin | 34 mg/mL | kanamycin | stored at -20 °C |
| Lysozyme | 100 mg/mL | lysozyme | stored at -20 °C |
| IPTG | 1 M | IPTG | steril filtrated, stored at -20 °C |

## 2.1.6 Buffers and solutions

The buffers and solutions used for protein purification and NMR experiments in this work are listed in Table 2.9.

**Table 2.9:** Buffers and solutions

| Method | Name | Amount | Chemical |
|---|---|---|---|
| **Purification of** | 1 x PBS pH 7.4 | 140 mM<br>2.7 mM<br>10 mM<br>1.8 mM | NaCl<br>NaCl<br>$Na_2HPO_4$<br>$K_2HPO_4$ |
| **GST-tagged protein** | elution buffer 1 | 20 mM<br>500 mM<br>10 mM | Tris/HCl pH 7.9<br>NaCl<br>reduced gluthatione |
| **Purification of** | Lysis bufer | 20 mM<br>150 mM<br>0.2 %(w/v)<br>1 mM | Tris/HCl pH 7.4<br>NaCl<br>IGEPAL (NP40)<br>DTT |
| **MBP-tagged protein** | wash buffer | 20 mM<br>300 mM<br>1 mM | Tris/HCl pH 7.4<br>NaCl<br>DTT |
| | elution buffer | 20 mM<br>150 mM<br>10 mM<br>1 mM | Tris/HCl pH 7.4<br>NaCl<br>maltose<br>DTT |
| **Purification of** | Lysis bufer | 20 mM<br>150 mM<br>0.2 %(w/v)<br>10 mM | Tris/HCl pH 8.0<br>NaCl<br>IGEPAL (NP40)<br>imidazol |
| | wash buffer 1 | 20 mM<br>1 M<br>10 mM | Tris/HCl pH 8.0<br>NaCl<br>imidazol |
| **His-tagged protein** | wash buffer 2 | 20 mM<br>150 mM<br>10 mM | Tris/HCl pH 8.0<br>NaCl<br>imidazol |
| | wash buffer 3 | 20 mM<br>150 mM<br>50 mM | Tris/HCl pH 8.0<br>NaCl<br>imidazol |
| | elution buffer | 20 mM<br>150 mM<br>250 mM | Tris/HCl pH 8.0<br>NaCl<br>imidazol |
| **SDS-PAGE** | separating gel buffer (4x) | 1.6 M<br>0.4% (w/v) | Tris/HCl pH8.8<br>SDS |
| | stacking gel buffer (4x) | 500 mM<br>0.4%(w/v) | Tris/HCl pH6.8<br>SDS |
| | separating gel solution (10%) | 1x<br>10 mL<br><br>100 $\mu$L<br>100 $\mu$L<br>up to 30mL | separating gel buffer<br>rotiphorese gel 30 (30% acry-<br>lamide : 0.8% bisacrylamide)<br>TEMED<br>10%(w/v) APS<br>$H_2O$ |

| Method | Name | Amount | Chemical |
|---|---|---|---|
| **SDS-PAGE** | stacking gel solution (5%) | 1x | stacking gel buffer |
| | | 1.66 mL | rotiphorese gel 30 (30% acrylamide : 0.8% bisacrylamide) |
| | | 25 $\mu$L | TEMED |
| | | 50 $\mu$L | 10%(w/v) APS |
| | | up to 10mL | $H_2O$ |
| | destaining solution | 100 mL | acetic acid |
| | | 900 mL | $H_2O$ |
| | 4 x protein loading buffer | 1.7 g | SDS |
| | | 7.5 mL, 1 M | Tris/HCl pH 6.8 |
| | | 23 mL | glycerol |
| | | 50 mg | bromphenol blue |
| | | 0.5 mL | $\beta$-mercaptoethanol |
| | | up to 50 mL | $H_2O$ |
| | running buffer | 1 g | SDS |
| | | 3.03 g | Tris |
| | | 14.4 g | glycine |
| | | up to 1 L | $H_2O$ |
| | staining solution | 2.2 g | Coomassie Brilliant blue G250 |
| | | 100 mL | acetic acid |
| | | 250 mL | isopropanol |
| | | 650 mL | $H_2O$ |
| **native RNA gel** | 10 x TBE buffer | 108 g | Tris |
| | | 55 g | boric acid |
| | | 40 mL, 0.5 M | EDTA pH 8.0 |
| | | up to 1 L | $H_2O$ |
| | gel solution (6%) | 10mL | acrylamide gel A (30%) |
| | | 1.875mL | bisacrylymide (2%) |
| | | 2.5mL | 10 x TBE |
| | | 250 $\mu$L | 25% (w/v) APS |
| | | 25 $\mu$L | TEMED |
| | | up to 50mL | $H_2O$ |
| **Gel filtration** | gel filtration buffer | 20 mM | HEPES pH 7.6 |
| | | 120 mM | NaCl |
| | | 1mM | DTT |
| **NMR** | sample buffer | 20 mM | HEPES pH 7.6 |
| | | 120 mM | NaCl |
| | | 1mM | DTT |
| **PCR** | Cloned Pfu buffer | 20 mM | Tris/HCl pH 8.8 |
| | | 2 mM | $MgSO_4$ |
| | | 10 mM | KCl |
| | | 10 mM | $(NH_4)_2SO_4$ |
| | | 0.1 % | Triton X-100 |
| | | 0.1 mg/mL | BSA |

### 2.1.7 Kits

Commercially available kits used in this work are listed in Table 2.10.

**Table 2.10:** Bacterial strains used in this work

| Strain | Company |
|---|---|
| Bradford assay | Bio-Rad, Munich |
| QIAgen Plasmid Mini/Maxi Preparation Kit | QIAGEN, Hilden |
| QuickChange$^{TM}$ Site-Directed Mutagenesis Kit | Stratagene, USA |

### 2.1.8 Working and chromatography materials

Working and chromatography materials used in this work can be found in Table 2.11.

**Table 2.11:** Working and chromatography materials

| Materials | Company |
|---|---|
| Amicon Ultra centrifugal filter devices 10,000 MWCO and 30,000 MWCO | Millipore, USA |
| Amylose resin | New England Biolabs, Ipswich,USA |
| Baffled flask 250 mL, 3000mL | Schott Duran, Germany |
| BioMax MR film 35x43 cm | Kodak, USA |
| Dialyses membranes MWCO 6000-8000 Da | SpektraPor, USA |
| Econo-Pac column 20mL | Bio-Rad, Munich |
| GSTrap$^{TM}$ FF 1mL and 5mL | GE Healthcare life Sciences, Germany |
| Glutathione sepharose | GE Healthcare life Sciences, Germany |
| Hi-load 26/60 Superdex 200 pg | GE Healthcare life Sciences, Germany |
| Hi-load 26/60 Superdex 75 pg | GE Healthcare life Sciences, Germany |
| Ni-NTA Agarose | QIAGEN, Hilden |
| Pipettes | Gilson Medical Electronics, Eppendorf, Germany |
| Small Tube-O-Dialyzer | Millipore, USA |

| Materials | Company |
|---|---|
| Shigemi NMR tube 5 mm | Shigemi Corp., Japan |
| Slide-A-Lyzer Dialysis Cassettes MWCO 3500 (0.1-0.5 mL) | Pierce , USA |
| Sterile filters 0.2 m or 0.45 m | Millipore, USA |
| Whatman 3MM Paper | Whatman Paper, UK |

## 2.1.9   Equipment

Laboratory instruments and consumables used in this work are tabulated in Table 2.12.

**Table 2.12:** Equipment and supplier

| Common Name | Identifier/Company |
|---|---|
| **Balances** | Sartorius BP211D, Sartorius, Göttingen, Germany<br>Sartorius BP410S, Sartorius, Göttingen, Germany<br>Sartorius BP4100, Sartorius, Göttingen, Germany |
| **Centrifuges** | Sorvall rotors: JLA-6000, SS 34, Kendro, USA<br>Eppendorf Centrifuge 5415D, Wesseling-Berzdorf, Germany<br>Eppendorf Centrifuge 5804, Wesseling-Berzdorf, Germany<br>Heraeus Centrifuge Biofuge fresco, Kendro, Hanau, German<br>Heraeus Centrifuge Evolution RC, Kendro, Hanau, Germany<br>Heraeus Centrifuge Megafuge 1.0R, Kendro, Hanau, Germany |
| **Electrophoresis** | Kodak Electrophoresis documentation and analysis system 120, Eastman Kodak Co., New York, NY, USA<br>Power Pac 300, BioRad, München, Germany<br>Polyacrylamide gel electrophoresis: Mini-PROTEAN 3 Cell, BioRad, München, Germany<br>Agarose gel electrophoresis: Mini-Sub Cell GT, BioRad, München, Germany |
| **-80 °C freezer** | MDF-U71V Ultra-low temperature freezer, SANYO Electric Co., Ltd, Osaka, Japan |
| **Filtering** | sterile filter 0,20 $\mu$m, Sartorius, Göttingen, Germany |
| **Gel developer** | X-OMAT 2000 processor, Kodak, USA |
| **Gel dryer model 583** | Bio-Rad, Munich |
| **HPLC** | Äkta prime, GE Healthcare life Sciences, Germany |
| **Incubator** | Infors Multitron HT, Einsbach, Germany<br>Certomat R, B. Braun Biotech International, Melsungen, Germany |
| **Lyophylisation** | Christ Alpha 2-4, B. Braun Biotech International, Melsungen, Germany |
| **NMR** | AVANCE 400, Bruker, Karlsruhe, Germany<br>AVANCE 600, Bruker, Karlsruhe, Germany<br>DRX 600, Bruker, Karlsruhe, Germany<br>AVANCE 700, Bruker, Karlsruhe, Germany<br>DRX 800, Bruker, Karlsruhe, Germany |

| Common Name | Identifier/Company |
|---|---|
| | AVANCE 900, Bruker, Karlsruhe, Germany |
| Milli-Q-water | Milli-Q-water supply apparatus, Millipore, USA |
| pH-Meter | PB11 PY-P10, Sartorius, Göttingen, Germany |
| Spectroscopy | UV/VIS-Spectrophotometer, Hewlett-Packard 8453, Böblingen, Germany<br>CD spectropolarimeter J-720, JASCO International, Groß-Umstadt, Germany<br>Mass spectrometer, Water Micromass ZQ single quadrupole, Waters, Saint-Quentin, France |
| Sonifier | Digital sonifier, Branson, USA |
| Scintillation | Scintillation analyzer LS 1701/TRI-CARB 2100TR, Bechman/Packard, USA |

## 2.1.10   Software packages

A list of the software packages and web servers employed for processing and analysing NMR experiments, calculating electrostatic potential surfaces, structural calculation and for viewing 3D structures is given in Table **??**.

**Table 2.13:** Software packages

| PROGRAM | REFERENCE/ORIGIN |
|---|---|
| UCSF Chimera | http://www.rbvi.ucsf.edu (139) |
| Clustal X (v.1.83) | http://www.ebioinformatics.org (140) |
| FELIX-ND (2000.1) | Accelrys, San Diego, USA |
| HADDOCK2.0 | http://www.nmr.chem.uu.nl/haddock/ (141) |
| InsightII | MSI 2000 release, San Diego, CA, USA |
| MOLMOL | (142) |
| NCBI-BLAST | http://www.ncbi.nlm.nih.gov |
| PALES | (143) |
| PDB2PQR | http://pdb2pqr.sourceforge.net (144) |
| PredictProtein | http://www.predictprotein.org (145) |
| ProFit (v.2.5) | http://www.bioinf.org.uk |
| PyMOL | http://www.pymol.org |
| Sparky 3 | T. D. Goddard and D. G. Kneller, University of California, San Francisco |
| SWISS-MODEL | http://swissmodel.expasy.org (146; 147) |
| Topspin 2.0 | Bruker, Karlsruhe, Germany |
| XPLOR-NIH 2.9.7 | (148) |
| X-WINNMR 3.5 | Bruker, Karlsruhe, Germany |

## 2.2 Methods

### 2.2.1 Molecular biology methods

**Concentration determination of nucleic acids**

To determine the concentration of nucleic acids, the absorbance was measured at a wavelength of 260 nm in comparison to the corresponding buffer without nucleic acids. The concentration was calculated from the following equations:

- 1 $OD_{260}$ = 50 ug/mL double stranded DNA = 0.15 mM (in nucleotides)

- 1 $OD_{260}$ = 33 ug/mL double stranded DNA = 0.10 mM (in nucleotides)

- 1 $OD_{260}$ = 40 ug/mL double stranded DNA = 0.11 mM (in nucleotides)

- 1 OD260 = 45 ug/mL double stranded DNA = 0.12 mM (in nucleotides)

**Site-directed mutagenesis**

Single and double amino acid changes were generated using the QuikChange site-directed mutagenesis kit.

Reaction mix:

- 20 pmol of each primer

- 20-25 ng of plasmid DNA template

- 6.25 $\mu$mol of dNTPs

- 1.5 units of Turbo Pfu DNA polymerase

- 3.5 $\mu$L 10x Cloned Pfu buffer

  add to 35 $\mu$L with $H_2O$.

PCR steps:

- 1x 95 °C for 30 s (initial denaturation)

- 12x 95 °C for 30s (denaturation)

- 53 °C for 1min (annealing)

- 68 °C for 11min (elongation)

Elongation time was estimated using suggested 2 min per kb of plasmid length ($\sim$5.3 kb) by QuickChange protocol. For creating single nucleotide changes and for creating double nucleotide changes the number of cycles was 12 and 18, respectively. Following temperature cycling, 1.5 units of *Dpn* I restriction enzyme were directly added to each PCR reaction mixture and incubated for 1.5 h at 37 °C. 5 $\mu$L of the reaction mix was transformed into E. coli XL1-blue. The plasmid DNA is obtained by the miniprep procedure using Miniprep kit.

### DNA sequencing

Sequencing of purified plasmid DNA was performed via the extended Hot Shot DNA sequencing service of Seqlab (Goettingen, Germany). 200 $\mu$L PCR tubes with flat lids were loaded with a total volume of 7 $\mu$L containing 20 pmol of primer and 0.6-0.7 $\mu$g plasmid DNA in $H_2O$.

**Radioactive 5'-end labeling of RNA oligonucleotides**

For band shift assays RNA oligonucleotides were labeled at the 5'-end using r$^{32}$P ATP (5000Ci/mmol) through 5'- end phosphorylation by T4-polynucleotide kinase.

Reaction mix:

- 1$\mu$L RNA oligonucleotide (10 pmol/$\mu$L)

- 1$\mu$L 10x PNK buffer

- 5$\mu$L r$^{32}$P ATP 5000 (Ci/mmol)

- 1$\mu$L T4-polynucleotide kinase

add to 10$\mu$L with ddH2O.

After incubating for 1 hr at 37 °C, 40$\mu$L of CE buffer was added to the reaction mix. G-25 micro-centrifugation column was used for purification of the product. 1$\mu$L of the eluate was taken for determining the radioactivity in the scintillation counter.

**Native polyacrylamide gel electrophoresis**

In order to test protein-RNA binding interaction using band shift assay, the condition is chosen so that protein and RNA form complexes in their native state. Therefore native rather than denatured polyacrylamide gels were used. In this work 6%-9% polyacrylamide gels (80:1 acrylamide : bisacrylamide) were used. The gel was left to 'pre-run' without loading of the samples to ensure homogeneity of the gel and to achieve a better migration profile. The samples were mixed with glycerol. The gel electrophoresis procedure

was carried out at 4 °C in 0.5 x TBE buffer. Protein-RNA complexes were separated at 9 W for 1.5 hrs.

**Band shift assay**

30-50 fmol of radioactively 5'- end labeled RNA oligos was used in band shift assays. The *in vitro* reconstitution of protein-RNA complexes was carried out by adding 15.5K to 100 times and hPrp31 up to 300 times molar excess respectively in comparison with RNA. The protein concentration was determined by Bradford assay. Additionally 10 $\mu$g of *E. coli* tRNA and 0.2% of triton X-100 was added for decreasing unspecific binding. Gel filtration buffer containing DTT was then added to the mixture to an end volume of 20 $\mu$L. The samples was incubated at 4°C for 30-60 min and loaded onto 6-9% native polyacrylamide gels as described above. The gel was then transfered onto a Whatman paper and dried at 80 °C for 30 min. The result was visualized by exposing the radioactive gel to a Kodak film.

**Transformation of *E.coli***

The chemical competent cells were prepared with the $CaCl_2$-method. 1 $\mu$L of plasmid DNA were added to 50 $\mu$L chemical competent cells and incubated for 10 min on ice. Cells were subjected to a heat shock at 42 °C for 50 s and kept on ice for 5 min before supplemented with 1 mL of LB medium. After incubation for 30min at 37 °C and centrifugation for 0.5 min at 6000 x *g*, the cell pellet was resuspended in 200 $\mu$L medium and plated onto LB-agar plates containing appropriate antibiotics.

### 2.2.2 Protein biochemistry methods

**Protein expression in labeled media**

**Expression and isotope labeling of GST-15.5K** During the course of this work, only GST-15.5K wildtype protein was labeled with stable isotopes, ie. $^{15}$N, $^{13}$C and $^2$H. *E. coli* BL21(DE3) cells were transformed with pGEX-15.5K. Agar plates and all growth media contained ampicillin. A single colony of *E. coli* BL21(DE3) was used to inoculate a 2 mL LB pre-culture, which was grown for 4hrs. This pre-culture was transferred to a 50 mL overnight pre-culture. The 50 mL overnight pre-culture was then used to inoculate a 1 L LB culture. For producing labeled protein, the 50 mL LB and the 1 L LB culture were replaced by a 50 mL and a 1 L M9-minimal medium culture containing the required nitrogen and carbon source, namely ammonium chloride (>98 % $^{15}$N), $^{13}$C$_6$-D-glucose. In the case of deuterium labeling, M9-minimal medium culture was prepared with 99.9% D$_2$O. For perdeuteration, all buffers and stock solutions were prepared in 99.9% D$_2$O. All media contained 10 % v/v of Silantes rich growth medium. 2 mL M9-minimal medium pre-cultures with the deuterium concentration increasing from 33 %, 50 %, 75 % to 99.9 % and 4 g/L D$_8$-glycerol as the sole carbon source were grown for ca. 10 h each. Each pre-culture was centrifuged at 3500 x g. Cells from the pellets were picked using a pipette tip and transfered into the medium with the next deuterium concentration. Cell growth was monitored by measuring the OD$_{600}$. At an OD$_{600}$ of 0.5-0.6 the protein expression was induced by addition of 1 M IPTG to a final concentration of 1 mM. The cells were harvested after overnight incubation (ca 23 hrs) at an

$OD_{600}$ of 1.2-1.6 by centrifugation at 4000 x g and 4 °C for 30 min. The cell pellet from 1L of culture was resuspended in 5 mL of 1x PBS buffer and fast frozen in liquid nitrogen before being stored at -80 °C.

**Expression of hPrp31-His**   E. coli Tuner (DE3) pLacI cells were transformed with pTriEx-hPrp31 and were grown in LB medium with chloramphenicol (30 $\mu$g/mL), ampicillin (100 $\mu$g/mL) and 1 % glucose to an $OD_{600}$ of 0.8-1.2. After induction with 0.2 mM IPTG, the cells were incubated for an additional 3 hrs at 30 C before harvesting at 4000x g for 30min. The cell pellet from 1L of culture was resuspended in 10 mL of lysis buffer for His-tagged protein purification (see Table 2.9) before being frozen in liquid nitrogen and stored at -80 °C.

**Expression of GST-hPrp31$^{78-333}$**   The plasmid was transformed into E. coli BL21 (DE3)-RIL cells and the GST-hPrp31$^{78-333}$ fusion protein was expressed overnight at 20°C using the auto-induction method (138). With the auto-induction method, the use of IPTG to initiate the expression was not needed and the yield is several fold higher than the conventional IPTG induction method. A 200 mL overnight culture in LB-medium was pelleted at 3500g at 4°C for 10min. The pellet was resuspended in the 'ZY' medium with 20 mL of 50x5052 medium, 20 mL of 50xM medium, 1 mL of 2M $MgSO_4$ and 200 $\mu$L 1000x trace metal mixture added prior to usage (see Table 2.8). The 'ZY' medium used also contained chloramphenicol (30$\mu$g/mL) and carbenicillin (50$\mu$g/mL). The initial expression was at 37°C until an $OD_{600}$ of $\sim$ 0.4. The temperature was then decreased to 20°C and the culture continued to grow for a total expression time of 8 hrs. The cell pellet from 30 min

of centrifuging at 4000x g was resuspended in 1xPBS buffer (see Table 2.9) with pH 7.4 and stored at -80 °C.

**Expression of MBP-hPrp31**  HMS 174 (DE3) cells were transformed with pETM-40-hPrp31 and were grown in LB medium with kanamycin (34 $\mu$g/mL) to an $OD_{600}$ of 0.5-0.6. After induction with 1 mM IPTG, the cells were incubated for an additional 5 hrs at 22 C before harvesting at 4000x g for 30min. The cell pellet from 1L of culture was resuspended in 10 mL of lysis buffer for MBP-tagged protein purification (see Table 2.9) before being frozen in liquid nitrogen and stored at -80 °C.

**Protein purification**

**Purification of 15.5K and 15.5K mutants**  25 mL of cell resuspension in 1 xPBS buffer at pH 7.4 was incubated with lysozyme (100 $\mu$g/mL), 25 $\mu$L of RQI DNAse and DTT (1 mM) for 30 min at 4°C before sonification. The sonified lysate was then centrifuged at 30000g at 4°C for 1 hr. The supernatant was purified through a 1 mL GSTrap FF column using the Äkta Prime system at 4°C. After loading the supernatant, the column was washed with 20 mL of 1xPBS and the GST-tagged fusion protein was eluted with the elution buffer for GST-tagged protein purification (see Table 2.9). The peak fractions were combined and concentrated to a concentration of 1-2 mg/mL. The GST tag was cleaved using 1mg/U of thrombin protease at 22°C for 5-6 hrs. 15.5K and 15.5K mutants were then separated from the GST tag and thrombin using gel filtration method on a Superdex 200 26/60 column equilibrated with the gel filtration buffer (see Table 2.9).

**Purification of hPrp31-His**   160 mL of cell resuspension was incubated with DNase I (0.6 $\mu$g/mL), lysozyme (10 $\mu$g/mL) and Pefabloc SC (80 $\mu$g/mL) at room temperature for 30 min before sonification. The soluble extract after 1hr of centrifuge at 30000g at 4°C was purified via 3.3 mL of Ni-NTA agarose (Qiagen) using a 20 ml economy column (Bio-Rad). After incubating the supernatant and the agarose for 1hr at 4°C, the agarose was washed with 180 mL of wash buffer 1 for His-tagged protein purification (see Table 2.9). This was followed by washing with 180 mL of wash buffer 2 and wash buffer 3. The His-tagged fusion protein was then eluted from the agarose using the elution buffer (see Table 2.9). For the purpose of our study, the protein was dialyzed using dialysis memberane with 8000 MWCO against the gel filtration buffer.

**Purification of GST-hPrp31$^{78-333}$**   30 mL of cell resuspension was incubated with lysozyme (10 $\mu$g/mL), DNase I (0.6 $\mu$g/mL), Pefabloc (1 mM) and DTT (5 mM) for 30 min at 20°C before sonification. The soluble extract was obtained by centrifuging at 30000g at 4°C for 1 hr. After incubating the supernatant and the glutathione sepharose for 1hr at 4°C, the sepharose was washed with 200 mL of 1x PBS followed by 100 mL of the gel filtration buffer (see Table 2.9). The GST-tag was cleaved overnight at 4°C in 4 mL of gel filtration buffer using 300 $\mu$L of PreScission protease (2 $\mu$g/mL). hPrp31$^{78-333}$ was then eluted by washing the resin with gel filtration buffer and further purified using Superdex 75 26/60 gel filtration column equilibrated with the gel filtration buffer (see Table 2.9).

**Purification of MBP-hPrp31**   25 mL of cell resuspension was incubated with lysozyme (100 $\mu$g/mL), 30 $\mu$L of RQI DNAse and DTT (1 mM) for 30 min at 4°C before sonification. The soluble extract was obtained by centrifuging at 30000g at 4°C for 1 hr. After incubating the supernatant and the amylose resin for 1hr at 4°C, the resin was washed with 200 mL of wash buffer for MBP-tagged protein purification (see Table 2.9). MBP-hPrp31 protein was then eluted using the elution buffer described in Table 2.9. For the purpose of our study, the protein was dialyzed using dialysis memberane with 8000 MWCO against the gel filtration buffer.

**In *vitro* reconstitution of RNPs**   HPLC-purified U4 5'-SL RNA oligomer consisting of nucleotides 20-52 of U4 snRNA was obtained from iBA BioTAG-nology and was dialyzed overnight against the gel filtration buffer. Before adding to the protein, the RNA was annealed at 65°C for 90 sec. The complex of 15.5K-U4 5'-SL RNA was formed by mixing the two components in a 1:1 ratio and incubating at 4 C for 30 min. The binary complex was then added to dilute hPrp31 (including the C-terminal His6-tag,   0.4 mg/mL) purified as described above and incubated at 4°C for 1 hr. The ternary complex was concentrated using a 15 mL 10,000 MWCO Amicon centrifugal filter device. The final sample (300 $\mu$L) contained close to 0.1 mM $^2$H, $^{15}$N-labeled 15.5K in complex with an equimolar ratio of unlabeled U4 5'-SL RNA and a 1.5 molar ratio of unlabelled hPrp31.

To form the hPrp31$^{78-333}$-15.5K- U4 5'-SL complex, hPrp31$^{78-333}$ ($\sim$ 2.1 mg/mL), $^2$H, $^{15}$N-labeled 15.5K protein ($\sim$ 3.2 mg/mL) and U4 5'-SL RNA

oligomer were mixed in 1:1:2 molar ratio and incubated at 4 C for 1hr. The mixture was then applied to a Superdex 75 26/60 gel filtration column equilibrated with the gel filtration buffer. The peak fractions containing the ternary complex were then combined and concentrated to 300 $\mu$L with an end concentration of 0.2-0.3 mM for NMR studies.

## Concentration determination of proteins

Concentrations of proteins in solution were determined using Bradford assay. In this method proteins form complex with the dye, coomassie brilliant blue G in phosphoric acid solution. To carry out the assay, the protein sample is diluted with 800 $\mu$L of H2O, to which 200 $\mu$L of stain solution (Bio-Rad Protein Assay) is added. The mixture is homogenized by inverting vigorously and the colour is allowed to develop for 6 min. The absorbance of of the solution is measured at 595 nm wavelength in the UV photometer and concentration is calculated using calibration curve obtained from measuring the absorbance of BSA standards with different concentrations.

## Denaturation SDS-polyacrylamide gel electrophoresis

Using denatured SDS-polyacrylamide gel electrophoreses (SDS-PAGE) to analytically separate proteins was described by Laemmli (**?** ). According to different purpose, the thickness, the size and the degree of cross-linking are different. For this work, 10% and 13% polyacrylamide gels (37.5 : 1 acrylamide : bisacrylamide; 1 mm thickness) with 0.33% v/v of TEMED were used. Before loading, the protein samples were boiled with protein loading buffer at 95 °C for 3 min to ensure a complete denaturation. The gels were

stained with coomassie brilliant blue solution and dried at 80 °C for 1 hr.

### 2.2.3 NMR methods

**Sample preparations**

**Sample preparation of primary RNPs**    The ternary complexes for NMR studies were prepared as described above. As for ternary RNPs, for primary RNPs associated with 15.5K the RNA oligos obtained from iBA BioTAG-nology were all lyophilized and dialyzed against the gel filtration buffer. The buffer for all measured NMR samples was the gel filtration buffer (see Table 2.9). All RNA oligos were annealed at 65°C for 90 sec. For U4 snRNA constructs, 1:1 ratio of RNA and 15.5K were incubated for 20-30 min at 4°C. The complex was then concentrated using 10 mL 10,000 MWCO Amicon centrifugal filter device to $\sim$ 300 $\mu$L with concentration $\sim$ 0.3-0.5 mM. In the case of U14CD, U3CD and U3BC, the RNAs and 15.5K were added in a ratio of 2:1 and incubated as well as concentrated as for U4 snRNP. The final concentration of these RNPs was $\sim$ 0.3 mM.

**Preparation of anisotropic sample**    For RDC measurements, the isotropic sample of 15.5K-2-U4 5'-SL-33nt was aligned with the filamentous phage Pf1. Pf1 was purchased in a suspension of 10mM K-phosphate buffer pH /.6, 2mM MgCl$_2$ and 0.05% NaN$_3$ and was washed with the gel filtration buffer using ultra-centrifugation. The phage was then added to the isotropic samples to an end concentration of $\sim$ 10mg/ml. The sample was homogenized by gentle but extensive mixing and was then aligned under the magnetic field of the spectrometers for 1hr prior to the experiments.

**NMR experiments**

All NMR experiments were acquired using Bruker AVANCE 600 MHz and 900 MHz spectrometers equipped with cryo-probes. The experiments were run under X-WINNMR3.5 software. All NMR spectra were processed and analyzed using Felix-ND (2000.1), X-WINNMR3.5, Topspin2.0 and Sparky 3, in which Felix-ND was the most used program especially for 3D spectra processing and analysis.

**Experiments for backbone assignment**   Backbone resonance assignment was carried out on $^{15}$N, $^{13}$C-labelled and 70% deuterated, $^{15}$N, $^{13}$C-labelled 15.5K in complex with unlabelled U4 5'-SL-24nt. 70% deuteration was carried out to improve the quality of the 3D spectra. The sample concentration was $\sim$ 0.4mM. A more detailed description of the theory and applications of the 3D experiments for protein assignment is given in the Result chapter.

**Backbone assignment using non-TROSY type experiments**

The experiments used for obtain the backbone assignment of $^{15}$N, $^{13}$C-labelled 15.5K in complex with unlabelled U4 5'-SL-24nt were all carried out on a Bruker AVANCE 600 MHz spectrometer equipped with cryo-probe at 308 K and pH 7.6. The pulse programs were modified from the corresponding standard Bruker pulse sequences to suit this complex. These experiments are listed here in Table 2.14.

**TROSY type experiments used for backbone assignment in 70% deuterated, $^{15}$N, $^{13}$C-labelled 15.5K bound to U4 5'-SL-24nt**

The experiments used for obtain the backbone assignment of 70% deuter-

**Table 2.14:** Non-TROSY 3D experiments for backbone assignment

| Experiment | Spectrometer | Pulse program |
|---|---|---|
| 2D HSQC | AVANCE 600, cryo-probe | tchsqcwg.txt |
| 3D HNCA | AVANCE 600, cryo-probe | tchncagp3d.txt |
| 3D HN(CO)CA | AVANCE 600, cryo-probe | tchncocagp3d.2.txt |
| 3D HNCACB | AVANCE 600, cryo-probe | tchncacbgp3d.txt |
| 3D HNCOCACB | AVANCE 600, cryo-probe | tchncocacbgp3dnew.txt |
| 3D HNCACO | AVANCE 600, cryo-probe | tchncacogp3d.txt |
| 3D HNCO | AVANCE 600, cryo-probe | hncogp3d.txt |

ated, $^{15}$N, $^{13}$C-labelled 15.5K in complex with unlabelled U4 5'-SL-24nt were all carried out on a Bruker AVANCE 900 MHz spectrometer equipped with cryo-probe at 308 K and pH 7.6. The pulse programs were modified from the corresponding standard Bruker pulse sequences to suit this complex. These experiments are listed here in Table 2.15.

**Table 2.15:** TROSY version 3D experiments for backbone assignment

| Experiment | Spectrometer | Pulse program |
|---|---|---|
| 2D HSQC | AVANCE 900, cryo-probe | trosyetf3gpsi |
| 3D HNCA | AVANCE 900, cryo-probe | tctrhnca2h |
| 3D HNCACB | AVANCE 900, cryo-probe | tctrhncacbgp2h3d |
| 3D HNCOCACB | AVANCE 900, cryo-probe | tctrhncocacbgp2h3d |
| 3D HNCACO | AVANCE 900, cryo-probe | tctrhncacogp2h3d |
| 3D HNCO | AVANCE 900, cryo-probe | tctrhncogp2h3d1 |

**Other experiments**    Further experiments including TROSY-HSQC experiments and cross-saturation experiments on the ternary complexes hPrp31-15.5K-U4 5'-SL and hPrp31$^{78-333}$-15.5K-U4 5'-SL complexes, as well as the 3-interleaved IPAP experiments used for extracting RDCs (residue dipolar couplings) are listed in Table 2.16. The experiments on the hPrp31$^{78-333}$-15.5K-U4 5'-SL complex were carried out at 303 K. A more detailed description on these experiments is given in the Result chapter.

**Table 2.16:** Other NMR experiments

| Experiment | Spectrometer | Pulse program |
| --- | --- | --- |
| 2D TROSY-HSQC | AVANCE 900, cryo-probe | trosyetf3gpsi |
| cross-saturation | AVANCE 900, cryo-probe | tcsattrosy |
| IPAP-HSQC | AVANCE 600, cryo-probe | hsqc-diffAB.vivi.cf |

**Chemical shift analysis**

Changes in chemical shift values give important structural information about the molecules. It provides a rapid and easy way to study the interacting molecules. The chemical shift values of $^{1}$H and $^{15}$N are especially sensitive to the change in the local environment of the nuclei. Therefore, chemical shift perturbations recorded using $^{15}$N-$^{1}$H-HSQC spectra through an NMR titration are commonly used for revealing the structural changes upon a binding event. A combined average change in the $^{15}$N and $^{1}$H dimensions of the 2D spectra is used and is expressed as

$$\delta_{NH} = \sqrt{\frac{(\delta_N/5)^2 + (\delta_H)^2}{2}} \qquad (2.1)$$

where $\delta_{NH}$ is the average amide chemical shift perturbation whereas $\delta_N$ and $\delta_H$ are the chemical shift perturbations of amide nitrogen and amide hydrogen respectively (149). The residues, which show large $\delta_{NH}$ upon NMR titration are likely to be involved at the interaction surface of the molecule. All the chemical shift analysis in this work was calculated using this formula.

### 2.2.4 Structural calculation

During the frame of this work, a 3D model of the hPrp31$^{188-332}$ was generated using comparative modelling via the web server SWISS-MODEL (see Table 2.13), molecular dynamics simulation protocols from XPLOR-NIH and HADDOCK2.0 program (141). The detailed procedure and the essence of the methods used are explained in the related sections in the Result chapter.

# Chapter 3

# Results

## 3.1 Backbone assignment of 15.5K

### 3.1.1 Introduction to protein backbone assignment

In NMR spectroscopy nuclear spins experience a magnetic field which is the superposition of an externally applied magnetic field $B_0$ and an internal field which is induced by the orbiting electrons. As the result, the observed resonance frequencies of the individual nuclei differ from each other and depend on the local environments (i.e. electron densities) around the nuclei. This phenomenon is called the chemical shift and is a basic feature of NMR. The crucial first step in the NMR investigation of proteins is to assign the chemical shift observed for each atom in the molecules. As the chemical shift is highly sensitive to changes in the local environment, it is routinely used to monitor changes in the protein upon binding events. To assign all proton resonances of proteins with MW > 8 kDa was once an impossible task due
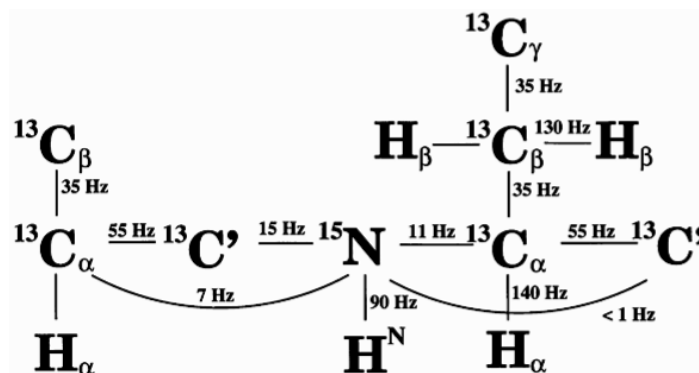
**Figure 3.1:** An example of a protein peptide chain is shown here with $^1$J and $^2$J couplings between the atoms indicated. These J couplings are used in triple-resonances experiments on $^{15}$N and $^{13}$C labelled proteins (figure taken from Fig.2 in (153)).

to signal overlap (4). A big step forward was made with the development of heteronuclear three-dimensional experiments, for which the protein sample needs to be $^{13}$C and $^{15}$N labelled. These experiments rely exclusively on $^1$J/$^2$J couplings between H, N and C nuclei of the protein and aim to simplify the spectrum content by disentangling overlapped peaks of the two-dimensional spectrum by separating them into a third dimension (150; 151; 152; 18; 153) (Figure 3.1). As in triple resonance experiments the chemical shifts of $^1$H, $^{15}$N and $^{13}$C are correlated, the protein need to be first of all labelled with $^{15}$N and $^{13}$C isotopes. Isotope labelling of proteins requires the over-expression to be carried out in minimal media containing $^{15}$N and $^{13}$C enriched nitrogen and carbon sources, e.g. $^{15}$NH$_4$Cl and $^{13}$C-glucose.

To assign the backbone resonance of a protein, a $^1$H$_N$-$^{15}$N HSQC (heteronuclear single-quantum coherence) spectrum is first recorded. In HSQC experiments

the magnetization of a backbone amide proton ($^1H_N$) is transferred to the connected nitrogen ($^{15}N$) through $^1J$ coupling via the INEPT (insensitive nuclei enhanced by polarization transfer) step, then follows the chemical shift evolution of $^{15}N$ for $t_1$ (evolution time) and finally the magnetization on $^{15}N$ after the $t_1$ period is transferred back to $^1H_N$ for observation (154). In the resulting spectrum the amino acids of the protein, each corresponding to a single peak, is located with the chemical shifts of its amide protein in one dimension and that of the directly connected amide nitrogen in the second dimension. The task is then assignment: identifying the corresponding residue in the protein sequence, which gives a peak in the spectrum (i.e. assignment). This task is carried out with the help of triple resonance experiments usually including HNCA, HN(CO)CA, HNCACB, HN(CO)CACB, HN(CA)CO and HNCO experiments.

In HNCA experiments, the chemical shifts from $^1H_N$ and $^{15}N$ are correlated with the intra-residue $^{13}C_\alpha$ ($^{13}C_{\alpha i}$) shifts through the $^1J$ coupling (7-11 Hz) between $^{15}N$ and $^{13}C_\alpha$ (see Figure 3.1). Furthermore, it gives sequential connectivity by transferring the coherence from the $^{15}N$ spins to the $^{13}C_\alpha$ of the previous residue ($^{13}C_{\alpha(i-1)}$) (155). In the spectrum, for each of $^1H_N$ and $^{15}N$ chemical shift combinations, which are as the ones shown in the HSQC spectrum, there are two peaks resulting from $^{13}C_{\alpha i}$ and $^{13}C_{\alpha(i-1)}$ correlations. To distinguish the peak of $^{13}C_{\alpha(i-1)}$ from $^{13}C_{\alpha i}$, the HN(CO)CA experiment is recorded in which only the resonance from $^{13}C_{\alpha(i-1)}$ is correlated by transferring coherence via the intervening $^{13}CO$ of the previous residue (156; 157). However, the chemical shift of a $^{13}C_\alpha$ alone is usually

not characteristic enough for determining the amino acid type of the corresponding residue. For this purpose, experiments involving chemical shifts of the side-chain carbon, $^{13}C_{\beta}$, have been developed. Information about the chemical shifts of both $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ of a residue is extremely important because they are much more characteristic for different amino acid types. The HNCACB experiment is such an experiment in which magnetization starting from $^{1}H_{N}$ to $^{15}N$ is transferred to $^{13}C_{\alpha i}$ and is then further transferred to $^{13}C_{\beta i}$ as well as to $^{13}C_{\beta(i-1)}$ (158). In this spectrum there are four peaks which correspond to $^{13}C_{\alpha i}$, $^{13}C_{\alpha(i-1)}$, $^{13}C_{\beta i}$ and $^{13}C_{\beta(i-1)}$ for each $^{1}H_{N}$ and $^{15}N$ combination with the exception of glycines, which do not posses a $^{13}C_{\beta}$. As in the case of HNCA experiments, one needs to distinguish the peaks of the residue itself from the ones of the previous residue. To achieve this, the HN(CO)CACB experiment is used in which only $^{13}C_{\alpha(i-1)}$ and $^{13}C_{\beta(i-1)}$ peaks are observed. HNCO and HN(CA)CO experiments help to confirm the assignment even further by correlating $^{1}H_{N}$ and $^{15}N$ with the carbonyl groups $^{13}CO_{i}$ and $^{13}CO_{i-1}$ in which HNCO, the most sensitive experiment, only shows the peaks from $^{13}CO_{i-1}$'s (Figure 3.2). Having analyzed all these spectra, one could group the chemical shift values of $^{13}C_{\alpha}i$, $^{13}C_{\alpha(i-1)}$, $^{13}C_{\beta i}$, $^{13}C_{\beta(i-1)}$, $^{13}CO_{i}$ and $^{13}CO_{i-1}$ together for each amino acid. The exact location of this amino acid in the protein sequence is then found out by tracking upstream in the peptide sequence, namely finding an amino acid with the chemical shift values matching the $^{13}C_{\alpha(i-1)}$, $^{13}C_{\beta(i-1)}$ and $^{13}CO_{i-1}$ chemical shifts of this amino acid. Fragments of such connectivities can then be pieced together to finally obtain the assignment of the complete backbone.

| Experiments | Correlations observed | Magnetization transfer |
|---|---|---|
| **HNCA** | $^1H_i^N - {}^{15}N_i - {}^{13}C_i^\alpha$ $^1H_i^N - {}^{15}N_i - {}^{13}C_{i-1}^\alpha$ | |
| **HN(CO)CA** | $^1H_i^N - {}^{15}N_i - {}^{13}C_{i-1}^\alpha$ | |
| **HNCACB** | $^1H_i^N - {}^{15}N_i - {}^{13}C_i^\alpha - {}^{13}C_i^\beta$ $^1H_i^N - {}^{15}N_i - {}^{13}C_{i-1}^\alpha - {}^{13}C_{i-1}^\beta$ | |
| **HN(CO)CACB** | $^1H_i^N - {}^{15}N_i - {}^{13}C_{i-1}^\alpha - {}^{13}C_{i-1}^\beta$ | |
| **HN(CA)CO** | $^1H_i^N - {}^{15}N_i - {}^{13}CO_i$ $^1H_i^N - {}^{15}N_i - {}^{13}CO_{i-1}$ | |
| **HNCO** | $^1H_i^N - {}^{15}N_i - {}^{13}CO_{i-1}$ | |



**Figure 3.2:** Schematic drawing indicating the magnetization transfer in standard triple resonance experiments required for backbone assignment. The nuclei, between which correlations can be observed, are indicated with red dots. The transiently involved nuclei are in open circles. Magnetization transfers are indicated with double-headed arrows.

### 3.1.2   Backbone assignment of 15.5K

15.5K was at first uniformly labelled with $^{15}$N, whereas U4 5'-SL constructs used in all experiments of this work are unlabelled. Two RNA constructs were used in the experiments namely the U4 5'-SL-24nt and the U4 5'-SL-33nt construct. The 24nt construct comprises of the 22nt construct (nucleotides 26-47 of U4 snRNA) with one additional C-G base pair at the end of stem I for stabilization (Figure 3.3 A). The U4 5'-SL 33nt construct, representing the entire U4 5'-SL (nucleotides 20-52 of U4 snRNA), was shown to contain the minimal length of stem I required for hPrp31 binding (134). As the RNAs are non-labelled, the observed resonances in two-dimensional and three-dimensional experiments are solely from the 15.5K protein. $^{1}$H$_N$-$^{15}$N HSQC of the protein alone and the protein bound to 1:1 molar ratio of the 24nt U4 5'-SL construct were measured at 308 K and pH 7.6 (Figure 3.3 B). Peaks in both HSQC spectra showed good chemical shift dispersion which indicates that 15.5K is folded alone as well as upon binding to RNA. However, 15.5K protein alone showed much less stability and precipitated readily during the measurement. The stability was greatly improved when the protein was bound to the RNA. This observation is supported by the fact that 15.5K protein alone failed to crystallize, indicating an intrinsic instability of the free protein. Interestingly there are 102 peaks from backbone $^{1}$H$_N$ in the spectrum of 15.5K protein alone, whereas the spectrum with 15.5K-U4 5'-SL-24nt contains 114 resonances. 15.5K protein contains 128 amino acids in which 6 prolines and the first amino acid, methionine, will not be observed in this spectrum. Prolines lack backbone amide protons and the amine protons of the first amino acid in a peptide exchange very fast with

the solvent and are largely broadened. Taking this into consideration, there should be 121 peaks present in the $^1\text{H}_N$-$^{15}\text{N}$ HSQC. However, neither of the spectra recorded contains full number of peaks. There are additional 19 and 7 peaks missing in the spectra of the free and bound 15.5K respectively. $^1\text{H}_N$-$^{15}\text{N}$ HSQC of 15.5K bound to U4 5'-SL-33nt construct, which is the minimal RNA construct capable of forming ternary complex with 15.5K and hPrp31 protein, was also measured at 308K. The spectra of 15.5K bound to both U4 5'-SL constructs are identical, which confirms the previous findings that 15.5K only contacts the (5+2) internal loop and the stem II phosphate backbone.

15.5K was then uniformly labelled with $^{15}\text{N}$ and $^{13}\text{C}$. Due to the intrinsic instability of 15.5K protein alone, all triple resonance experiments were carried out on the 15.5K-U4 5'-SL-24nt complex. The sample concentration of the complex was ∼0.4-0.5 mM. Higher sample concentration could not be reached, as then precipitates were formed during the experiments. The complex size is ∼ 26 kDa. To decrease the relaxation effect from the relatively large complex size by increasing the molecular tumbling rate, all of the triple-resonance experiments were recorded at 308 K. Useful information for backbone assignment was obtained from HNCA, HN(CO)CA, HNCACB, HN(CO)CACB, HN(CA)CO and HNCO experiments as stated above. As mentioned before, there are 7 resonances missing in the $^1\text{H}_N$-$^{15}\text{N}$ HSQC of the 15.5K-U4 5'-SL-24nt complex. The resonances were assigned to amino acids 109-115, which are located in the loop connecting $\beta 4$ and $\alpha 5$ of the protein. For the observed backbone resonances, 87% could be assigned. 87.7% of the $^{13}\text{C}_\alpha$, 81% of the $^{13}\text{CO}$ and 81% of the $^{13}\text{C}_\beta$ resonances could be assigned
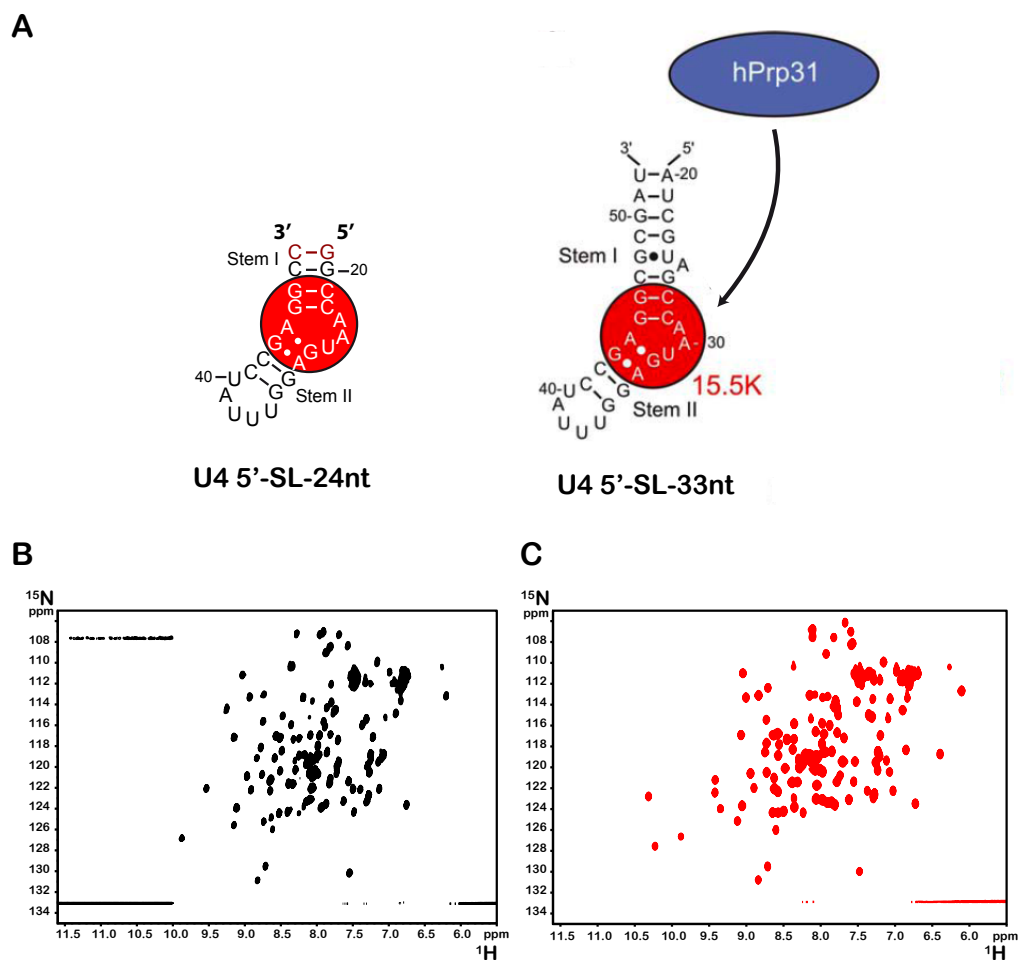
**Figure 3.3:** (A) The two U4 5'-SL constructs used in the NMR investigations. On the left, the U4 5'-SL-24nt has one additional C-G base pair (red) added to the construct in the 15.5K-U4 5'-SL crystal structure. On the right is the minimal RNA construct, which still enables ternary complex formation with hPrp31. (B) The HSQC spectrum of 15.5K alone. (C) The HSQC spectrum of 15.5K-U4 5'-SL-24nt complex.

unambiguously.

The sensitivity of all heteronuclear triple-resonance experiments decreases as the size of the protein increases. This is due to the result of increasing $T_2$ relaxation rates and affect the signal intensity of amide protons and aliphatic carbons in particular through the effective dipolar coupling between $^1$H and $^1$H as well as $^{13}$C and $^1$H (153). To overcome this problem, partial ($\sim 70\%$) or complete ($\sim 100\%$) deuteration of the protein is needed. In a deuterated protein, the relaxation time of aliphatic carbons is increased due to the suppression of the efficient relaxation mechanism caused by the dipolar coupling between $^2$H and $^{13}$C. This consequently leads to increased signal intensities and higher resolution. In addition, TROSY type experiments are used for large size proteins. This NMR technique selects the line in a coupled $^{15}$N-$^1$H correlation that has a longer relaxation time due to mutual cancellation of the $R_2$ relaxation rate of N and the CSA-dipole cross correlated relaxation rate of N-H$_N$ (9). 15.5K was 70% randomly deuterated and TROSY versions of the triple-resonance experiments mentioned above were recorded. This led to 99% of the observed backbone resonances being assigned. 99.1% the $^{13}$C$_\alpha$, 98.2% of the $^{13}$CO and 97.4% of the$^{13}$C$_\beta$ could be unambiguously assigned (Figure 3.4).

With the assignment result, the chemical shift changes exhibited in the $^1$H$_N$-$^{15}$N HSQC spectrum of RNA bound 15.5K in comparison to that of the unbound form could be identified (Figure 3.5) and the degree of the changes were colour-coded on the 15.5K crystal structure (Figure 3.6). As mentioned above, there are 19 resonances missing in the $^1$H$_N$-$^{15}$N HSQC spectrum of
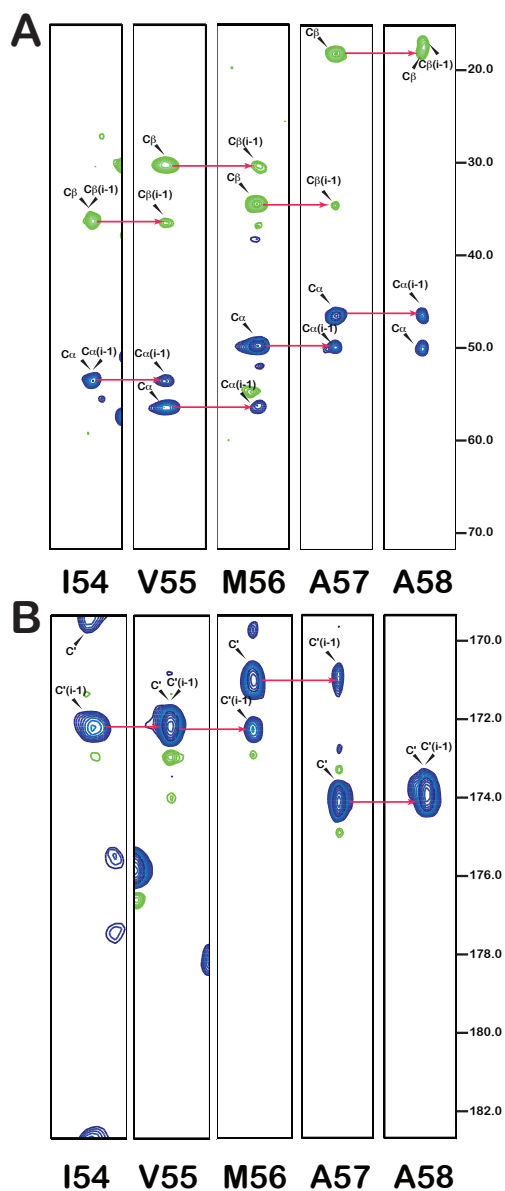
**Figure 3.4:** (A) 2D strips from the HNCACB experiment of 5 sequentially connected residues. Chemical shift of the $^{13}C_{\alpha(i-1)}$ of an amino acid takes the same value as the chemical shift of $^{13}C_{\alpha(i)}$ of the preceding amino acid, as is shown by red arrows. The same applies to $^{13}C_\beta$. Thus, the $^{13}C_\alpha$ and $^{13}C_\beta$ resonances can be assigned. (B) 2D strips from the HN(CA)CO experiment of 5 sequentially connected residues. The same principle allows the assignment of the $^{13}CO$ resonances.

the free 15.5K, 7 of which belong to amino acids 109-115 in the loop and 12 peaks seem to newly appear in the spectrum of the bound form. These peaks belong to K37, G38, A39, N40, E41, A42, T43, E61, K86, C93, R97 and V99 in which residue 37-42, 61, 86 and 97 have direct contact with the RNA in the crystal structure. Residue I65, V95 and I100, which showed large shifts, also form direct contact with the RNA. The remaining three RNA contacting residues namely R36, K44 and R48, showed medium sized shifts. Small chemical shift changes were observed for residues located away from the RNA binding site.

## 3.2 hPrp31-15.5K-U4 snRNA complex

### 3.2.1 Full deuteration of 15.5K, expression of hPrp31/hPrp31$^{78-333}$ and ternary complexes preparation

For the formation of ternary complex, the U4 5'-SL 33nt construct representing the entire U4 5'-SL (residues 20-52 of U4 snRNA) was used as it was shown to contain the minimal length of stem I required for hPrp31 binding (134). Constructs of hPrp31 were designed and cloned by our collaboration partner Dr. Sunbin Liu from the group of Dr. Markus Wahl. It was shown that treatment of the ternary complex containing the full length hPrp31 with trypsin generated a hPrp31 fragment that remained associated with 15.5K and the RNA. Isolated hPrp31, in contrast, was completely degraded. Based on N-terminal sequencing and the apparent size of the protease-resistant portion, hPrp31$^{78-333}$, which was C-terminally truncated after the predicted
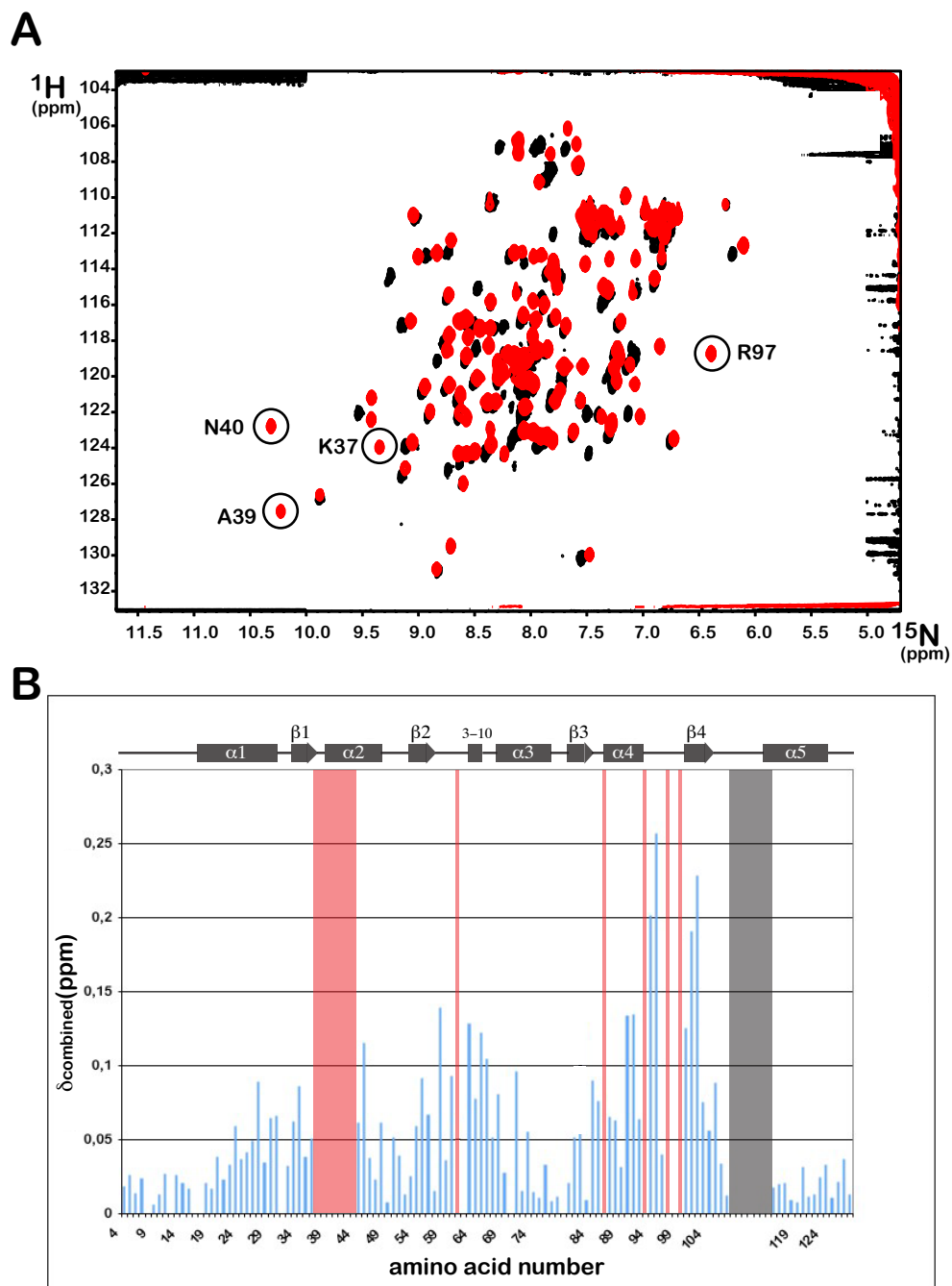
**Figure 3.5:** (A) Overlay of the HSQC spectrum of 15.5K alone (black) and
15.5K-U4 5'-SL-24nt (red). New resonances in the HSQC of the complex are
indicated by residue name. (B) Histogram plot of the chemical shift changes
observed in the HSQC of the complex as compared with that of the free protein.
Amino acid numbers are indicated on the x-axis. The combined chemical shift
changes in both $^1H_N$ and $^{15}N$ are plotted. Red bars indicate new peak in the
HSQC of the bound form.  Grey bar indicates the unassigned flexible loop.
The secondary structure of 15.5K is indicated above.
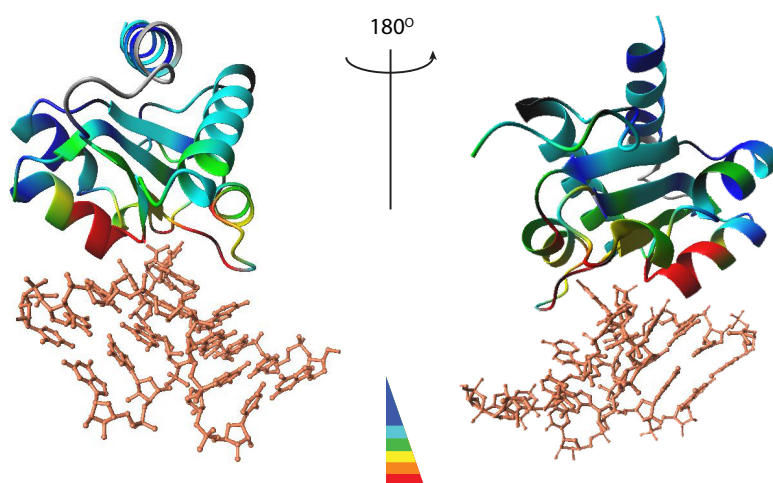
**Figure 3.6:** The degrees of chemical shift changes observed by comparing the HSQC spectra of the RNA bound and the free 15.5K are colour coded on the crystal structure of 15.5K-U4 5'-SL-22nt. Red indicates the new resonances in the HSQC of the bound form. The colours code ranges from orange to dark blue indicating decreasing chemical shift changes.

Nop domain, was chosen. hPrp31$^{78-333}$ was used in crystallization trails, because flexible regions in full-length hPrp31 possibly prevented crystallization of the ternary hPrp31-15.5K-U4 5'-SL complex. Gel filtration and electrophoretic gel mobility shift assays by Sunbin Liu showed that hPrp31$^{78-333}$ binds the 15.5K-RNA complex and that binding is strictly dependent on prior 15.5K binding. For NMR investigation, two constructs of hPrp31 were used namely the full length (499 amino acid long) construct with a His$_6$-tag and the hPrp31$^{78-333}$ construct.

In order to carry out a structural investigation of the ternary complex, the 15.5K protein needs to be fully deuterated. All non-exchangeable $^1$H of the protein were substituted with $^2$H. This is important mainly due to two reasons. Firstly, the sizes of the ternary complexes to be studied are very large; 81 kDa for the ternary complex containing hPrp31 full length and 57 kDa for the complex with hPrp31$^{78-333}$. As stated above, large molecular weight leads to fast relaxation and therefore low signal intensity and peak broadening. Higher resolution is achieved by deuteration through the lowering of homonuclear proton/proton relaxation pathways (spin diffusion) (11). Furthermore, to perform the cross-saturation experiment discussed Section 2.2.3 in detail, complete deuteration of 15.5K is required. 10% Silantes rich medium was added to the 100% $^2$H$_2$O containing minimal medium with d$_8$-gylcerol as the sole carbon source (see Material and Method chapter). The addition of the rich medium greatly improved the cell growth and consequently the yield was comparable to that of normal minimal medium, which is ∼8 mg/L. In the 1D $^1$H spectrum of a protonated protein, the S/N ratio of
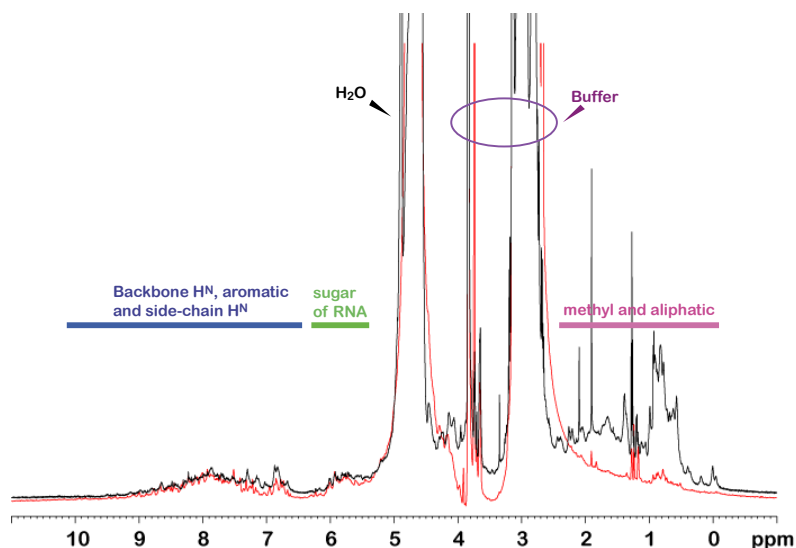
**Figure 3.7:** The 1D spectrum of $^{15}$N-15.5K-U4 5'-SL is shown in black and that of $^2$H, $^{15}$N-15.5K-U4 5'-SL is shown in red. Sources of the observed proton resonances are indicated as well as the peaks from the solvent. A $\sim$ 30 fold signal reduction was observed for the aliphatic protons in the sample with deuterated 15.5K.

the amide protons is usually less than one-third of the value for methyl groups in side chains. In the 1D $^1$H spectrum of fully deuterated 15.5K, the signal from the methyl groups are efficiently diminished by $\sim$30 fold (Figure 3.7).

The full length hPrp31 and hPrp31$^{78-333}$ were both expressed and purified as non-labelled form (see Material and Method chapter). For hPrp31-15.5K-U4 5'-SL complex, the complex of 15.5K-U4 5'-SL RNA was preformed in 1:1 ratio. The binary complex was then added to dilute hPrp31 (including the C-terminal His6-tag, $\sim$ 0.4 mg/mL). The ternary complex was concentrated and final sample (300 $\mu$L) contained close to 0.1 mM $^2$D, $^{15}$N-labelled 15.5K in complex with an equimolar ratio of unlabelled U4 5'-SL RNA and a 1.5 molar ratio of unlabelled hPrp31. In the case of the hPrp31$^{78-333}$-15.5K-U4 5'-SL

complex, the complex of 15.5K-U4 5'-SL RNA was preformed in 1:1 ratio and added to hPrp31$^{78-333}$ (without N-terminal GST-tag, $\sim$ 2.1 mg/mL). The ternary complex was then purified via gel filtration and concentrated to 300 $\mu$L with the ternary complex concentration of close to 0.3 mM (Figure 3.8).

### 3.2.2   Chemical shift analysis

$^1$H$_N$-$^{15}$N TROSY-HSQC ([9]; [159]) experiments on hPrp31-15.5K-U4 5'-SL and hPrp31$^{78-333}$-15.5K-U4 5'-SL complexes were carried out at 308 K and 303 K respectively on a Bruker 900 MHz spectrometer equipped with cryo-probe (Figure 3.9). The total experiment time was 12 h for the each experiment. Comparing the spectrum of hPrp31-15.5K-U4 5'-SL complex with that of the 15.5K-U4 5'-SL complex measured before the addition of the hPrp31, chemical shift changes could be observed and located primarily on helices $\alpha$2 and $\alpha$3 of 15.5K (Figure 3.10). The same experiment was carried out for hPrp31$^{78-333}$-15.5K-U4 5'-SL complex and showed strikingly similar chemical shift pattern to that observed in the ternary complex with full length hPrp31.

Helix $\alpha$2 of 15.5K showed the most changes in both complexes. Large changes on helix $\alpha$2 of 15.5K were also exhibited by residues K37, A39, N40, E41, A42, T43, K44, T45 and N47. Residues E61, L63 and I66 located close to and on the 3-10 helix and L72 on helix $\alpha$3 of 15.5K also showed large chemical shift changes upon ternary complex formation. Significant shifts were also observed for A10, which resides on the unstructured region before helix $\alpha$1 of 15.5K and K86, which locates on helix $\alpha$4 of 15.5K. However, chemical shift perturbations alone do not necessarily suggest direct contacts between the proteins, as the changes could be mediated through U4 5'-SL.
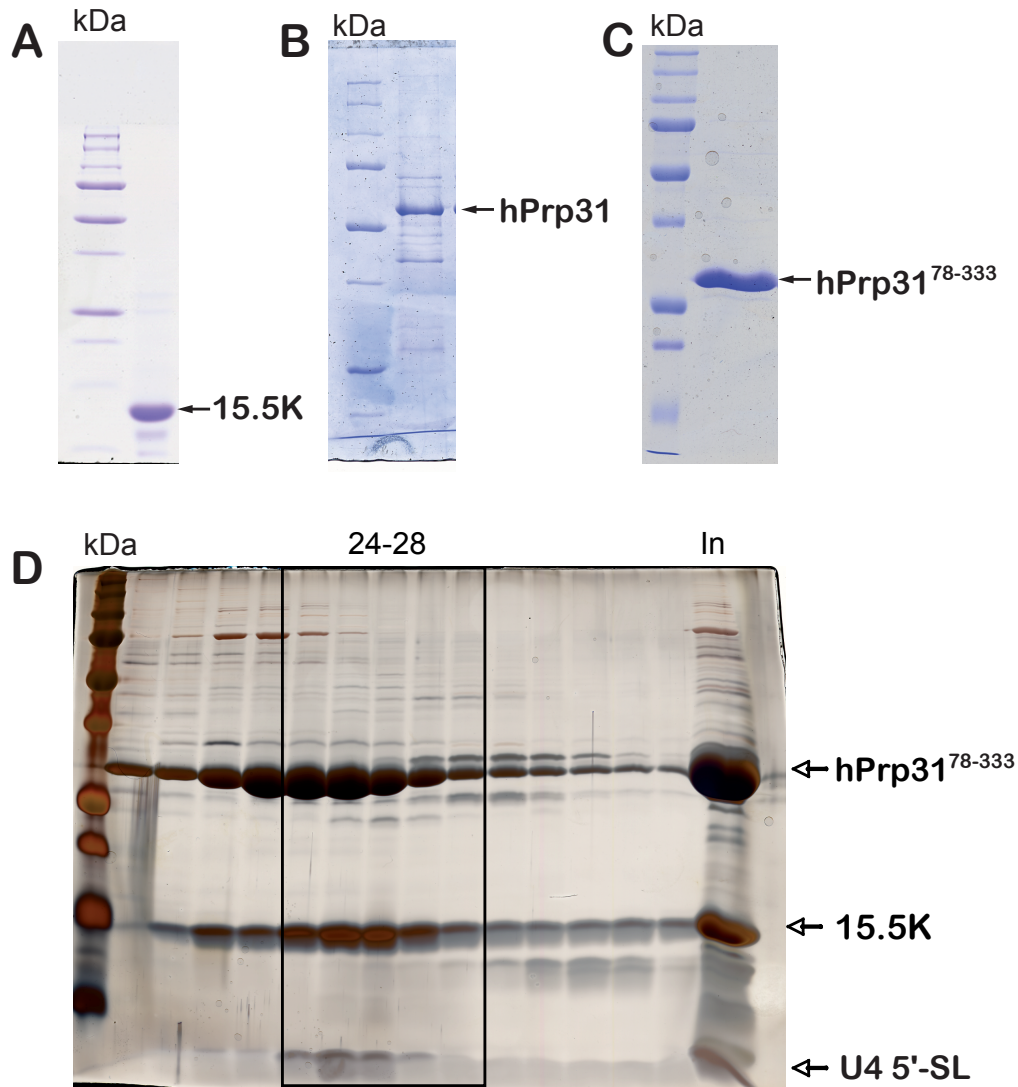
**Figure 3.8:** (A) $^2$H, $^{15}$N-15.5K purified using gel-filtration chromatography; (B) hPrp31 full length protein purified using affinity chromatography; (C)hPrp31$^{78-333}$ purified using affinity chromatography; (D) The peak fractions from the purification of hPrp31$^{78-333}$-15.5K-U4 5'-SL complex using gel-filtration chromatography are shown. The input is shown in the last lane on the right and indicated with n'. Fractions 24-28 containing all three components of the hPrp31$^{78-333}$-15.5K-U4 5'-SL complex were taken for the NMR experiments. (A,B,C) are pictures taken from coomassie stained gels. (D) was silver stained to also show the RNA. The molecular weight marker is indicated with 'kDa'.
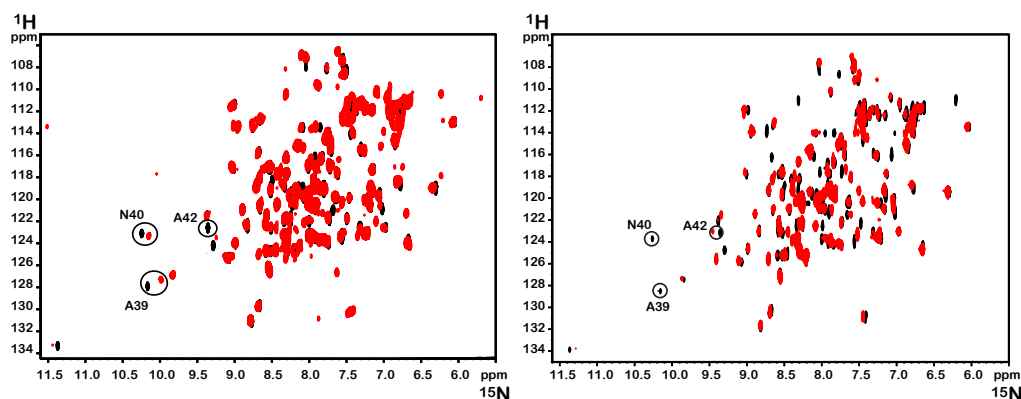
**Figure 3.9:** TROSY-HSQC spectra of the hPrp31-15.5K-U4 5'-SL complex (left, red) and the hPrp31$^{78-333}$-15.5K-U4 5'-SL complex (right, red) overlaid with that of the 15.5K-U4 5'-SL (black). Three of the resonances, which showed large chemical shift changes upon ternary complex formation are indicated in open circles.

To verify that the chemical shift changes observed on 15.5K were elicited by direct protein-protein contact, cross-saturation experiments were performed.

## 3.2.3   Saturation transfer analysis

**The saturation transfer experiment**

The saturation transfer experiment as described by Takahashi *et al* in year 2000 (160) was performed on both ternary complexes. In this experiment the protein for which the interaction surface will be defined, in this case the 15.5K protein, is uniformly labelled with $^2$H and $^{15}$N. This protein then forms complex with a non-labelled target protein, in this case hPrp31 or hPrp31$^{78-333}$. The complex now contains protein with lower (15.5K) and higher (hPrp31 or hPrp31$^{78-333}$) proton densities. When the aliphatic protons of the protein
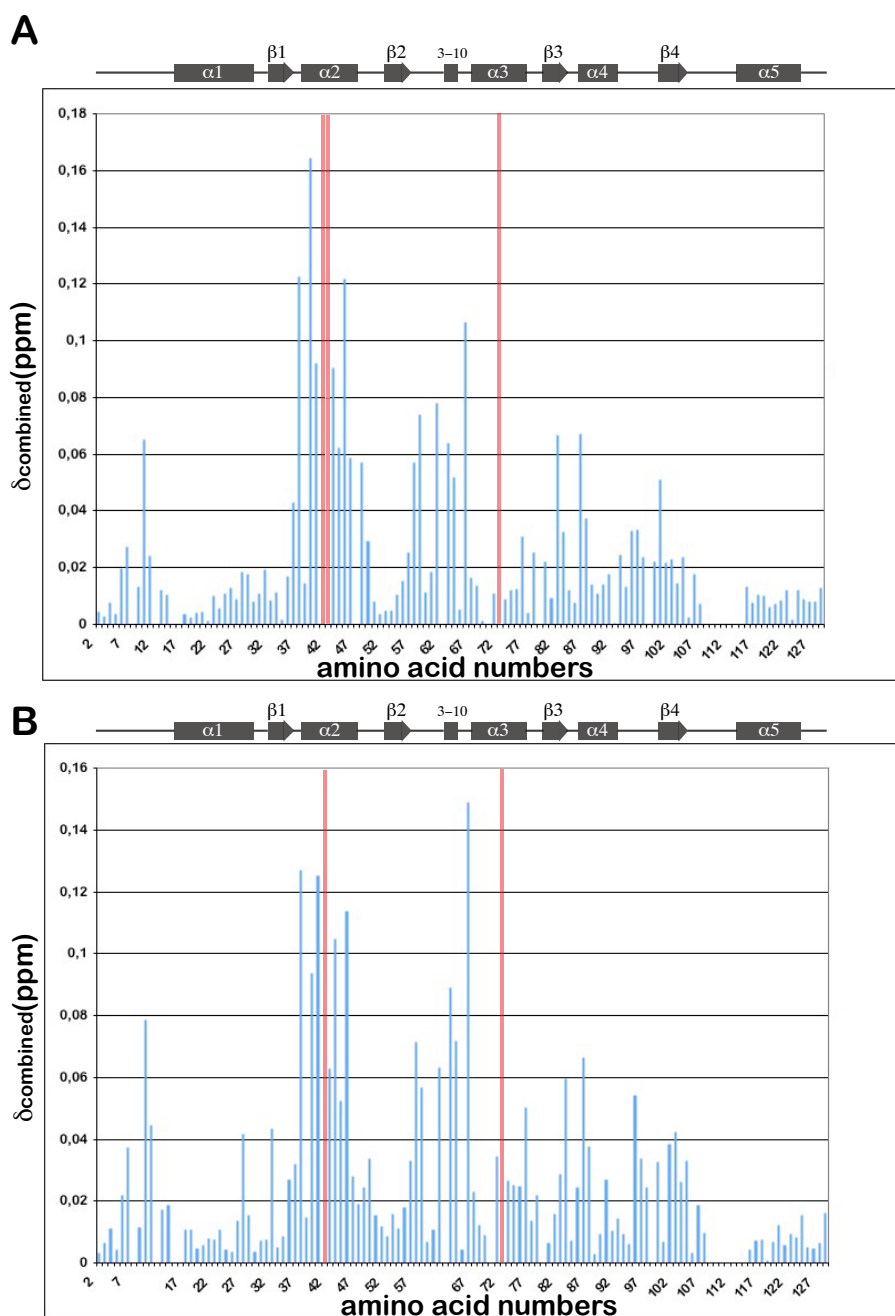
**Figure 3.10:** Combined chemical shift changes in H-N resonances of 15.5K observed upon ternary complex formation are presented here as histograms. The changes are with respect to the chemical shifts of 15.5K in the primary complex with U4 5'-SL-33nt. (A) Combined chemical shift changes observed upon hPrp31-15.5K-U4 5'-SL ternary complex formation. (B) Combined chemical shift changes observed upon hPrp31$^{78-333}$-15.5K-U4 5'-SL complex formation. The red bars indicate resonances with so large chemical shift changes that were not able to be identified upon ternary complexes formation. These residues locate on helices $\alpha 2$ and $\alpha 3$ of 15.5K
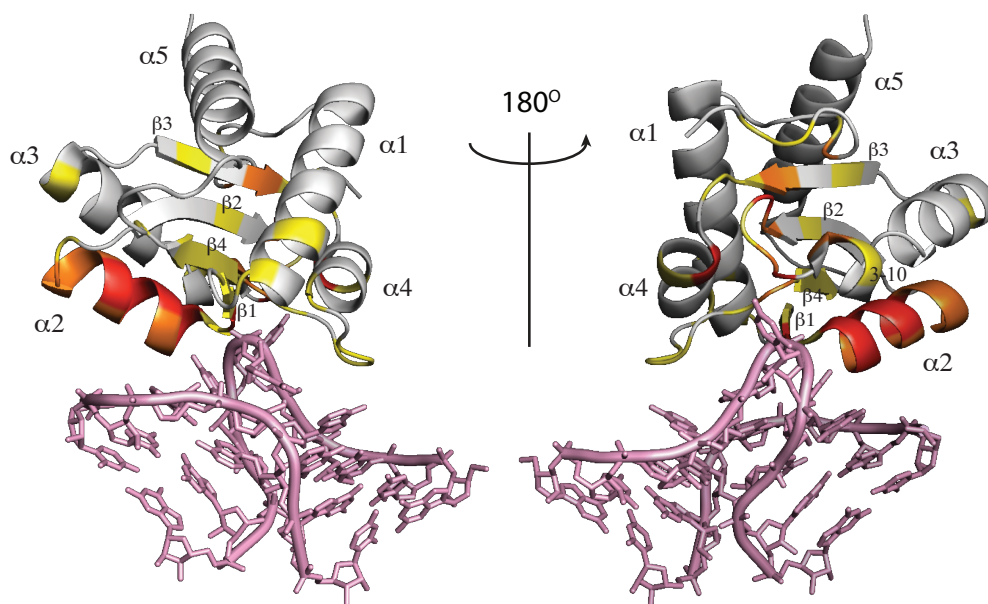
**Figure 3.11:** The degrees of combined chemical shift changes observed in both the hPrp31-15.5K-U4 5'-SL and the hPrp31$^{78-333}$-15.5K-U4 5'-SL complexes are plotted on the crystal structure of 15.5K. Red indicates a change bigger than 0.06 ppm; orange, a change between 0.04 and 0.06 ppm; yellow, a change between 0.02 and 0.04 ppm.

with higher proton density are irradiated with a selective radio frequency (RF) field, the aliphatic protons will be saturated instantaneously(160) (Figure 3.12 A). The deuterated 15.5K protein, on the other hand, is not affected directly by the RF field, as $^2$H and $^1$H have very different resonance frequencies because of the difference in their gyromagnetic ratios: that of $^2$H is about one-sixth of that of $^1$H. However, the saturation effect can be transferred from the irradiated protein onto the deuterated protein via cross-relaxation and when the deuteration level is high enough, 100% deuteration in this case, the saturation transfer will be confined to the interaction surface between the two proteins. Residues at the interaction surface can thus be identified by monitoring the reduction of signal intensities in the $^1$H$_N$-$^{15}$N HSQC spectrum. $^1$H$_N$-$^{15}$N HSQC spectrum can be measured for fully deuterated protein, because the amide $^2$H are back exchanged to $^1$H in water. The pulse sequence consists of an alternative band selective WURST-2 decoupling scheme (161), followed by a water flip-back TROSY-HSQC experiment (9) (Figure 3.12 B).

In this work, cross-saturation experiments on both hPrp31-15.5K-U4 5'-SL and hPrp31$^{78-333}$-15.5K-U4 5'-SL complexes were carried out on a Bruker 900 MHz spectrometer equipped with a cryo-probe and the total experiment time was about four days for each experiment. Cross saturation experiments were conducted in a mixture of 40 %$^1$H$_2$O, 60 %$^2$H$_2$O (containing 20 mM HEPES, pH 7.6, 120 mM NaCl and 5 mM DTT) to minimize spin diffusion effects due to bound water. As the amide protons are in exchange with water, the saturation effect can be carried by water to parts of the protein other than the interaction surface. This was observed by Takahashi *et al* that in a buffer with 90 %$^1$H$_2$O, 10 %$^2$H$_2$O the distinctive saturation transfer pattern could
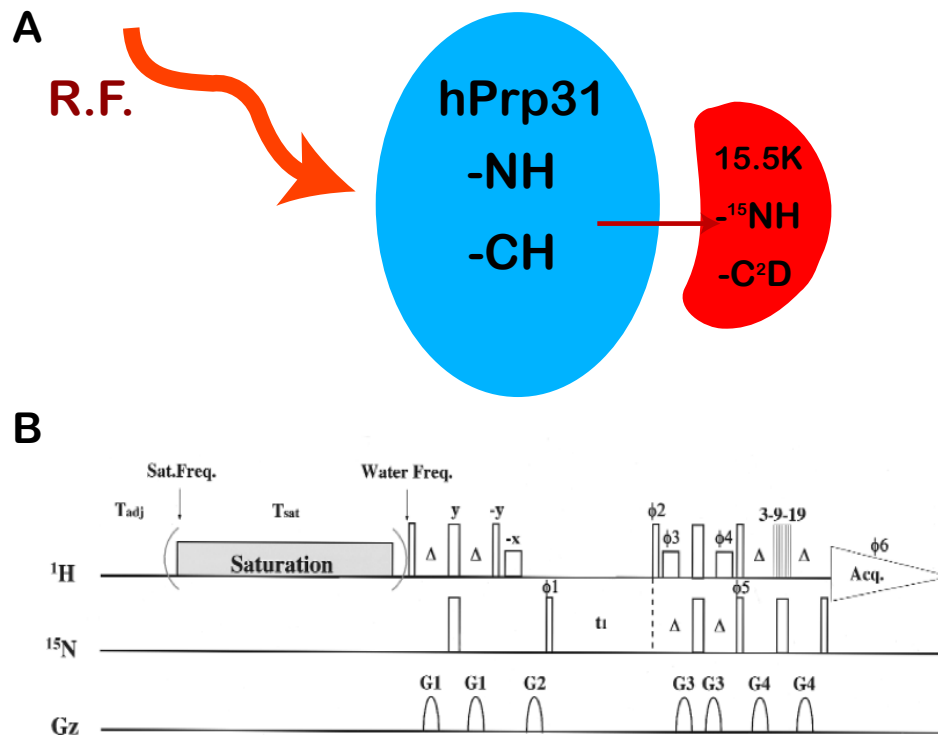
**Figure 3.12:** (A) Schematic representations of the cross-saturation effect. The aliphatic protons of hPrp31 are irradiated with a band-selective R.F. pulse and the saturation effect is transferred at the interaction surface to the fully deuterated 15.5K through cross relaxation. hPrp31 is unlabelled and 15.5K is fully deuterated and $^{15}$N labelled. The amide protons of 15.5K are exchanged to $^{1}$H in water (picture modified from Figure.1 in (160). (B) Pulse scheme for the cross-saturation method (picture taken from Figure.2 in (160).
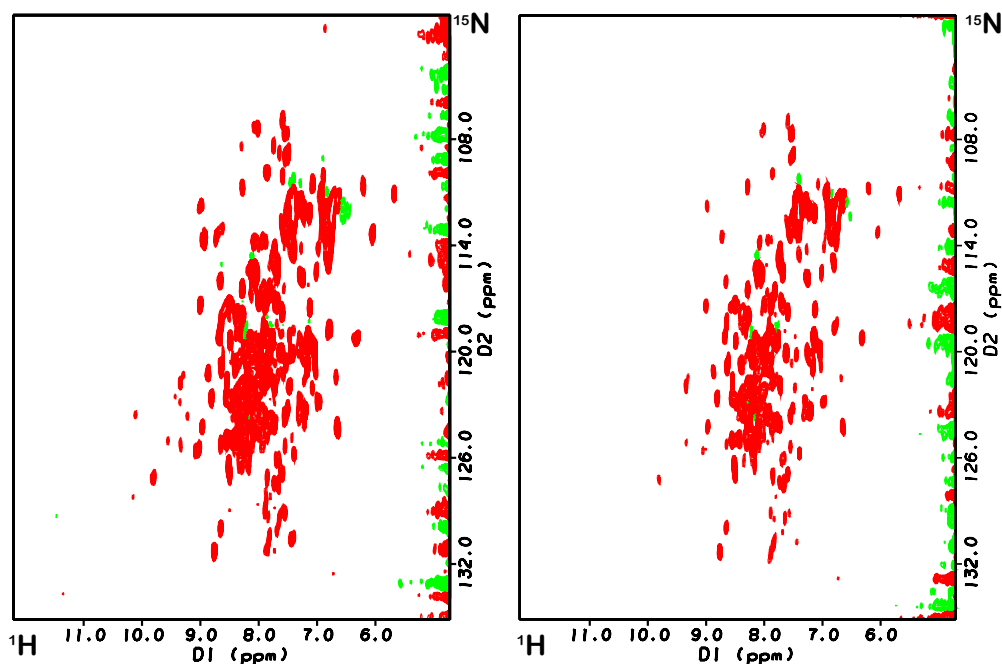
**Figure 3.13:** (A) The reference spectrum without cross-saturation (B) The spectrum with cross-saturation effect. These spectra are from the experiment on the hPrp31-15.5K-U4 5'-SL ternary complex.

no longer be observed (160). However, saturation of the aliphatic protons of the hPrp31 was performed using the WURST-2 decoupling scheme centred at -0.1 ppm with the excitation bandwidth of $\pm$ 3ppm. The maximum RF amplitude was 6 kHz with a pulse length of 40 ms and a saturation time of 800 ms. The experiments were recorded in 2-interleaved manner, with one reference experiment without saturation and a second experiment with the use of the decoupling scheme recorded alternately in one measurement. The experiments were processed in XWINNMR (Bruker) and the resulting spectra were analyzed in Felix (Accelrys) (Figure 3.13).

Peaks were picked in the reference spectrum (unsaturated). The picked

peaks were indicated as boxes and the box size i.e. peak area were manually adjusted to cover two-thirds of the peak widths in both $^1$H and $^{15}$N dimensions. The peak centres and peak areas from the reference spectrum were transferred onto the spectrum with saturation effect (saturated). The signal intensities were calculated for both spectra by Felix using 'peak integration' option as peak volume (I) included in the peak area. Changes in signal intensity (I) were calculated by taking the ratio between the intensity of the peak in the reference spectrum ($I_{unsat}$) and that of the corresponding peak in the spectrum with saturation ($I_{sat}$), shown as

$$I_{change} = \frac{I_{unsat}}{I_{sat}}. \tag{3.1}$$

Errors in peak volumes were also analyzed. The peak areas of the picked peaks in both spectra were divided into small, medium and large categories. Boxes with areas ($Area_{noise}$) adjusted to the average size of each category were then placed at different regions of the spectra without peaks to extract the noise. There were 7 boxes of the same size for each size category placed at difference positions in the background of the spectrum. The RMSD value of the noise intensity (volume) for each size category was calculated as well as the average peak area, which is the ratio between the $Area_{peak}$ and the peak area ($Area_{noise}$) of that size category. The noise (N) of each peak in both spectra was calculated as

$$N = RMSD_{noise} \cdot \frac{Area_{peak}}{Area_{noise}}. \tag{3.2}$$

Finally, the error ($\delta$) in signal intensity change for both spectra was then

calculated as

$$\delta = \left( \left( \left( \frac{N_{unsat}}{I_{unsat}} \right)^2 + \left( \frac{N_{sat}}{I_{sat}} \right)^2 \right) \right)^{\frac{1}{2}} \cdot I_{change}. \tag{3.3}$$

Considering that spin diffusion could still be mediated by 40% of water in the solvent and through the protonated U4 5'-SL RNA as well as residual non-exchangeable protons of the 15.5K side-chain, a control cross-saturation experiment was carried out on [$^2$H,$^{15}$N]15.5K-U4 5'-SL in the absence of hPrp31 or hPrp31$^{78-333}$. The signal intensity changes observed in the control experiment were then multiplied by 3.35 and 1.73 for comparison with the intensity changes observed in hPrp31-15.5K-U4 5'-SL and hPrp31$^{78-333}$-15.5K-U4 5'-SL complexes respectively. The factors account for the size ratios between the ternary complexes and the 15.5K-U4 5'-SL complex, which result different relaxation behaviors and different rates of spin diffusion in the complexes. The weighted intensity changes from the control experiment were subtracted from the intensity changes for each of the ternary complexes. Errors were also scaled accordingly. The resulting intensity changes for both ternary complexes were then presented as histograms (Figure 3.15).

Large intensity changes were observed on helix $\alpha$2 of 15.5K in both ternary complexes, with most pronounced changes from R36, A39, A42 and K44 of 15.5K in the hPrp31-15.5K-U4 5'-SL complex and K44 of 15.5K in the hPrp31$^{78-333}$-15.5K-U4 5'-SL complex. Residues on the 3-10 helix and helix $\alpha$3 of 15.5K also exhibited significant intensity changes. These are L63, L71 and L72 in the case of the hPrp31-15.5K-U4 5'-SL complex and L63, I66, I67, L71 and L72 for the hPrp31$^{78-333}$-15.5K-U4 5'-SL complex. V95 and I102, located on the loop region before and on $\beta$4 of 15.5K also showed
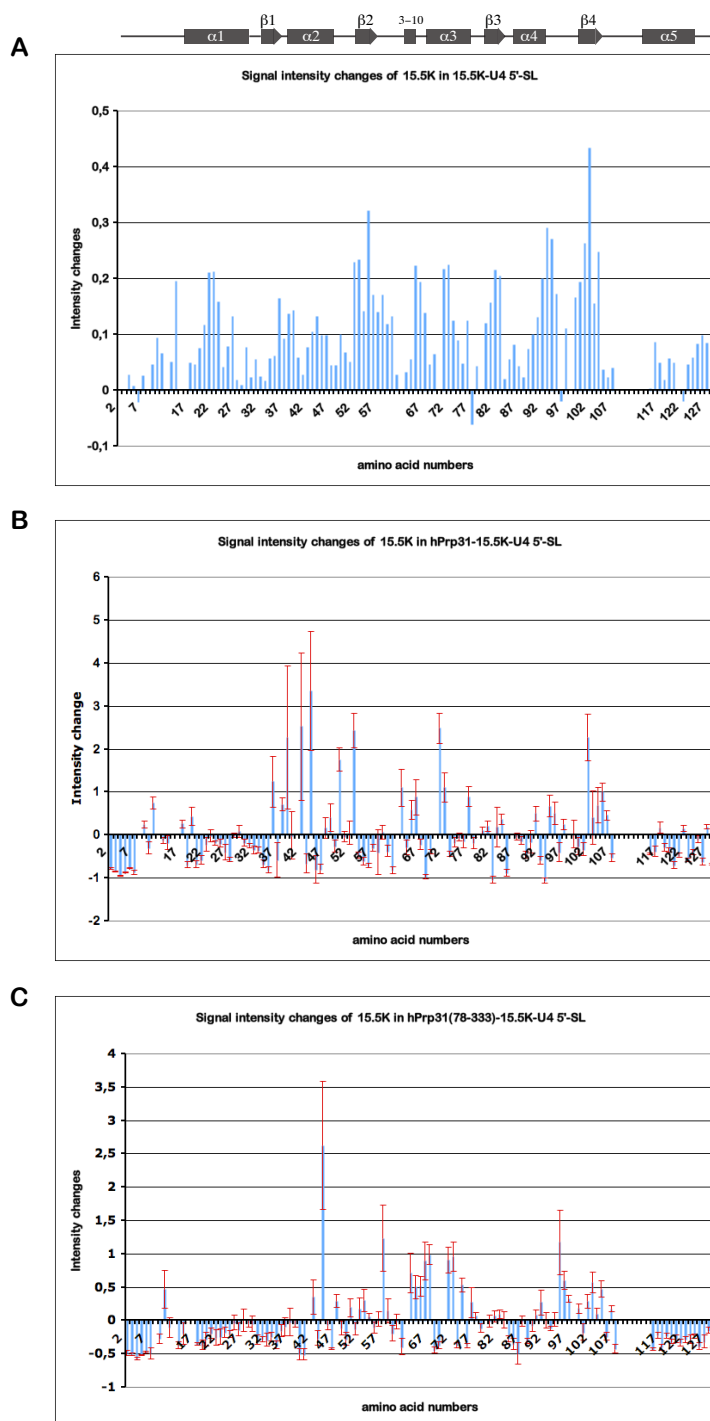
**Figure 3.14:** The intensity changes observed in the control experiment on the primary RNP and the cross-saturation experiments on the ternary complexes are plotted as histograms. The analyzed errors in the experiments on the ternary complexes are indicated as red error bars. Due to high quality of the control experiment, no error analysis was carried out. (A) 15.5K-U4 5'-SL; (B) hPrp31-15.5K-U4 5'-SL; (C) hPrp31$^{78-333}$-15.5K-U4 5'-SL
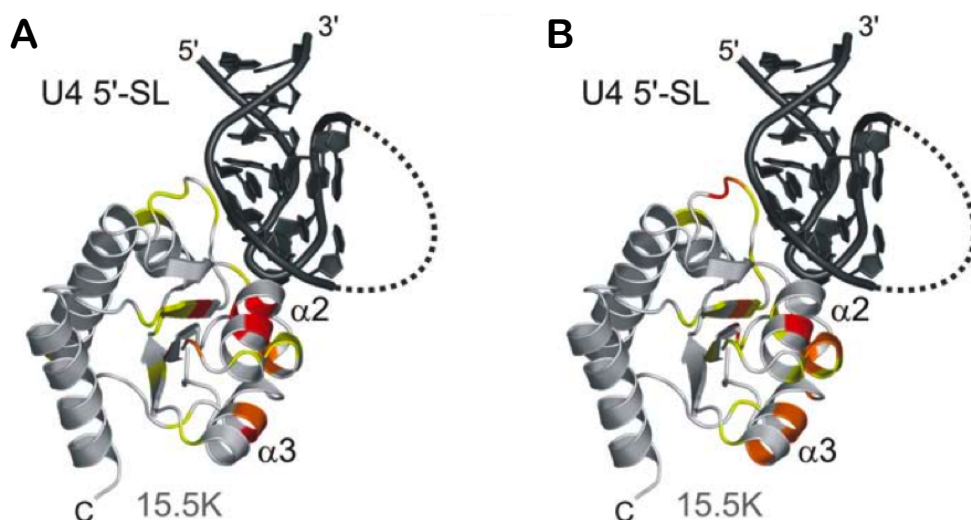
**Figure 3.15:** Signal intensity changes plotted as the colour codes on the crystal structure of 15.5K-U4 5'-SL. Red indicates a change larger than the value of 2; Orange indicates a value between 1 and 2 and yellow, a value between 0.5 and 1. (A) hPrp31-15.5K-U4 5'-SL; (B) hPrp31$^{78-333}$-15.5K-U4 5'-SL

considerable changes. However, they did not show a large degree of change in the chemical shift analysis. Significant changes were also observed for I50, F53 and A57 on $\beta$2 of 15.5K which are not located on the surface of 15.5K. The degree of intensity change observed for both ternary complexes is colour-coded on the 15.5K crystal structure (Figure 3.15) and the trend matches that observed from chemical shift analysis namely large effects are primarily located on helices $\alpha$2 and $\alpha$3 of 15.5K protein.

With the information from chemical shift perturbations and cross-saturation experiments, a 3D model for the ternary complex was created. The only readily available structure was the crystal structure of the 15.5K-U4 5'-SL-22nt complex. Therefore, a 3D model of the hPrp31 protein needed to be gener-

ated.  Moreover, the 22nt construct of the U4 5'-SL fails to allow complex formation with hPrp31 or hPrp31$^{78-333}$.  Therefore, a model of the 33nt RNA construct needed to be created.  Firstly, a model of residue 188-331 of human hPrp31 protein was created using comparative modelling.  Based on the crystal structure of the 15.5K-U4 5'-SL-22nt complex, a model for 15.5K-U4 5'-SL-33nt complex was produced.  Using the information from chemical shift perturbation, cross-saturation experiments and biochemical findings, the 3D model of hPrp31$^{188-331}$ was then docked onto the 15.5K-U4 5'-SL-33nt complex model with HADDOCK2.0.

# 3.3   Modelling and docking of hPrp31$^{188-332}$-15.5K-U4 snRNA complex

## 3.3.1   Comparative modelling of hPrp31$^{188-332}$ from its archaea homolog Nop5p

**Introduction to comparative modelling**

The aim of comparative protein structure modelling is to build a 3D model for a protein of unknown structure (the target) on the basis of the known structures of homologous proteins (the templates) with considerable amount of sequence similarity (162; 163; 164).  It usually requires a minimum sequence identity of 30% between the target and the templates to ensure sufficient accuracy in sequence alignment.  Although significant progresses have been made in *ab initio* protein structure prediction (165), comparative modelling is able to predict structures much more accurately.  3D structures of a

family of proteins are more conserved than their sequences (166). Furthermore, proteins with non-detectable sequence similarity can still adopt similar structures. With the exponentially increasing number of new structures being solved experimentally and the limited unique structural folds in proteins, comparative modelling plays an increasingly important role.

All current comparative modelling methods consist of four sequential steps: search for homologs using BLAST methods (167) and templates selection; template-target alignment; model building and model evaluation. This procedure is repeated until a satisfactory model is obtained.

## Crystal structure of *Archaeoglobus fulgidus* AF Nop5p and comparative modelling of hPrp31$^{188-332}$

In 2003 the structure of the archaea *Archaeoglobus fulgidus* (AF) homolog of Nop56/Nop58 namely Nop5p in complex with fibrillarin and the methyl donor S-adenosyl-$_L$-methionine (AdoMet) was solved in the group of Prof. Hong Li (100). The Nop5p and fibrillarin are specific proteins of the archaeal box C/D snoRNP in which L7ae, the archaeal homolog of 15.5K, plays the role as the nucleation factor for RNP assembly. The box C/D snoRNPs are known to carry out the task of 2'-O-methylation of pre-rRNAs with fibrillarin as the putative methyltransferase (Figure 3.16). The crystal structure of Nop5p consists of 2 two $\beta$ sheets at the beginning of the N-terminus followed by 12 $\alpha$ helices, which was an unique structural fold.

The crystal structure of the AF Nop5p provided us with the only available template for comparative modelling of human hPrp31 Nop domain. Firstly, sequence alignment of human hPrp31 and AF Nop5p was carried
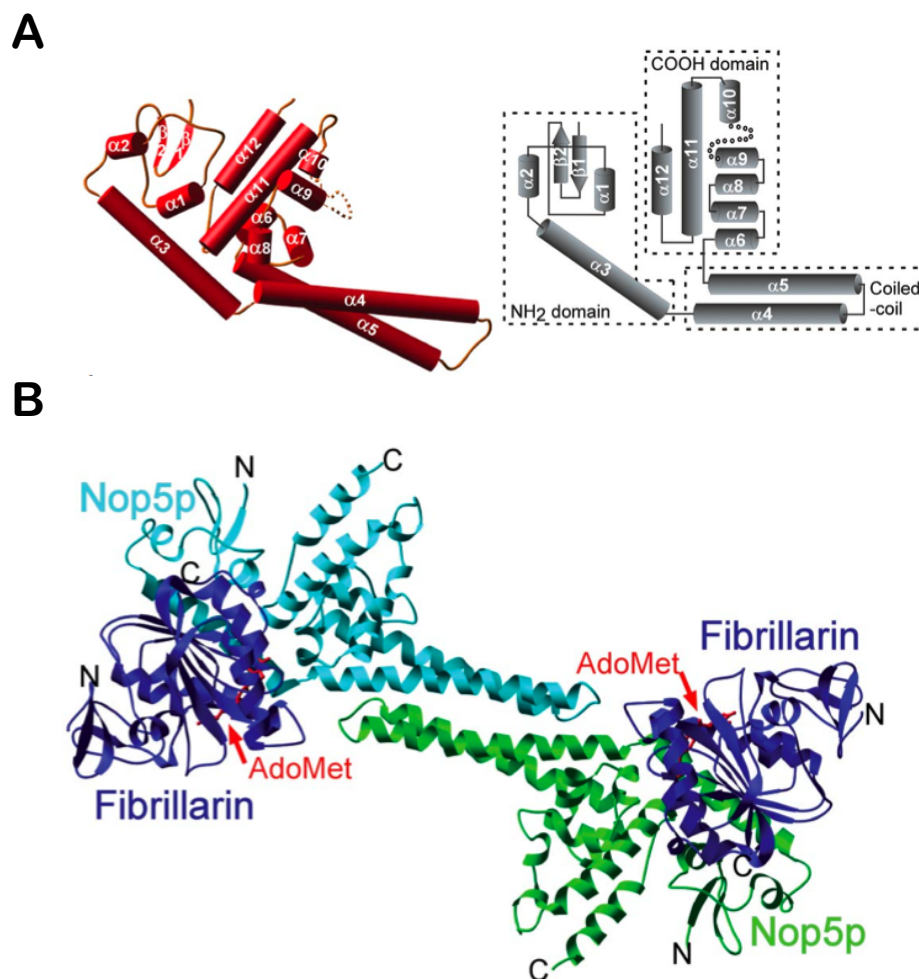
**Figure 3.16:** (A) The crystal structure of Nop5p coloured in red and represented as cylindar. A schematic representation of the secondary structure of the protein on the right. (B) The two homodimer of the Nop5p-fibrillarin heterodimer. AdoMet is presented in red (figure modified from Figure.1 in (100)).

out using Clustal method. Amino acids 188-334 in hPrp31 and 111-258 in Nop5p showed strong alignment, except for the stretch of 5 amino acids, which located on the unstructured loop region (Figure 3.17). The two sequences showed 28.0% sequence identity, which is at the minimal sequence identity required for comparative modelling. Therefore, secondary structures of hPrp31$^{188-332}$ and Nop5p$^{111-258}$ were compared before carrying out comparative modelling. Secondary structure prediction on both proteins was performed using PredictProtein server (http://www.predictprotein.org/) (145) (Figure 3.18). In this case sequences of both fragments were supplied to the server as single letter codes. PredictProtein server uses over 20 different methods for database search, sequence alignment as well as domain, structure and function prediction. The output was sent to the user in html format. Results of the secondary structural prediction for each protein fragment showed strikingly similar structural motifs. hPrp31$^{188-332}$ was predicted to contain $\alpha$ helices, which reside in regions that match the ones in the Nop5p$^{111-258}$ protein. The loop region, which was not crystallized in Nop5p, include residues 179-198 (KSLYKAFARMKKGKKAKIPK) and was aligned to residues 256-270 (RKTLSGFSSTSVLPH) in hPrp31$^{188-332}$. For this region in both protein fragments neither helix nor strand was predicted.

As the secondary structures of these two proteins showed high similarity, comparative modelling was carried out with confidence. The 'first approach mode' provided by SWISS-MODEL server (http://swissmodel.expasy.org//SWISS-MODEL.html) was used to perform this task (146; 147). In this mode, the amino acid sequence of hPrp31$^{188-332}$ was provided and the PDB ID of AF Nop5p, namely, 1nt2B was given to specify the template. The SWISS-

```
hPrp31     78   SEVMGPV.EAAPEYRVIVDANNLTVEIENELNIIHKFIRDKYSKRFPELESLVPNALDYI
afNop5p    73   ...........EDRYVVALVKALEEIDESINMLNEKLEDIRAVKESEIT..........
```
                                                     α4

```
hPrp31    137   RTVKELGN..SLDKCKNNENLQQILTNATIMVVSVTASTTQGQQLSEEELERLDEACDMA
afNop5p   111   ..............................................EKFEKKI
```

```
hPrp31    195   LELNASKHRIYEYVESRMSFIAPNLSIIIGASTAAKIMGVAGGLTNLSKMPACNIMLLGA
afNop5p   118   RELRELRRDVEREIEEVMEKIAPNMTELVGAKVAAKLLERAGSMERLVRLPASKIQVIGA
```
                         α5                α6            α7           α8        α9

```
hPrp31    255   QRKTLSGF.....SSTSVLPHTGYIYHSDIVQSLPPDLRRKAARLVAAKCTLAARVDSFH
afNop5p   178   EKSLYKAFARMKKGKKAKIPKHGIIFLHPFIRTLPKAKRGKMARFLAAKLAIAAKIDYFR
```
                                             α10                      α11

```
hPrp31    310   ESTEGKVGYELKDEIERKFDKWQE
afNop5p   238   GEIDESLYESIRRRYEELRRK...
```
                         α12

**Figure 3.17:**  Sequence alignment of human Prp31 and the *Archaeglobus fulgidus* AF Nop5p over the region 78-333 of hPrp31. The secondary structural motif of the Nop5p in this region is shown below the sequences. The sequence corresponding to the uncrystallized loop is shown in the red box. Residue H270, which was shown to directly contact the RNA, is indicated with a red triangle on the top.

**(A)**  AF Nop5p$^{111-258}$



**(B)**  hPrp31$^{187-334}$



**Figure 3.18:** (A) The secondary structural prediction results carried out using the PredictProtein server (http://www.predictprotein.org/). The histogram in red indicates the possibility for assigning helix (1=high, 0=low); in blue, the probability for assigning strand (1=high, 0=low); in green, the probability for assigning neither helix nor strand (1=high, 0=low) and in grey, the predicted relative solvent accessibility. (A) Nop5p$^{111-258}$ and (B) hPrp31$^{187-334}$

**Figure 3.19:** The comparative model of hPrp31$^{188-332}$ obtained from the SWISS-MODEL server is colour in green. Three $\beta$ sheets (red) were predicted for the region corresponding to the uncrystallized loop in AF Nop5p. The helices are numbered as the corresponding helices in AF Nop5p.

MODEL server performs four modelling steps: template superposition (not needed in this case, as only one template was given); target template alignment; model building and energy minimization. The model building step include backbone modelling, modelling for insertion or deletions in the target-template alignment using constraint space programming (CSP) as well as side chain modelling by iso-sterically replacing template structure side chains. As the result, a 3D model of hPrp31$^{188-332}$ was generated (Figure 3.19).

The model consisted of 9 helices in the regions predicted for helix motif by PredictProtein server. To allow comparison, the helices in hPrp31$^{188-332}$ model are numbered according to those of the corresponding helices in Nop5p. The helices corresponding to helices $\alpha8$, $\alpha9$, $\alpha11$ and $\alpha12$ of the Nop domain

in Nop5p form a roughly flat surface. For residues 253-272 (GAQRKTLSGF-SSTSVLPHTG), which reside in the unstructured loop region, a stretch of 3 $\beta$ sheets were modeled (coloured red in Figure 3.19), which is unrealistic in this case. Moreover, in the secondary structure prediction, residues 253-262 (GAQRKTLSGF) also showed high probability to form $\alpha$ helix. Although absent in the crystal structure, H270 in this loop region, which is highly conserved in different hPrp31 orthologs as well as in its homologs Nop56 and Nop58, was found to form direct contact with U44 of the U4atac snRNA in UV-crosslinking study and this cross-link was identified and confirmed by mass spectrometry (95; 136). Therefore, remodelling of this loop region is required for docking.

Energy minimization and remodelling of the unstructured region of hPrp31$^{188-332}$ were carried out using Xplor-NIH (148; 168). Xplor-NIH is a program suite for structural calculations. First of all, information on the molecular structure and energy terms during molecular dynamics (MD) need to be defined. The topology files define atoms (mass, charge), assignments of covalent bonds, bond angles and improper terms, and parameter file contain various force constants for bonds, angles, impropers, dihedrals and non-bonded terms. These two files are used in conjunction to produce the molecular structure files marked with the file extension '.psf'. Standard topology and parameter files are provided by Xplor-NIH suite. In this case, topall-hdg5.3.pro (topology file) and parallhdg5.3.pro (parameter file) were used for defining the terms mentioned above in the protein. The PDB file generated by SWISS-MODEL was used as the input coordinate file. In this PDB file, all hydrogen atoms are absent. The standard generation protocol from Xplor

was used for generating a initial PDB (coordinate file) of hPrp31$^{188-332}$ with hydrogen atoms and the psf (structure file). The initial PDB was also minimized in this protocol with 70 steps of Powell minimization. This minimized initial PDB was then subjected to a sequence of simulated annealing (SA) steps. The SA protocol consists of three steps, a heating stage from 100 to 2000 K in 200 steps with 1 fs time-step and refinement stage (hot step) for 3000 steps at 2000K with 2 fs time-steps and a cooling stage from 2000K to 25K in 200 steps. The protein was fixed, while residues 253-272 were left free with $\alpha$ helical dihedral restraints for residues 250-262. The output structure was subsequently used as input structure in the water refinement protocol. In this water refinement protocol, a shell of water molecules is simulated around the protein structure and simulated annealing is carried out consisting of three stages, a heating stage from 100 to 500K in 10 steps with 3 fs time step and short refinement stage (hot step) for 100 steps at 500K with 4 fs time-steps and a cooling stage from 500K to 25K in 10 steps. The resulting water refined structure was superimposed to that of AF Nop5p$^{111-258}$ in structured regions with an RMSD of 0.4 Å (Figure 3.20).

### Generation of 15.5K-U4 5'-SL-33nt model

As the U4 5'-SL-33nt construct contains the minimal required length of stem I, a 3D model of 15.5K-U4 5'-SL-33nt complex was created based on the crystal structure of the 15.5K-U4 5'-SL-22nt complex. Moreover, the penta-loop is not present in the crystal structure but has been shown to form direct contacts with hPrp31 protein. As mentioned above, H270 of hPrp31 was identified to form direct contact with U44 in the penta-loop of U4atac snRNA. U44
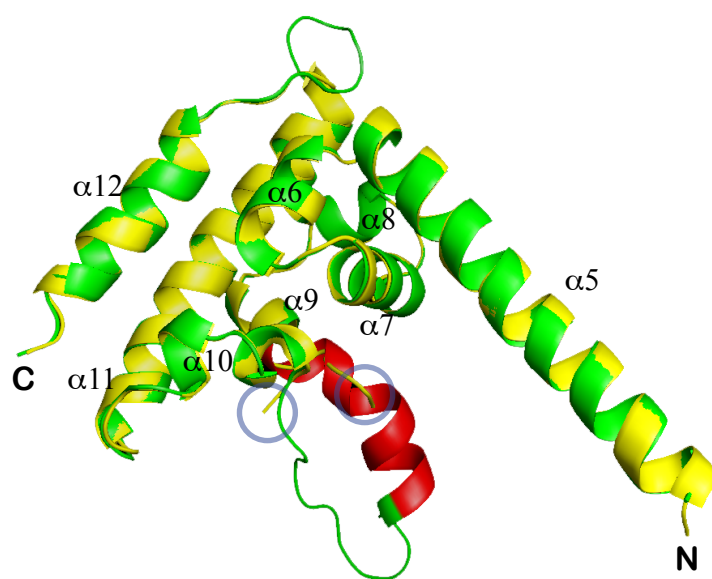
**Figure 3.20:** The water refined model of hPrp31$^{188-332}$ (in green) is overlaid on the crystal structure of AF Nop5p (residues 111-258, in yellow). The elongated helix $\alpha$9, which is absent in Nop5p, is shown in red. The breaks in the crystal structure of Nop5p at this region are shown in grey circles.

in U4atac corresponds to A39 in the penta-loop of the U4 snRNA. There-
fore, to be able to include this important piece of biological information in
the docking, the presence of the penta-loop is needed in the model. The
biopolymer module of InsightII (Accelrys) program was used to generate the
missing parts of the RNA. Nucleotides of the penta-loop (nucleotides 36-40)
and of the missing part of stem I, namely nucleotides 20-25 and nucleotides
48-52, were appended using 'A_RNA_1_Strand' in the biopolymer module as
three separate A-form strands and the structures were saved as three sepa-
rate PDB files. The crystal structure of 15.5K-U4 5'-SL-22nt complex was
then displayed in a new session and the three PDBs for the missing parts
were also displayed. The three pieces were assembled to the appropriate
positions on the 22nt RNA manually. The whole ensemble was then saved
together as one PDB file. Hydrogen atoms in the PDB file were removed to
ensure compatibility of the file to CNS system, which is used by the dock-
ing program. Psf files were generated separately for 15.5K protein and the
U4 5'-SL-33nt. The 'RNA_DNA_amber' topology and parameter files down-
loaded from the Moore Lab Homepage (http://proton.chem.yale.edu/) was
used for generation of the psf file for the RNA. In the psf generating pro-
cedure, hydrogen atoms were added to the structures. PDB files containing
hydrogens for 15.5K and the RNA were also generated (Figure 3.21). The
penta-loop was allowed to adopt random structures through three rounds of
verlet dynamics with 3000 steps and 1 fs time-step at 1000K, 600K and 400K
with the loop randomization protocol. For connecting the penta-loop to the
stem II, random NOEs were given between G35 and U36, as well as U40 and
C41. The rest of the RNA was fixed in this protocol and planarity restraints

were given to the nucleotides in the loop. Subsequently three rounds of rigid body dynamics with 3000 steps and 1 fs time-step at 1000K, 600K and 400K for the whole RNA. During this procedure, contacts between 15.5K and the RNA were fixed and the planarity, the sugar pucker and the hydrogen bonding distance of the RNA were restrained. As the result, 200 structures of 15.5K-U4 5'-SL-33nt complex with randomized penta-loop were generated. In these structures, the stem I of the RNA was not yet properly formed. The whole complex structure was then subjected to water refinement with 500 steps and 2 fs time-step at 2000K. The complete stem I was restrained to A-form with using dihedral restraints for standard A-form RNA helix. The planarity restraints for the base-pairs were given a weight of 1000.0 to ensure that the bases of the RNA were planar. Again 200 structures were generated after this water refinement protocol. In these structures, the elongated stem I showed A-form and the penta-loop adopts different structures (see Appendix).

**Docking contacts**

Before carrying out the docking task, the electrostatic potential of the surfaces of 15.5K-U4 5'-SL and hPrp31$^{188-332}$ were analyzed. The charge properties of the molecular surfaces provide crucial hints on the interaction surfaces of the molecules to be docked, as clearly, surfaces with complementary charges are favoured to find and interact with each other in solution. As little information was available on the interaction surface of hPrp31 in the ternary complex, this investigation provided us a crucial clue for selecting the right surface on hPrp31$^{188-332}$ for docking. Electrostatic surfaces of
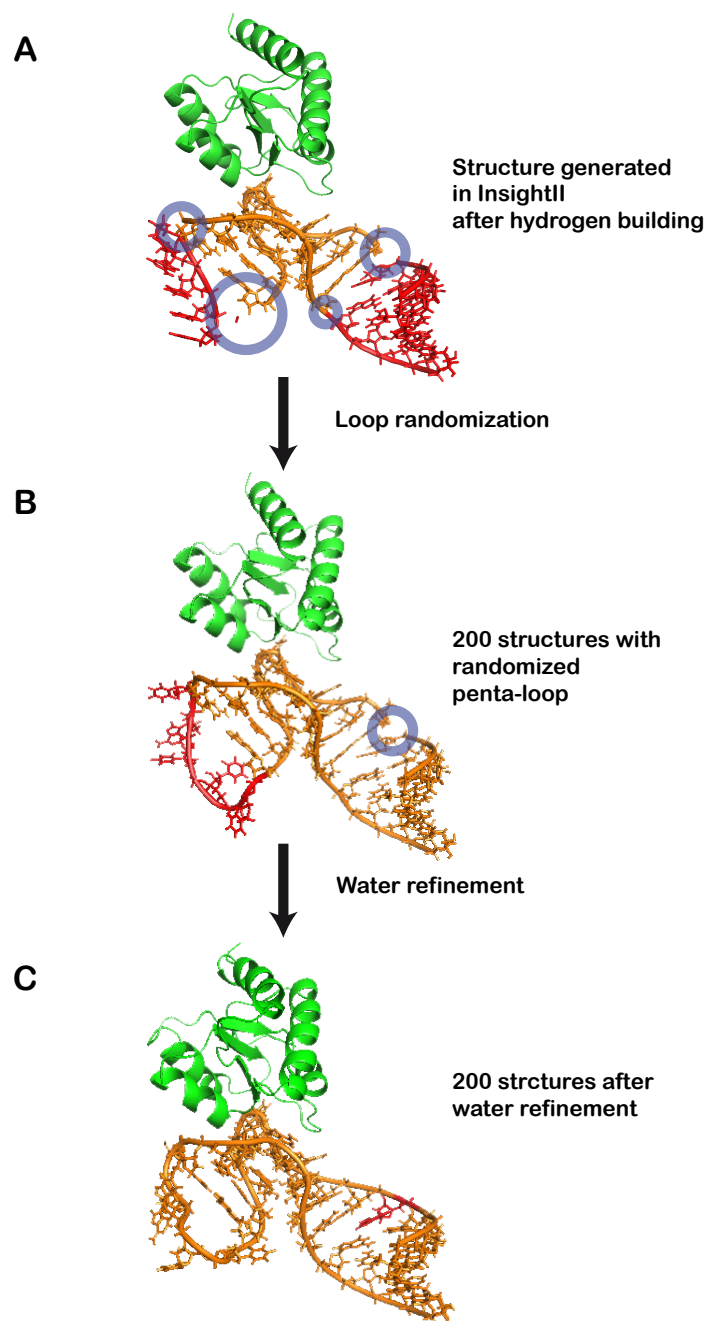
**Figure 3.21:** Models of 15.5K-U4 5'-SL-33nt complex were generated in three steps: (A) Build the elongated stem I and the penta-loop in the U4 5'-SL-33nt using Insight II. The fragments generated by Insight II are shown in red. The connections of these fragments to the crystal structure of the U4 5'-SL-22nt are shown in circles. (B) Randomization of the penta-loop. Penta-loop is coloured in red and the break in the backbone of stem I is shown in the circle. (C) One of the 200 structure generated after the water refinement. A27 (red) in the RNA, which is likely to be bulged out, is stacked in due to the A-form helix restraints on neighbouring nucleotides.

15.5K-U4 5'-SL-33nt and hPrp31$^{188-332}$ were calculated from the PDB files using the PDB2PQR server (http://pdb2pqr.sourceforge.net/) (144). Amber force field was chosen and the results were given as PQR files in which the occupancy column of a PDB file was replaced with atomic charge and the temperature column with the radius. The surfaces were then displayed by Pymol using the APBS function (169). The electrostatic potential calculation for hPrp31$^{188-332}$ resulted a strongly electropositive surface and a relatively more electronegative surface (Figure 3.22 A). A large portion of the electropositive surface is the roughly flat surface formed by helices corresponding to helices $\alpha 8$, $\alpha 9$, $\alpha 11$ and $\alpha 12$ of the Nop domain in Nop5p. Two residues located on this surface, namely C247 and R293, could form direct contact with the 15.5K-U4 5'-SL RNP. C247 of hPrp31 was identified by mass spectrometry to be cross-linked to the snRNA (data not published). Mutagenesis studies carried out on fibrillarin-Nop5p complex showed that R224 of Nop5p, which corresponds to R293 of hPrp31, was essential for the association of the protein to L7ae-box C/D snoRNA complex (100). It can be deduced that R293 of hPrp31 would be similarly involved in the interaction with 15.5K-U4 5'-SL-33nt complex. A third piece of contact information came from the confirmed UV-crosslink between H270 of hPrp31 and U44 of U4$_{atac}$ 5'-SL, which corresponds to A39 in U4 5'-SL (136). H270 is not strictly located on the same surface as C247 and R293, but on the loop which is undefined in the crystal structure of Nop5p. It can, therefore, orient towards either surface. The electrostatic surface of 15.5K-U4 5'-SL-33nt complex, showed a significant electronegative flat surface formed mainly by the stem II and the penta-loop of the U4 snRNA as well as the N-terminal part of the

helix $\alpha2$ of 15.5K (Figure 3.22 C). Helix $\alpha3$ of 15.5K, which shows mainly neutral charges, is directly above the helix $\alpha2$ and is, therefore, also located on this electronegative surface. It was a very obvious choice for us to pick the strongly positively charged surface on hPrp31 and the overall negatively charged surfaced provided by the stem II and the penta-loop of the RNA and the interaction surface defined on 15.5K as the surfaces for docking.

The relative orientation of the two charge complementary surfaces on hPrp31[188−332] and 15.5K then had to be defined. A rough measure of the distance between H270 and R293 of hPrp31 showed that these two residues are about 15 Å apart in space. The distance between A39 of U4 snRNA to the bulged out U31 at the peak of the kink-turn also gave a rough measure of 15 Å. Therefore, it is reasonable to propose that with H270 contacting A39, R293 will be contacting the stem II of the RNA and as the side chain of an arginine is long, it could also form contacts with the helix $\alpha2$ of 15.5K. If the third anchoring point, C247, which contacts the RNA is also considered, from geometric compatibility, R293 could only contact the stem II of RNA. Consequently two docking runs were performed. In one, the information that C247 cross-links to RNA was not used, thus, C247 was assigned to contact only the helix $\alpha2$ of the 15.5K protein and R293 to both the 15.5K and the stem II of the RNA. In the other run, the cross-linking information was considered and therefore C247 was given contacts to both the helix $\alpha2$ of 15.5K protein and the stem II of the RNA whereas R293 was only assigned to the stem II of the RNA. Using the information that helix $\alpha2$ and $\alpha3$ of 15.5K contact hPrp31, other surface residues located on the interface of the hPrp31 could also be assigned under geometry considerations. Taking all the above into
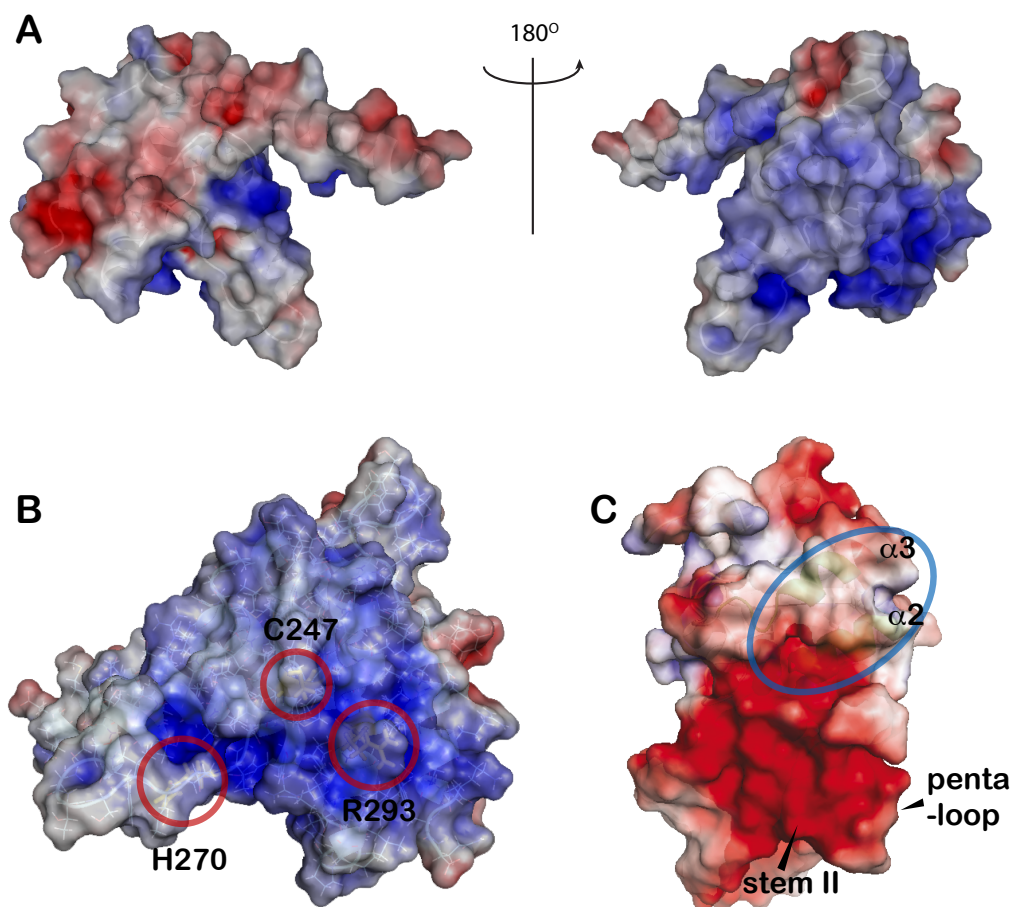
**Figure 3.22:** (A) hPrp31$^{188-332}$ posses a strongly electropositive surface (right) and a relatively more electronegative surface (left). (B)The three residues of hPrp31 studied by biochemical means namely C247, H270 and R293. C247 and R293 both locate on the electropositive surface. H270 locate on the loop region before helix $\alpha$10. The three residues are indicated in red circles. (C) The electrostatic surface of 15.5K-U4 5'-SL. Helices $\alpha$2 and $\alpha$3 are shown in the circle. The stem II and the penta-loop of the RNA contribute to a highly electronegative surface.

consideration, the interaction surface on hPrp31$^{188-332}$ was proposed to be primarily located on helices corresponding to helices $\alpha 8$, $\alpha 9$ and $\alpha 11$ and was divided into three zones. N240, V305 and F308 on the very N-terminal end of helix $\alpha 8$ and the very C-terminal end of $\alpha 11$ were only given contacts to N-terminal end of helix $\alpha 3$ of 15.5K (upper zone, shown in red in Figure 3.23 A), whereas residues 245-248 on helix $\alpha 9$, R293, L294, A297 and K298 on N-terminal region of helix $\alpha 11$ of hPrp31$^{188-332}$ were only given contacts to $\alpha 2$ of 15.5K (lower zone, shown in yellow in Figure 3.23 A). Residues K243 on the C-terminal end of the $\alpha 8$ and 299-304 on the C-terminal end of $\alpha 11$ of hPrp31 reside the middle region of the interaction surface (shown in orange in in Figure 3.23 A) and were given contacts to both helices $\alpha 2$ and $\alpha 3$ of 15.5K (Figure 3.23). Residues on helices $\alpha 2$ and $\alpha 3$ of 15.5K exhibited large chemical shift changes upon hPrp31 titration and significant intensity changes in the cross-saturation experiments. Therefore, the interaction surface on 15.5K primarily locates on helices $\alpha 2$ and $\alpha 3$. The contacting residues on 15.5K were chosen to be the surface residues namely N40, N47 and R48 of helix $\alpha 2$ (shown in yellow in Figure 3.23 B) and I56, I66, H68 and L71 of helix $\alpha 3$ (shown in red in Figure 3.23 B). These chosen residues are not identical to those showing largest intensity changes in the cross-saturation experiments namely R36, A39, A42, K44, L63, I66, I67, L71 and L72 (Figure 3.15), because the cross-saturation effect acts on backbone H$_N$, while the docking restraints are on the side chains. Therefore, it is very important that the side chains of the selected residues are solvent exposed. Moreover, several of these chosen surface residues could not be analyzed in the cross-saturation experiments, because the resonances of these residues shifted too
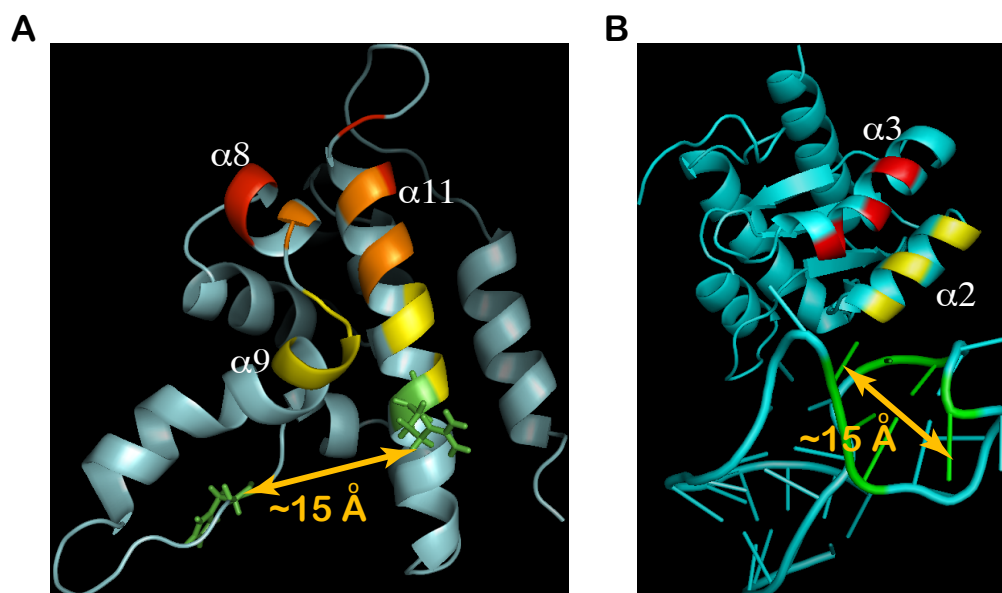
**Figure 3.23:** (A)The interaction surface defined on hPrp31$^{188-332}$ is shown in three coloured indicating the three contacting zones namely the upper zone (red), the middle zone(orange) and the lower zone (yellow). C247 and H270 are shown in green and the distance between these two residues is roughly 15Å. The rest of hPrp31$^{188-332}$ is in grey. (B) The interaction surface defined on 15.5K-U4 5'-SL is shown. The contacts on helix $\alpha$2 are indicated in yellow and those on helix $\alpha$3 in red. The stem II of the U4 5'-SL is also defined as part of the interaction surface and is coloured in green. The distance from A39 in the penta-loop to the top of the K-turn is roughly about 15Å.

much, became too broadened or possibly overlapped with other peaks. Due to the very limited data from biochemical studies and the highly ambiguous information about the interaction surface, especially on hPrp31$^{188-332}$, HAD-DOCK2.0 (high ambiguity driven biomolecular docking version2.0) was used to generate a model for the hPrp31$^{188-332}$-15.5K-U4 5'-SL-33nt complex.

### 3.3.2   Generation of the 3D model of hPrp31$^{188-331}$-15.5K-U4 5'-SL using HADDOCK2.0

As mentioned in the Introduction chapter, solving the structure of large bio-molecular complexes is still a difficult challenge due to their intrinsic complexity. Docking methods, such as HADDOCK, are becoming increasingly important for helping to carry out this task. Docking is essentially the process of fitting two molecules together in 3D space. HADDOCK2.0 is the latest version of the structural calculation program suite developed in the lab of Dr. Alexandre Bonvin (170; 171; 172; 173; 174; 141). It utilizes biochemical data from mutagenesis and cross-linking experiments, or biophysical interaction data such as chemical shift and cross-saturation data or bioinformatic predictions as structural restraints in the docking simulation of up to six bio-molecules including proteins, RNAs and DNAs. The information on the interacting residues from experimental studies is introduced as AIRs (<u>a</u>mibiguous <u>i</u>nteraction <u>r</u>estraints) to drive the docking between the bio-molecules. This involves finding the correct orientation of the molecules with respect to each other in space and allowing the experimentally observed residue contacts to be satisfied in the complex model. An AIR is defined as an ambiguous intermolecular distance ($d_{iAB}$) with a maximum value of 2 Å between any atom $m$ of an interacting residue i of protein A ($m_iA$) and any atom $n$ of the contacting residues k ($N_{res}$ in total) of protein B ($n_{kB}$). The effective distance $d_{iAB}^{eff}$ is calculated as

$$d_{iAB}^{eff} = \left( \sum_{m_{iA}=1}^{N_{atom}} \sum_{k=1}^{N_{resB}} \sum_{n_{kB}=1}^{N_{atom}} \frac{1}{d_{m_{iA}n_{kB}}^6} \right)^{-\frac{1}{6}}, \qquad (3.4)$$

where N$_{atoms}$ indicates all atoms of a given residue and N$_{resB}$ is the sum of all contacting residues for a given molecule. This formula accounts only for the attractive part of the Lennard-Jones potential and ensures that the AIRs are satisfied as soon as any two atoms of the two residues defined to have an AIR, are in contact. The actual docking consists of three steps: randomization of orientations and rigid body energy minimization (EM), semi-flexible simulated annealing (SA) and final flexible explicit solvent refinement (water in this case). In the rigid body EM step, the molecules are separated by a minimum of 15 Å and rotated randomly around their centre of mass. No changes in the structures of the molecules are allowed (rigid body). Usually 1000 structures are generated from this first step, from which 200 structures are selected for the semi-flexible SA step. This step consists of several stages: high temperature rigid body search, rigid body SA, semi-flexible SA with flexibilities allowed in side-chains at the interface defined by the user and semi-flexible SA with full flexibilities (both backbone and side-chains) allowed at the interface. All of the 200 structures resulted from this step will be subjected to explicit solvent refinement. The 200 solvent refined structures are then ranked according to their overall scores, which are calculated as a weighted sum of different terms expressed as

$$Score_{rigid} = 0.01E_{vdW} + 1.0E_{elec} + 0.01E_{AIR} + (-0.01)BSA, \qquad (3.5)$$

for the rigid body stage,

$$Score_{Semi-flexible} = 1.0 E_{vdW} + 1.0 E_{elec} + 0.1 E_{AIR} + (-0.01) BSA, \quad (3.6)$$

for the semi-flexible refinement and solvent refinement stage (141). Here $vdW$ is the van der Waals energy; $elec$, electrostatic energy; $AIR$, ambiguous interaction restraints; $BSA$, buried surface area. The importance of $E_{vdW}$ and 0.1 $E_{AIR}$ are emphasized in the refinement stages to give better scores to the models, which have better non-bonding interactions and can better satisfy the AIRs.

The solvent refined structures can then be grouped with an appropriate cut-off (7.5 Å cut-off by default) into clusters according to their pairwise RMSD values. These values are calculated by superimposing the structures on the backbone atoms of the flexible interface on molecule A and the RMSD is calculated on the backbone atoms of the flexible interface on its interaction partner(s). The clusters are then ranked by their average HADDOCK scores. Clusters with better scores in different HADDOCK2.0 docking attempts have been shown to contain models which are closer to the experimentally determined complex structures (141; 173).

Two docking runs with the AIRs as described in the previous section were carried out (see Appendix). The 'master file' of HADDOCK is the 'run.cns' file in which one can define all the parameters needed for each step of calculation, scoring and clustering. A default run.cns file from HADDOCK2.0 was used with appropriate modifications for each run. For both runs, the water refined comparative model of hPrp31$^{188-332}$ was defined as the molecule A

and the water refined model of 15.5K-U4 5'-SL-33nt complex was defined as molecule B. The semi-flexible interface on hPrp31$^{188-332}$ was defined as from residue 270 to 250 ($\alpha$8 to $\alpha$9) and from residue 291 to 310 ($\alpha$11) and that on 15.5K-U4 5'-SL included residues 38-50 ($\alpha$2) and 61-74 (helices 3-10 and $\alpha$3) on 15.5K and residues 32-35 and 41-43 (stem II up to the K-turn) of the RNA. Fully flexible fragments, which were kept free in all docking stages, were defined to be the flexible loop region in hPrp31$^{188-332}$ (residues 252-270) and the mobile penta-loop of the U4 5'-SL (nucleotides 36-40). During the calculation, as molecule B contains the RNA, the 'dna-rna_restraints.def' file (see Appendix) was used to constrain the planarity, sugar pucker, backbone dihedral and base-pair NOEs. For the topology and parameter files, 'topallhdg5.3.pro' and 'rna-prot.par' respectively were used for molecule A, whereas 'rna-prot.top' and 'rna-prot.par' were used for molecule B. The docking protocol, which includes the rigid body EM, the semi-flexible SA and the water refinement was used as the default parameters. The weights for the calculation of HADDOCK2.0 scores were also taken as default, except that the weight of $E_{elec}$ in the water refinement step was decreased to 0.1 to further emphasize the energy contribution from non-electrostatic terms. Finally the water refined structures in both runs were clustered with a 4 Å RMSD cutoff and a minimum of 4 structures per cluster using the 'cluster_struc' program. 15 clusters were calculated for run1 and 16 for run2. The HADDOCK2.0 score for each cluster and each structure was reported.The best scoring structures in the best scoring clusters of both runs fitted to each other with an average backbone RMSD of 2.55 Å and the average backbone RMSDs for the structures in the best scoring clusters of run1 and run2 are

2.96 Å and 2.17 Å respectively (Figure 3.24).

## 3.4   Investigation of the mutant 15.5K-2

The 15.5K-2 mutant was created by Dr. Annemarie Schultz (135) and exhibited a 4-fold reduction in binding affinity to hPrp31 in pull-down assays (135). In this mutant, 4 mutations, namely E74R/D75K/K76M/N77T located at the C-terminus of the helix $\alpha 3$ of 15.5K, were introduced. These residues, however, did not show direct contacts with hPrp31 in cross-saturation experiments and are not at the protein-protein interface in our docking model. To test the possibility that the dysfunction of the 15.5k-2 mutant is caused by structure changes in 15.5K protein, residual dipolar coupling (RDC) refinement was carried out on the 15.5K-2 mutant with the 15.5K crystal structure as the template.

### 3.4.1   Introduction to RDCs and RDC refinement

In solid-sate NMR, spatially anisotropic magnetic interactions such as magnetic dipole-dipole interaction, the chemical shift anisotropy, or the electric quadrupole interaction are often utilized for obtaining structural information. However, in liquid-state NMR experiments on isotropic samples, these spatially anisotropic interactions are averaged out due to the rapid and isotropic reorientation of the molecules in an isotropic environment. This spatial information could be regained in liquid-state NMR by partially aligning the molecule in anisotropic solvents or to external fields (magnetic or electric) (175; 176; 177; 178). In an anisotropic medium, a very small fraction of
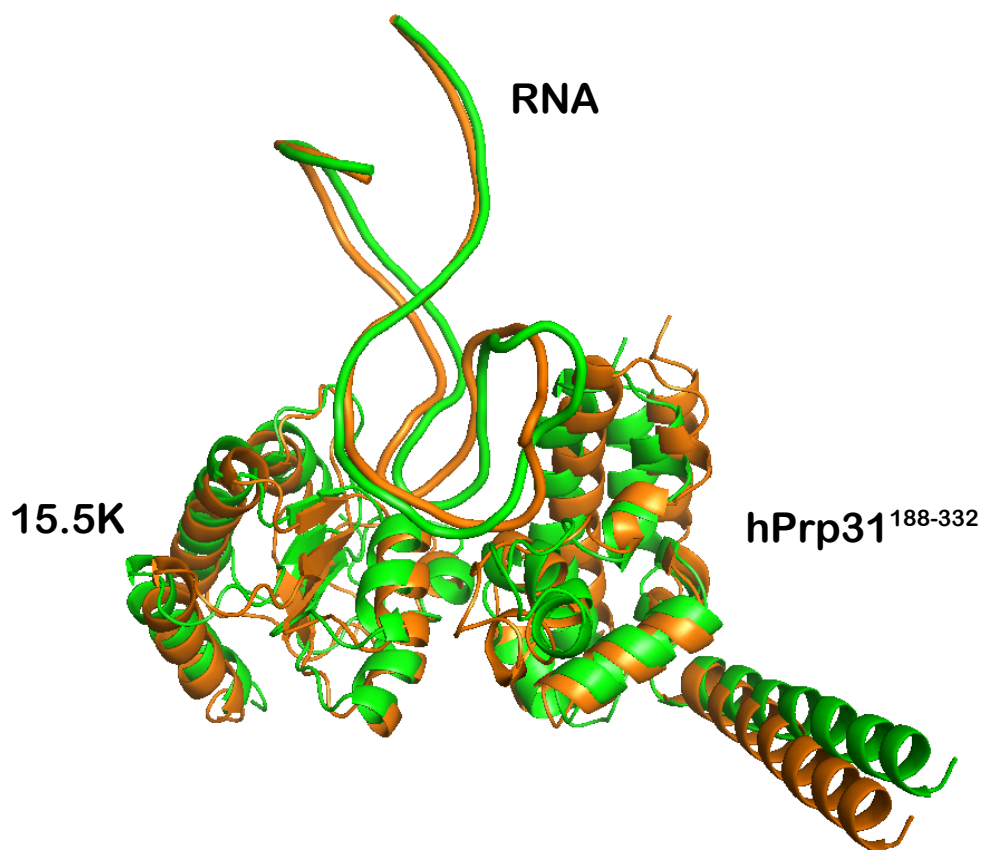
**Figure 3.24:** Superpostion of the best scoring docking models from two HADDOCK2.0 docking attempts. The best model of run1 is coloured in green and that of run2 is coloured in orange.

dissolved molecules adopt a preferred orientation through steric, anisotropic and electrostatic interactions with the alignment medium and dipolar couplings can be observed. Therefore, in an anisotropic sample, the measured $J_{total}$ splitting is

$$J_{total} = J_{scalar} + D, \tag{3.7}$$

where D is the RDC and is about 1/10000 of the size of the RDCs in solid state NMR. From the isotropic sample before adding alignment medium, $J_{scalar}$ couplings can be measured. By subtracting $J_{scalar}$ from $J_{total}$ in the anisotropic sample, the RDCs (D in equation 3.7), can be obtained. After their observation by Prestegard (179) and by Tjandra and Bax (180), RDCs have been playing an increasingly important role in liquid-state NMR as novel parameters for structural calculation. Different from the nuclear Overhauser effects (NOEs), which are effective in distances ranging from 1.8 Å to 5 Å, RDCs can provide long range and global structural information, e.g. relative orientations of helices in the protein and relative domain orientations. This is achieved because RDCs are labelled with the information of the relative orientation of the internuclear vector between two nuclei, e.g. N-H, with respect to the magnetic field and consequently also the relative orientations among the internuclear vectors throughout the molecule. RDCs in an anisotropic environment are described in terms of their orientation to the fixed alignment frame and the orientation averaging of the molecule can be described by an order matrix, or the alignment tensor (Figure 3.25).

The alignment tensor is a second order traceless tensor, which can be described by 5 independent elements: the orientation by $\alpha$, $\beta$, $\gamma$ angles and
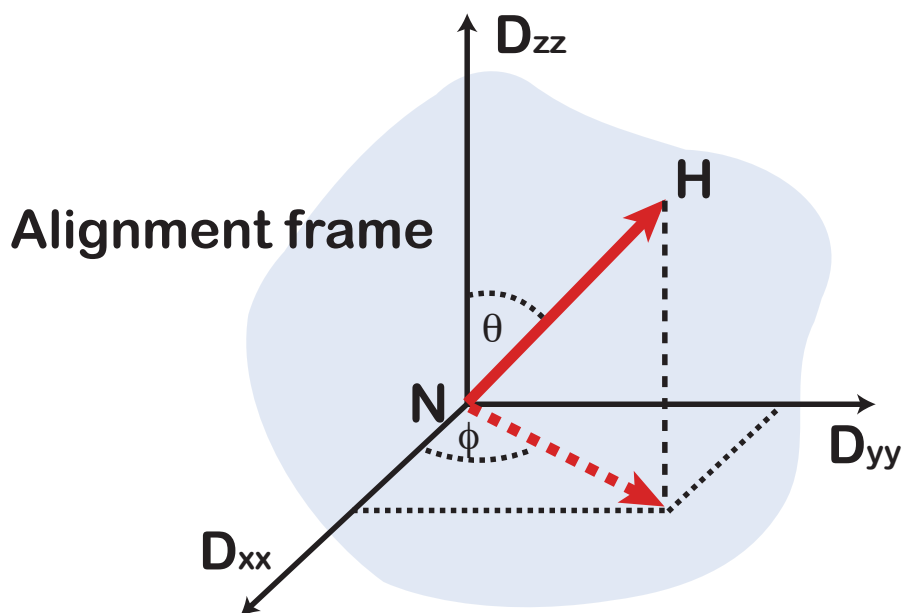
**Figure 3.25:** Schematic representation of a N-H vector oriented in the alignment frame is shown. The internuclear N-H vector is shown in solid red arrow and it projection in the x-y plane is shown as the dashed arrow. $\theta$ is the angle between the N-H vector and the $z$ axis of the tensor and $\phi$ is the angle between the project of the N-H vector in the x-y plane relative to the $x$ axis. The blue shade symbolized a globular protein like the 15.5K.

the degree of average alignment by the axial component ($D_a$) and the rhombic component($D_r$) with units in Hz (hertz). The expressions for the last two terms are

$$D_a = \frac{1}{3} \cdot \left( D_{zz} - (D_{xx} + D_{yy})/2 \right), \qquad \text{and} \qquad D_r = \frac{1}{3}(D_{xx} - D_{yy}). \quad (3.8)$$

$D_{zz}$, $D_{xx}$ and $D_{yy}$ are three eigenvalues of the alignment tensor and have the relation

$$|D_{zz}| \geq |D_{xx}| \geq |D_{yy}|. \tag{3.9}$$

From $D_a$ and $D_r$ values, the rhombicity R is defined as:

$$R = \frac{D_r}{D_a} \tag{3.10}$$

The RDC now can also be expressed in the spherical coordinates $\theta$, $\phi$ of the internuclear vector relative to the alignment tensor

$$D(\theta, \phi) = D_a \cdot \left\{ (3\cos^2\theta - 1) + \frac{3}{2}R \left( \sin^2\theta \cos 2\phi \right) \right\}, \qquad (3.11)$$

where $\theta$ is the angle between the internuclear vector and the $z$ axis of the tensor and $\phi$ is the angle between the projection of the internuclear vector in the x-y plane relative to the $x$ axis.

Experimentally RDCs can be measured by different NMR experiments. The one used in this work is an IPAP type experiment developed in the group of Ad Bax (181). IPAP stands for in-phase/anti-phase and is measured in

interleaved manner. In the in-phase experiment (Figure 3.26), the coupled peaks with the separation of J (isotropic sample) or J+D (anisotropic sample) have the same sign whereas in the anti-phased experiment they have opposite signs. The spectra from the two experiments are then summed and subtracted to result two separate HSQC spectra, each containing one set of these coupled peaks and having only positive signs. In this way, double peaks in the coupled spectra are disentangled and the signal overlap problem is avoided. The chemical shift differences between the corresponding peaks from the summation and subtraction of the two coupled spectra are then extracted. They are the experimentally measured RDCs.

As described above, J couplings were first measured on an isotropic sample and then alignment medium was added to the sample. The same experiment was measured on this anisotropic sample in the same spectrometer and under the same conditions. The differences between the J+D and J couplings (RDCs) can then be extracted (Figure 3.27).

The measured RDCs have different uses in the structural determination process. They can be used in combination with NOE restraints in *ab initio* determination of the global fold of a protein or can be used in refinement which aims at improving the structural quality of an existing 3D structure of the molecule determined by NMR or X-ray crystallography. The essence of RDC refinement is to change the existing 3D structure until the difference between calculated RDCs from the structure and the experimental RDCs is minimized. Again there are different ways to carry out this task. In this work, this task was achieved with the help of PALES program and the water refinement protocol from Xplor.
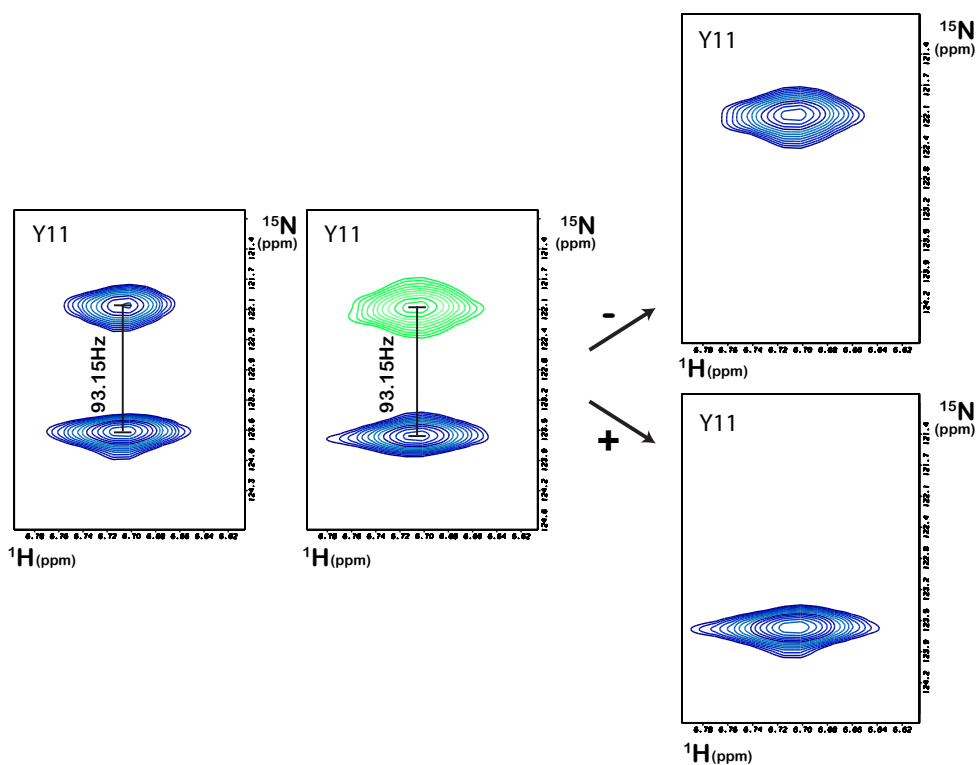
**Figure 3.26:** Residue Y11 is taken as an example here. The peaks in the in-phase and anti-phase spectra from the IPAP experiment are added and subtracted to give the two result peaks. The splitting between the two peaks equals to the J or J+D coupling value.
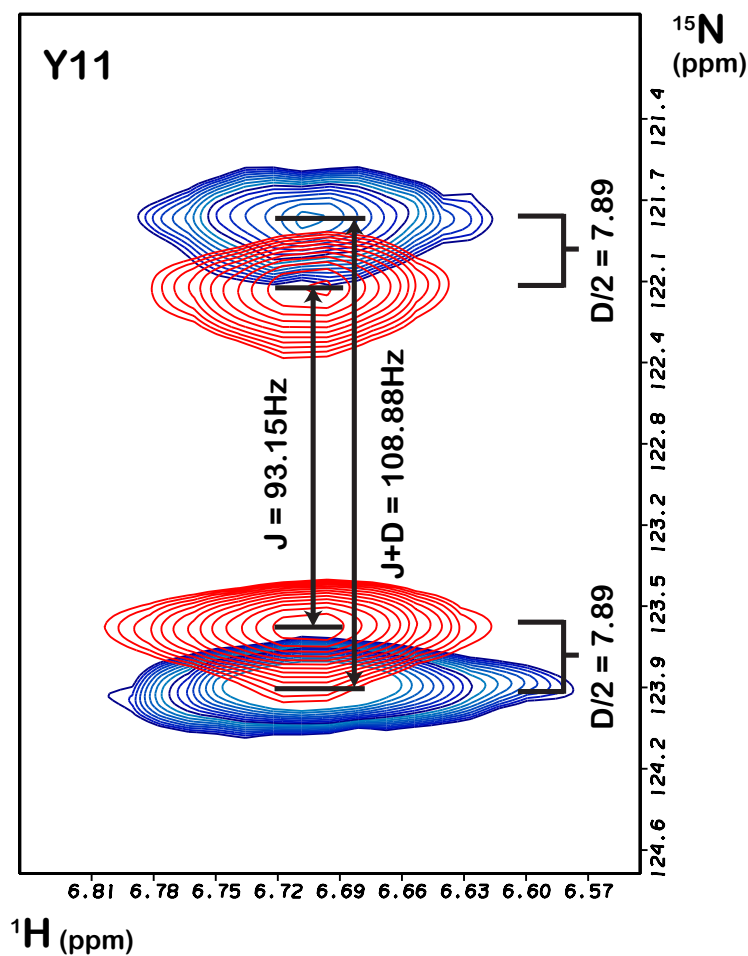
**Figure 3.27:** Residue Y11 of 15.5K-2 is again taken as an example here. The salt concentration of the sample is 0.12 M and the concentration of Pf1 phage as the alignment medium in the anisotropic sample is ∼10 mg/mL. The difference between the J+D coupling (doublet indicated in blue) measured on the anisotropic sample and the J couplings (doublet indicated in red) measured on the isotropic sample is the RDC (D).

PALES (prediction of alignment from structure) is a program developed in the group of Dr. Markus Zweckstetter for analyzing RDCs (143). One of its options namely the 'Best-fit' option, calculates the alignment tensor from a given 3D structural template and the experimental data. The calculated alignment tensor and the template structure are then used to back-calculate a set of RDCs. The back-calculated RDCs are compared with the experimentally obtained RDCs. The correlation between the calculated and measured RDCs are defined as the Pearson correlation coefficient (R) and the quality (Q) factor. The Q-factor can be expressed as

$$Q = \sum_j \frac{(D_j^{exp} - D_j^{calc})^2}{(D_j^{exp})^2}, \tag{3.12}$$

where $j$ stands for residues of the protein.

Calculated alignment tensor is also reported including the $D_a$ and $D_r$ values. $D_a$ and $D_r$ values and the measured RDCs are required for the water refinement protocol. After the refinement in Xplor, the resulting structure of the molecule is used as the template in the 'Best-fit' option in PALES and as before alignment tensor is calculated for this refined structure and difference between the calculated and measured RDCs should become smaller, hence smaller $R_{dip}$ and Q values. This procedure is repeated until the smallest $R_{dip}$ and Q values have been reached.

## 3.4.2 HSQC experiment and RDC refinement of 15.5K-2

Before carrying out RDC measurement on the mutant 15.5K-2, the protein needed to be $^{15}$N and $^{13}$C labelled and assigned. The HSQC spectrum of 15.5K-2 bound to 1:1 ratio of U4 5'-SL-33nt showed good signal dispersion like the wild type. As the overall pattern of the HSQC spectrum was not changed, a complete set of triple-resonance experiments is not needed for the backbone assignment. HNCA and HNCOCA experiments were carried out to establish the $^{13}$C$_\alpha$ connection whereas the residue type was confirmed using the combined information obtained from the backbone assignment of wild type 15.5K and the CBCACONH experiment which correlates the resonances of $^{13}$C$_{\beta(i-1)}$, $^{13}$C$_{\alpha(i-1)}$ to the amide proton. The signals from the backbone amide of 15.5K were unambiguously assigned to 97.5% (Figure 3.28). The chemical shift changes observed in the HSQC spectrum of 15.5K-2-U4 5'-SL could be identified. The largest chemical shift change resulted from the N77T mutation. Large chemical shifts are mainly located on helix $\alpha$3 of 15.5K, which encompasses the mutation site. N47 and G49 on helix $\alpha$2 of 15.5K are close to the mutation site and also exhibited high degree of changes. The degree of chemical shift changes was colour-coded on the crystal structure of 15.5K (Figure 3.29).

As the chemical shift pattern in the HSQC of 15.5K-2 did not differ dramatically from the wild type protein, the crystal structure of 15.5K could be employed as the starting template structure for the RDC refinement. The refinement of the crystal structure of 15.5K using RDCs measured from the 15.5K-2 mutant should reveal the subtle structure differences between the two proteins. For obtaining the anisotropic sample, filamentous phage
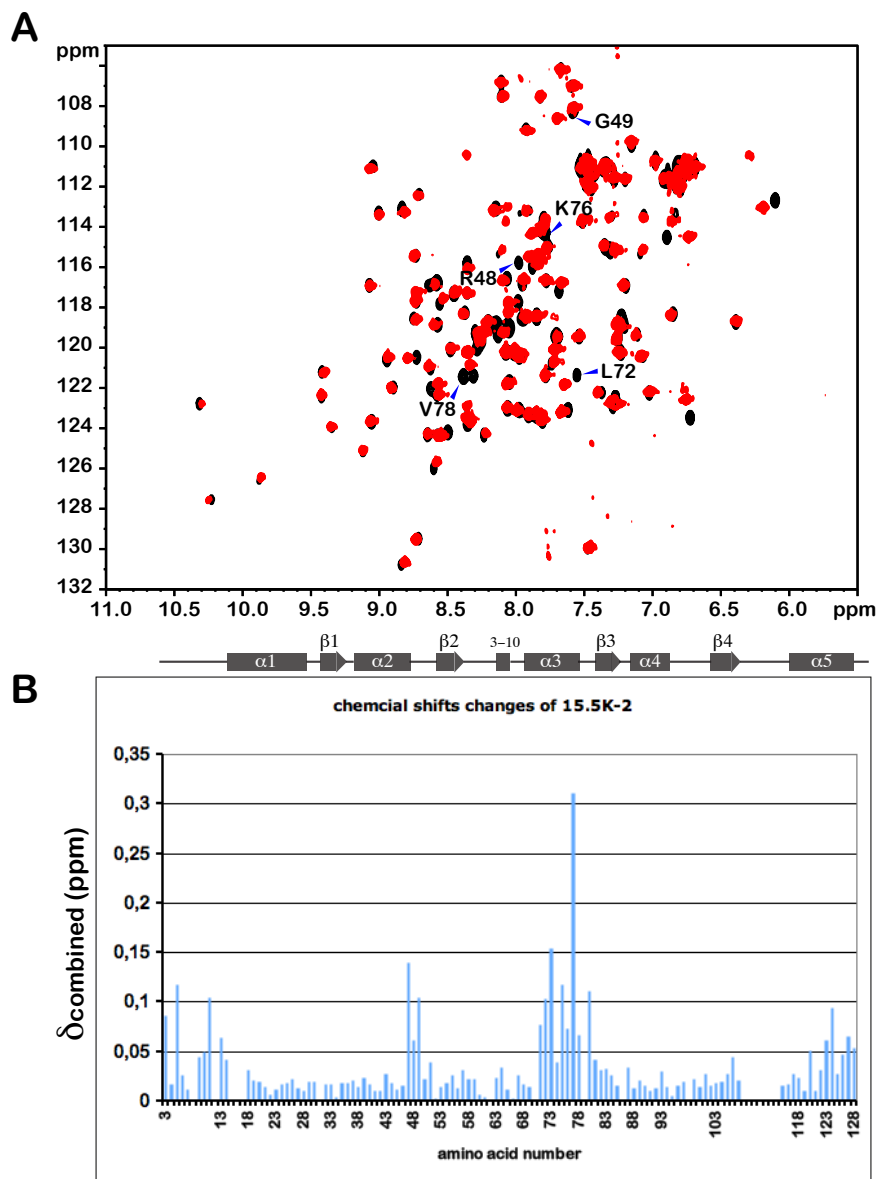
**Figure 3.28:** (A) Overlay of the HSQC spectrum of the wild type 15.5K-U4 5'-SL complex (black) and that of the 15.5K-2-U4 5'-SL complex (red). 5 of the resonances showed large chemical shift changes are indicated in the spectrum. (B) Histogram plot of the combined chemical shift changes observed by comparing the HSQC spectrum of 15.5K-2-U4 5'-SL and that of wild type 15.5K-U4 5'-SL.
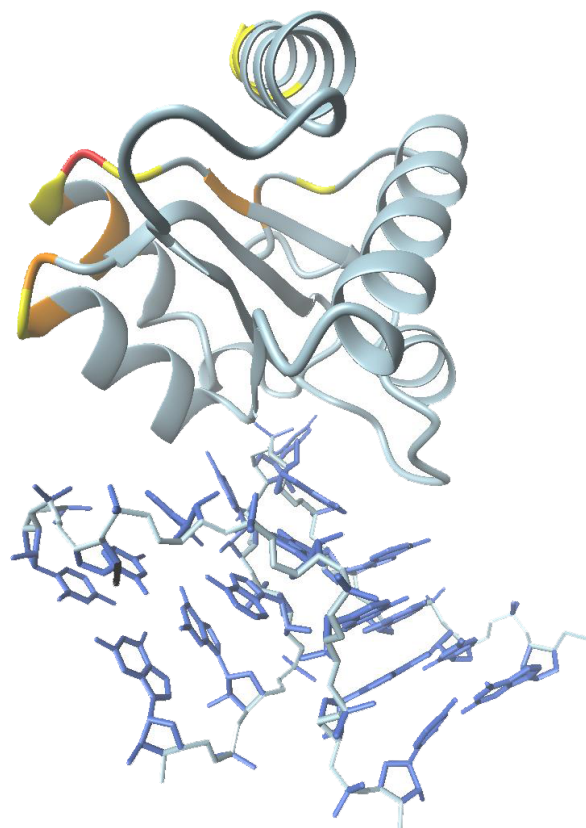
**Figure 3.29:** Chemical shift changes in 15.5K-2-U4 5'-SL are colour coded on the 15.5K-U4 5'-SL crystal structure. Red colour indicates a combined chemical shift change greater than 0.15 ppm. Orange colour is given to changes between 0.1 and 0.15 ppm. Yellow shows small changes between 0.05 to 0.1 ppm (see Figure 3.29)

Pf1 was used. The phage particles are highly negatively charged and bind non-specifically to macromolecules which have significant positively charged surface patches (182). It is therefore expected to cause a strong alignment of RNA binding proteins such as 15.5K, which usually possess positively charged surface patches. For RNA bound 15.5K, the electrostatic interaction between the phage and the protein is expected to be less severe. However, very strong alignment by the addition of an excessive amount of Pf1 could also cause line broadening due to the decreased transverse relaxation time and should, therefore, be avoided. $J_{N-H}$ scalar coupling was measured on the 15.5K-2-U4 5'-SL-33nt sample with $\sim$ 0.3 mM concentration. For the $(J+D)_{N-H}$ coupling measurements, filamentous phage Pf1 from ASLA biotech was added to the same isotropic sample to an end concentration of $\sim$ 10 $\mu$g/$\mu$l and the salt concentration was 120 mM. A deuterium splitting value in the range of 10-13 Hz was obtained. The chemical shift pattern of 15.5K-2 in the anisotropic sample did not vary from that of the isotropic sample upon the addition of Pf1. All experiments for J and J+D measurements were carried out at 308K on a 600MHz Bruker spectrometer equipped with a cryo probe. The couplings were measured using a 3-interleaved IPAP-HSQC experiment, modified from that developed in Ad Bax's group (181), was used. The 3-interleaved experiment, including an in-phase, an anti-phase and a decoupled HSQC experiment, was measured with SW/SWH values of 13.984 ppm (8389.262 Hz) and 32.000 (1945.431Hz) for F2 ($^1$H) and F1($^{15}$N) respectively. TD2 was set to be 2048 and TD1 648. The intensities of the signals in anti-phase spectra tend to be weak for molecule with large size due to the fast relaxation of anti-phase magnetization, thus, a 3-interleaved

experiment is favoured over a 2-interleaved experiment, as this problem is not present in a decoupled HSQC spectrum.

The spectra were processed with Felix using a Felix macro, which disentangles the 3 experiments resulting an in-phase, an anti-phase and an HSQC spectrum. It then performs zero-filling to 1k in the F2 dimension and to 2k in the F1 dimension. As described above, the in-phase and anti-phase spectra were added and subtracted and couplings with values of J/2 or (J+D)/2 were extracted between the spectrum from the summation and the HSQC spectrum. This allows one to manually take 1D traces of the coupled peaks through the centres of their F1 dimension. It then performs zero-filling to 64k in the F1 dimension, which resulted in a digital resolution of 0.03Hz/point in the F1 dimension. The couplings were then extracted by shifting one of the 1D traces with respect to the other until the two traces completely overlap with each other. The shifted amount in Hz is then the desired coupling.

The crystal structure of 15.5K wildtype protein was used as the template PDB and 'Best-fit' option provided by PALES was used for the calculation. The alignment tensor from measured RDCs fit to the calculated alignment tensor from the template structure with an R correlation coefficient of 0.85 and a Q factor of 0.35 (Figure 3.29). The calculated alignment tensor, with $Da_{HN}$= 15.155 Hz and rhombicity of 0.248, was used for the subsequent water refinement. A starting input structure of the 15.5K-2 mutant with the appropriate mutations, namely E74R/D75K/K76M/N77T, was needed for the water refinement. This structure was generated by exchanging residues

74-77 in the crystal structure of 15.5K to the corresponding residues in 15.5K-2 using Chimera (University of California San Francisco). Therefore, apart from the differences in the side chains of these 4 amino acids, the resulting starting structure of 15.5K-2 is identical to the crystal structure of the wild type protein. Hydrogen atoms were built in and the psf files were generated using a standard generation protocol from Xplor-NIH. The initial 15.5K-2 structure was minimized with 70 steps of Powell minimization. The water refinement with the measured N-H RDCs as 'sani' restraints was then carried out. The $Da_{HN}$ and rhombicity values were used as the coefficients for the sani term. Square potential with a force constant of 1.0 was used. This water refinement was carried out with 200 steps and 0.5 fs time-step for the heating stage, 2000 steps and 1 fs time-step for the hot stage at 500K. During this refinement intra-helical dihedral angle restraints and intra-helical NOE restraints were given to prevent the secondary structure of the protein from distorting. These restraints were taken from standard values for protein $\alpha$ helices. The resulting PDB was then used as the new template PDB for the alignment tensor calculation with the 'Best-fit' option in PALES. The alignment tensor from experimental RDCs fitted to the newly calculated RDCs from the refined structure with an R correlation coefficient of 0.995 and a Q factor of 0.063. Normally after RDC refinement the expected Q factor is between 0.05 and 0.3. Therefore, the structure had converged after one round of refinement. Overlapping the refined structure of 15.5K-2 mutant and the wild type crystal structure gave a overall RMSD of 1.26 Å for the backbone. It could be seen that the overall structure of the 15.5K and the interaction surface were not significantly changed. This is consistent with

the chemical shift analysis (Figure 3.30).

### 3.4.3 Gel mobility shift assays of 15.5K mutants

From the chemical shift analysis and RDC refinement, we still could not deduce the reason for the dysfunction of 15.5K-2. Therefore, further investigation was carried out using biochemical methods. In order to pinpoint the residue responsible for the dysfunction of 15.5K-2 mutant, 4 mutants were created: E74R, D75K, K76M and E74R/D75K using the Quickchange protocol from Quiagen. These mutants were expressed in *E. coli* as GST-tagged fusion proteins and were purified in the same manner as for the wild-type 15.5K. The GST-tag was then cleaved using thrombin protease yielding all 4 soluble proteins (Figure 3.31).

Gel mobility shift assays were carried out with these mutants to test their abilities in forming ternary complexes. 20 fmol of 5'-end radioactively labelled U4 5'-SL-33nt was incubated with 20 pmol of 15.5K or 15.5K mutants (15.5K-2, 15.5K-E74R, 15.5K-D75K, 15.5K-E74R/D75K) and 60 pmol of MBP-hPrp31. The *in vitro* reconstituted protein-RNA complexes were then applied to a 6% native gel (Figure 3.32). Surprisingly, all of the mutants were able to form ternary complexes even in the case of 15.5K-2. 15.5K-2, 15.5K-D75K and 15.5K-E74R/D75K mutants showed a barely detectable amount of reduction in binding affinity with the most noticeable change shown by 15.5K-D75K. To ensure that this result was not an artifact from over saturation of 15.5K/15.5K mutant-U4 5'-SL complexes, a gel mobility shift titration assay with increasing amounts (6, 12, 30 and 60 pmol) of MPB-hPrp31 was carried out (Figure 3.33). Again the D75K mutation alone could
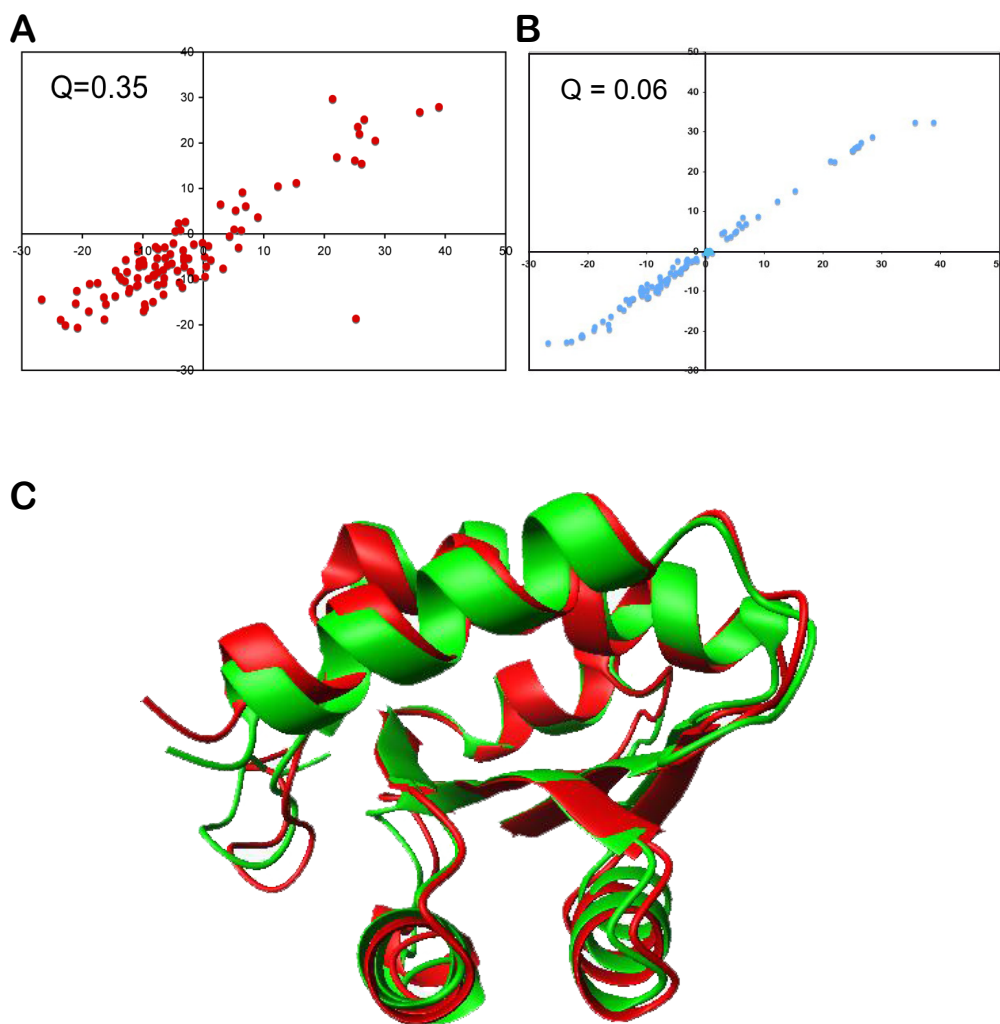
**Figure 3.30:** (A) Correlation of the calculated and experimental RDCs before refinement. The Q factor was 0.35. (B) Correlation of the calculated and experimental RDCs after refinement. The Q factor became 0.06. (C) Superimpose the RDCs refined structure of 15.5K-2 and the crystal structure of 15.5K in 15.5K-U4 5'-SL.

**Figure 3.31:** Purification of 15.5K and 15.5K mutants. (A) Lanes 1-5 correspond to the wild type 15.5K, the 15.5K-2, the 15.5K-E74R, the 15.5K-D75K and the 15.5K-E74R/D75K mutant protein after purification respectively. (B) shows the 15.5K-K76M protein after purification.

account for the slight reduction ($\sim$25%) of hPrp31 binding in 15.5K-2 and 15.5K-E74R/D75K. A gel mobility shift assay with the 15.5K-K76M mutant also showed no reduction of binding affinity to His-61K in comparison to the wild type protein supporting the observation that the D75K mutation is likely to be the most important change (data not shown).

## 3.5 15.5K associated to U3 box B/C, box C'/D and U14 box C/D

15.5K associates with different RNAs containing the K-turn motif and nucleates the binding of other secondary binding proteins. It provides the link between pre-mRNA and pre-rRNA processings. It also challenges us with
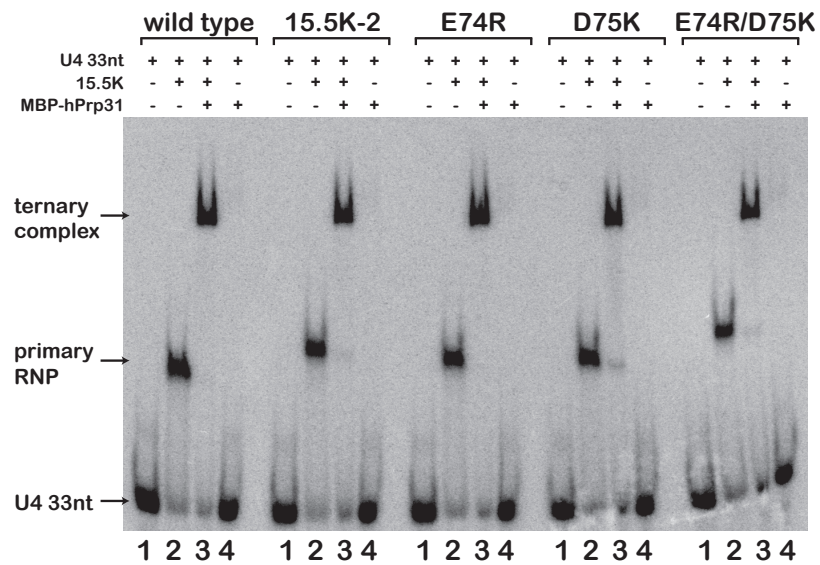
**Figure 3.32:** Gel mobility shift assays exhibits the relative migration of 5'-end radioactively labelled U4 5'-SL-33nt in free form, in the primary RNP with 15.5K and in the ternary complex with MBP-hPrp31. Assays have been carried out for wild type 15.5K, 15.5K-2 mutant, 15.5K-E74R, 15.5K-D75K and 15.5K-E74R/D75K. In each case, lane 1 contains U4 5'-SL-33nt alone; lane 2: the primary RNP; lane 3: the ternary complex; lane 4: negative control. The '+' or '-' sign above indicates the presence or absence of the component respectively.

**Figure 3.33:** Gel mobility shift titration assays were carried out on the primary RNPs containing the wild type 15.5K and the 15.5K mutants with increasing addition of MBP-hPrp31 protein. U4 5'-SL-33nt was 5'-end radioactively labelled and its migration in different complexes are shown. '+' and '-' sign indicates the presence or absence of the component respectively. For MBP-hPrp31, the increasing amounts in pmol are shown. In each case, lane 1 contains U4 5'-SL-33nt alone; lane 2: the primary RNP; lane 3-6: the ternary complex with increasing amount of MBP-hPrp31 added; lane 4: negative control.

the interesting question: namely how do these primary RNPs (snRNP and snoRNP) discriminate between the secondary binding partners. Is this recognition primarily carried out by the RNAs or is it also caused by the structural differences exhibited by 15.5K in different RNPs? To answer these questions, possible changes in the backbone of 15.5K were monitored by chemical shift analysis using HSQC spectra of the 15.5K in different RNPs. The two snoRNPs, the so-called U14 box C/D snoRNP and U3 box C/D snoRNP, were studied by biochemical means (127; 131) and were shown to contain 15.5K as the nucleation factor for the RNP assembly. The U14 C/D snoRNA contains the K-turn motif and a large, structurally uncharacterized loop (Figure 3.34). For the HSQC experiment with 15.5K and U14 box C/D RNA, the K-turn region together with the adjacent stem I and stem II were included in the U14CD-26nt construct. A penta-loop identical to that in the U4 snRNA was added to stabilize the duplex. The penta-loop does not form direct contacts with 15.5K, and therefore is not expected to cause chemical shift changes. The U3 box C/D snoRNA is a special case. Instead of box C/D and box C'/D' sequence motifs as usually found in box C/D snoRNAs, it contains a U3 specific box B/C motif and a box C'/D motif (Figure 3.35). Biochemical studies have been carried out on the U3BC (box B/C part of U3 snoRNA) (131), whereas U3CD (box C'/D part of U3 snoRNA) is much less studied. 15.5K protein binds to the U3BC and initiates the binding of the hU3-55K protein. It is also expected to bind to U3CD and nucleate the association of other boxC/D specific proteins namely Nop56, Nop58, fibrillarin, Tip-48, Tip-49 and NIS proteins (127). As two different RNPs are formed on one piece of U3 boxC/D snoRNA, two RNA constructs encompassing the

U3BC (U3BC-62nt) and U3CD (U3CD-duplex-35nt) were used in the HSQC experiments. The 62nt long U3BC construct was used because a shorter construct lacking the stem V failed to bind to the 15.5K efficiently and resulted in double sets of peaks in the HSQC spectrum corresponding to the free and the bound form of 15.5K. The 15.5K protein was $^{15}$N labelled and allowed to form complexes with the three snoRNA constructs. The complex sizes for 15.5K-U14CD, 15.5K-U3BC, and 15.5K-U3CD are $\sim$ 24 kDa, $\sim$ 45 kDa, and $\sim$ 38 kDa, respectively. $^{1}$H-$^{15}$N HSQC spectra were recorded on these complexes with a sample concentration of $\sim$ 0.3 mM in each case. The spectra of 15.5K- U3CD-35nt and 15.5K- U14CD-26nt complexes were recorded at 308 K and pH 7.6 on a Bruker 600 MHz spectrometer equipped with a cryo probe. For the 15.5K-U3BC-62nt as the complex size is $\sim$ 45 kDa the spectra were recorded under the same conditions as above on a Bruker 900 MHz spectrometer. The resulting spectra of the three complexes were overlaid with the spectrum of 15.5K U4-5'-SL (Figure 3.36). All of the spectra showed good chemical shift dispersion as in the case of the wild type. The overall backbone chemical shift pattern was not dramatically changed as compared with the 15.5K-U4 5'-SL complex. Especially in the case for 15.5K-U3BC complex, the resulting spectrum showed an almost identical chemical shift pattern to the 15.5K-U4 5'-SL complex.

Backbone chemical shifts changes of 15.5K protein in snoRNA complexes were analyzed (Figure 3.37). In the case of 15.5K-U14CD complex, most prominent shifts changes were shown by residues K37, T45, V95, R97 and V99 which are located on helix $\alpha$2 and at the beginning of the helix $\alpha$4 of the 15.5K protein. For 15.5K-U3CD complex, the large shifts arose from
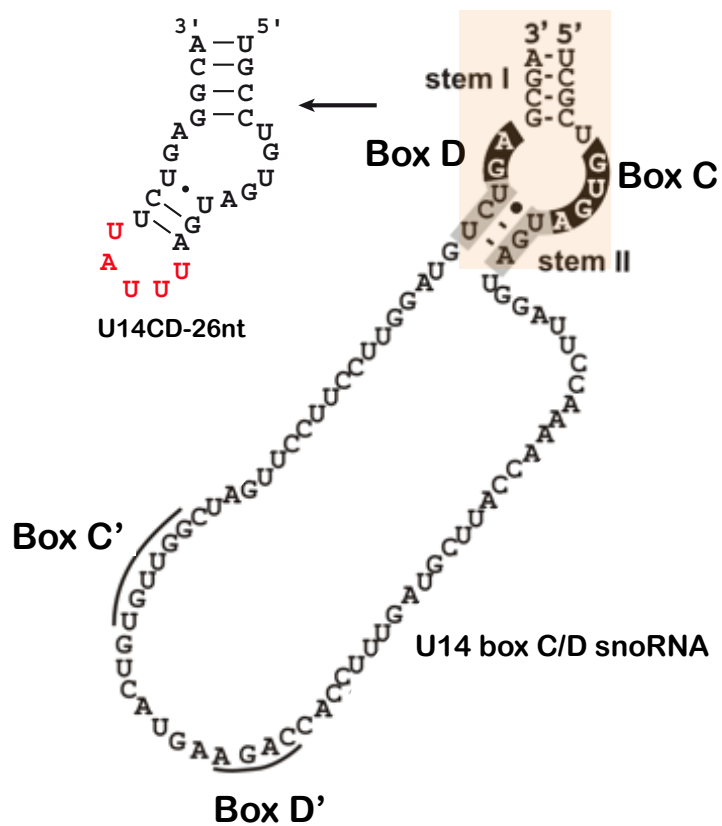
**Figure 3.34:** The origin of the U14CD RNA construct used in the NMR studies is shown. The position of the construct in the U14 snoRNA is shown with pink shaded box. The additional penta-loop in the construct is coloured in red (picture modified from Figure.1 in (135)).
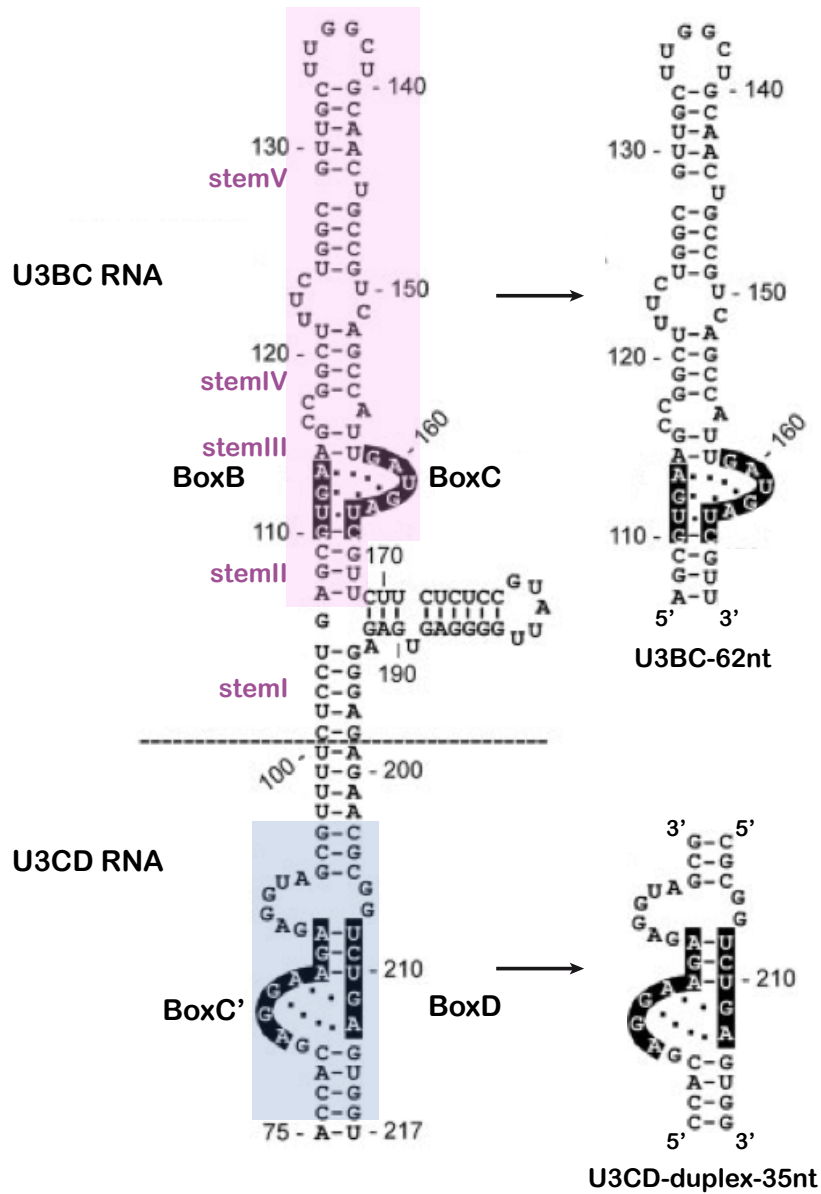
**Figure 3.35:** The origins of the U3BC and U3CD RNA constructs used in the NMR studies are shown. The positions of the constructs in the U3 boxC/D snoRNA are shown with pink and blue shaded boxs (picture modified from Figure.1 in (131)).
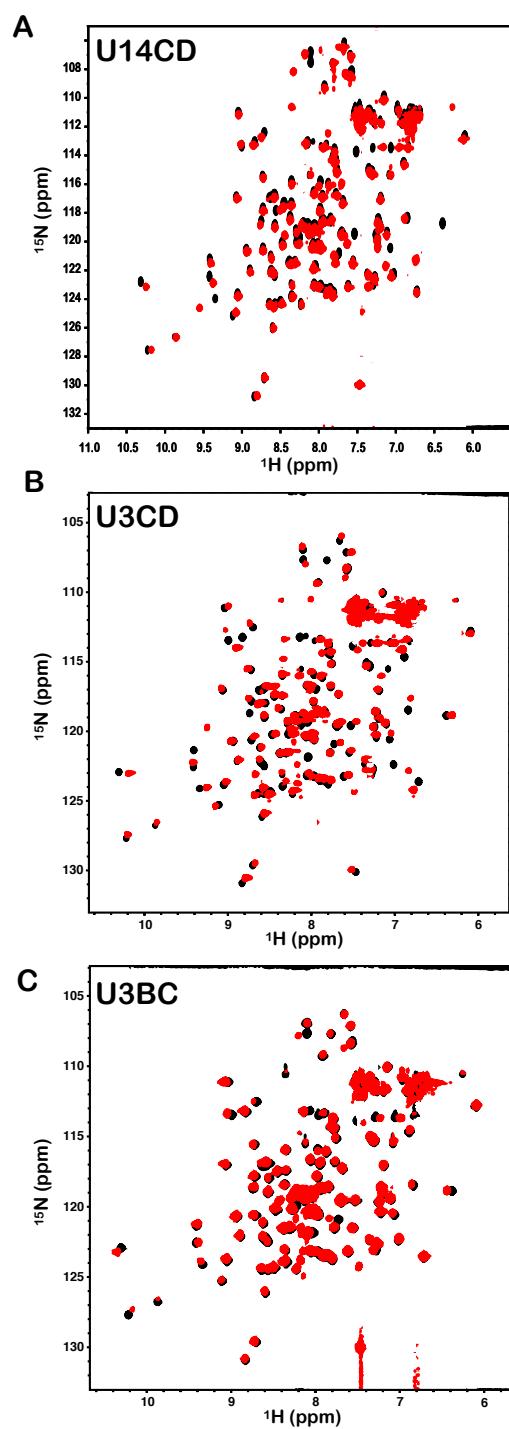
**Figure 3.36:** The HSQC spectra of 15.5K in complexes with U14CD,U3CD and U3 BC RNAs are shown in red and are overlaid on the HSQC spectrum of 15.5K in complex with U4 5'-SL. (A) 15.5K-U14CD; (B)15.5K-U3CD; (C) 15.5K-U3BC

residues A58, I65 and I66, which cluster in the region close to the 3-10 helix. In comparison to the complexes involving U14CD and U3BC RNAs, 15.5K-U3CD complex overall showed more deviation from the 15.5K-U4 5'-SL. In the case of the 15.5K-U3BC complex, all of the changes were much smaller than those in the complexes involving U14 and U3 snoRNAs. Noticeable changes arose from residues A39, E41, K44, T45, C93 and V95, which, as in the U14 case are located on the helix $\alpha 2$ and the loop region before the helix $\alpha 4$ of 15.5K. For all three complexes, large, medium and small chemical shift changes were defined as greater than 0.15 ppm, between 0.1 and 0.15 ppm, 0.05 and 0.1 ppm respectively. The degree of the chemical shift changes were colour-coded on the crystal structure of 15.5K-U4 5'-SL-24nt (Figure 3.38).
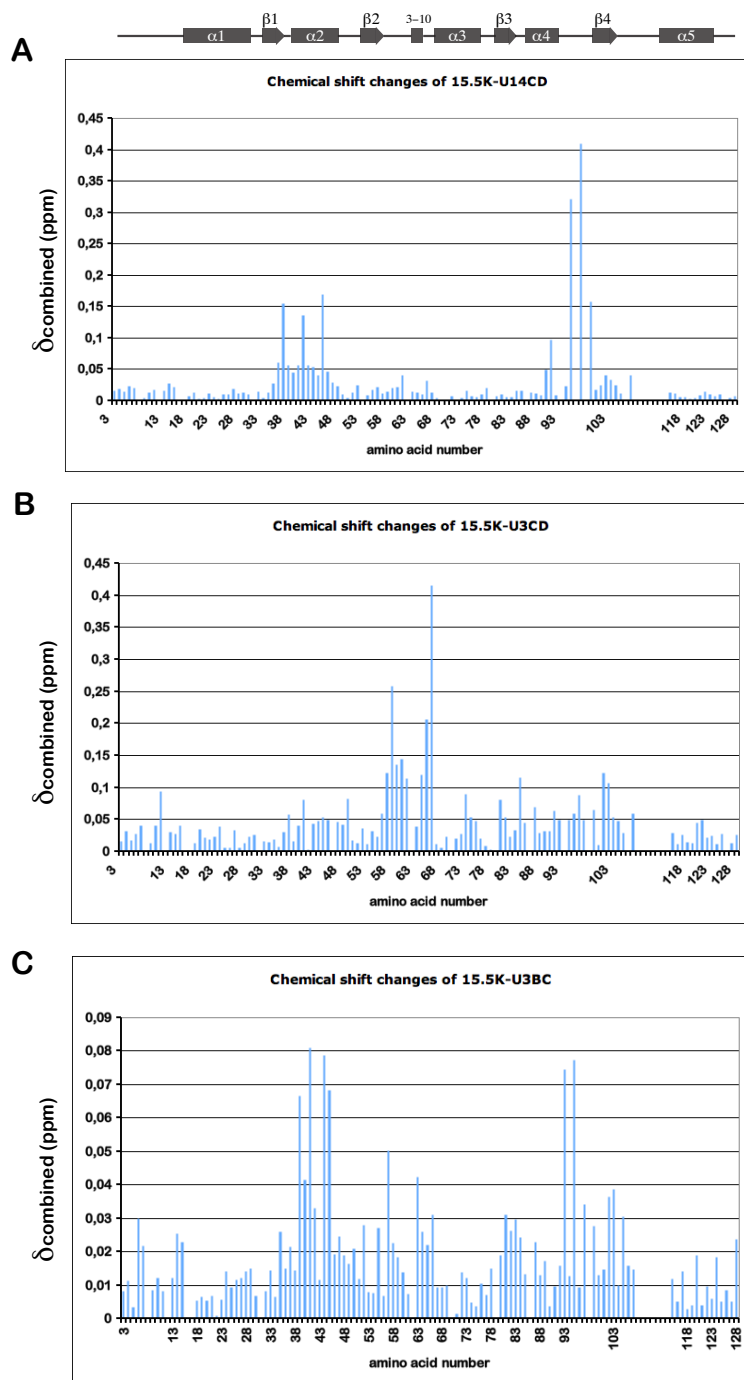
**Figure 3.37:** Histogram plot of combined chemical shift changes observed in 15.5K-snoRNA complexes in comparison to 15.5K-U4 5'-SL. (A) 15.5K-U14CD; (B)15.5K-U3CD; (C) 15.5K-U3BC
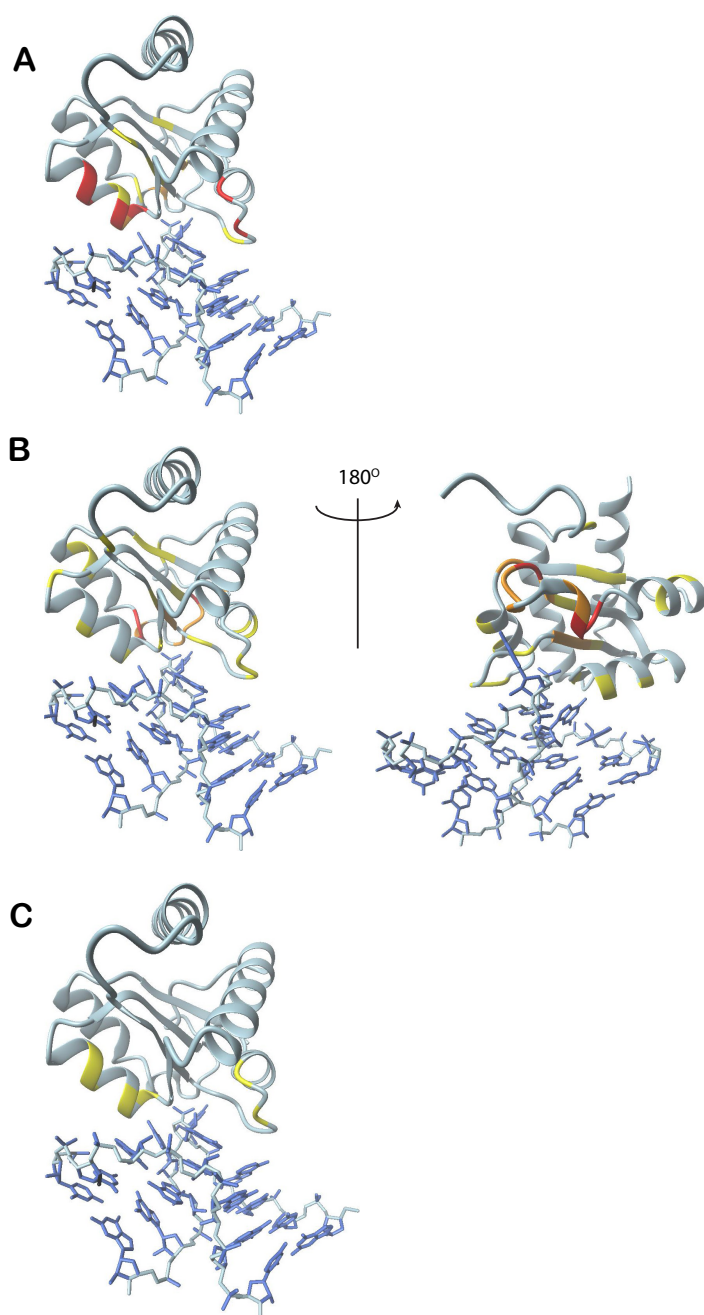
**Figure 3.38:** The degrees of combined chemical shift changes observed in 15.5K-snoRNA complexes in comparison to 15.5K-U4 5'-SL are colour coded on the crystal structure of 15.5K-U4 5'-SL. Red: a change larger than 0.15 ppm; organe: a change between 0.1 and 0.15 ppm; yellow:a change between 0.05 and 0.1 ppm (A) 15.5K-U14CD; (B)15.5K-U3CD; (C) 15.5K-U3BC

# Chapter 4

# Discussion

15.5K is a U4/U6 snRNP specific protein in the spliceosome which locates to the nucleus. Upon its binding to U4 5'-SL, it nucleates the binding of other U4/U6 specific proteins namely hPrp31 and the CypH/hPrp4/hPrp3 protein complex. 15.5K is also found to play a similar role in box C/D snoRNP assembly. Box C/D snoRNPs are involved in pre-rRNA processing and locate to the nucleoli. The different RNAs recognized by 15.5K all share the K-turn motif. It is of great interest to understand how this structurally similar primary RNPs discriminates among the secondary binding proteins. Nop56 and Nop58 which are the secondary binding proteins in box C/D snoRNPs, share the highly conserved Nop domain with hPrp31. Knowledge of detailed protein-protein interactions and protein-RNA interactions mediated by the Nop domain in these RNPs is essential for understanding the origin of the selectivity. In this work, NMR spectroscopy was employed to investigate the protein-protein interactions between hPrp31 and 15.5K in the presence of U4 5'-SL. At the beginning of the work, 15.5K-U4 5'-SL primary RNP was

studied by NMR, in which 15.5K was the only observable molecule. The first aim is to assign all the backbone resonances of 15.5K in the HSQC spectrum, which served as the basic knowledge for all further NMR investigations carried out in the frame of this work.

## 4.1  15.5K binds to the U4 5'-SL through an induced-fit mechanism

The free 15.5K protein exhibited much less stability than its RNA-bound form and precipitated readily during HSQC measurements, which lasted about 1 hour. Therefore, triple-resonance experiments, which required up to several days of experiment time, could not be carried out on the free 15.5K. This observation is supported by other studies, in which 15.5K alone failed to crystallize and showed significantly higher sensitivity towards protease digestions than its bound form, in complex with U4 snRNA (information from Dr. Henning Urlaub). TROSY versions of triple-resonance experiments as well as 70% uniform deuteration were performed to overcome the difficulties in backbone assignment caused by the large size of the 15.5K-U4 5'-SL complex ($\sim$ 25 kDa). 99% of the backbone resonances of the bound 15.5K could be assigned, which provided the basis for further structural investigations on the ternary complex. Although obtaining a 3D structure of the free 15.5K protein was not in the scope of this NMR study, important structural knowledge on 15.5K alone could be gathered by comparing the HSQC spectra of the free and bound form of the protein using the assignment information. 12 peaks, which were absent in the $^1$H-$^{15}$N HSQC spectrum of

unbound 15.5K, bacame visible upon binding the U4 RNA. These peaks correspond to residues K37, G38, A39, N40, E41, A42, T43, E61, K86, C93, R97 and V99. These resonances are likely to have been in exchange with water due to the corresponding residues being located at regions lacking a stable secondary structure. They could also have been heavily broadened due to exchange between different conformations in these regions of the protein in the absence of the RNA. These regions, however, seem to undergo conformational change upon the binding to U4 5'-SL. All of these peaks correspond to residues forming direct contact with the U4 5'-SL or residues at the RNA binding site, which involves $\beta$1-loop-$\alpha$2, 3-10 helix and $\alpha$4-loop-$\beta$4 of 15.5K (Figure 3.6). These observations point towards an induced-fit in the 15.5K in these regions upon U4 5'-SL association. The induced-fit mechanism is often favoured in protein-RNA interactions for the formation of intimate interaction surface, lowering the energy barrier to complex formation and, as the structures of free and bound molecules are different, it could be used for signaling an accomplished binding event (183). There are three possible ways for induced-fit in protein-RNA interactions, namely protein-induced RNA folding, RNA-induced protein folding and mutually induced fit (Figure 4.1). There have been many examples of induced-fit mechanisms documented including the complexes of L30, the ribosomal homlog of 15.5 K. Binding of L30 to its cognate mRNA results in conformational changes in both the protein and the RNA, which includes restructuring of the RNA internal loop and the formation of one $\beta$-strand-loop-helix at the C-terminus of the protein, which directly contact the RNA. In the case of 15.5K-U4 5'-SL complex formation, the conformation changes that possibly occur on the RNA have been studied
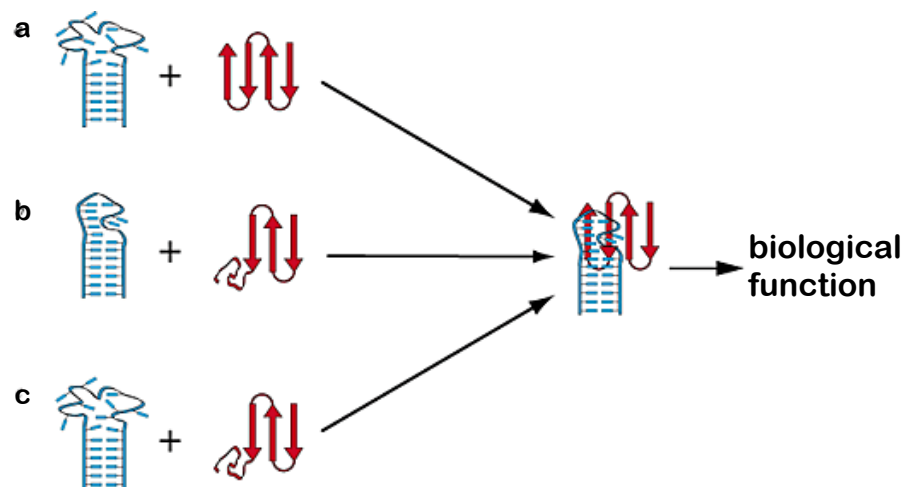
**Figure 4.1:** Schematic drawing of the formation of an RNA - protein complex. The RNA is drawn in blue, the protein in red. The three possible ways of indued-fit are: (a) protein-induced RNA folding; (b) RNA-induced protein folding and (c) mutual induced-fit (co-folding). (figure taken from Figure.1 in (183))

by biochemical, single-molecule FRET and molecular dynamics simulation (85; 94; 93). These studies suggested restructuring of the stem II and the k-turn of the U4 5'-SL upon binding to the 15.5K protein. Combined with the NMR observations on the 15.5K, it seems that a mutual induced-fit mechanism is possible in the case of 15.5K-U4 5'-SL complex formation. Currently, detailed structural studies using NMR spectroscopy on a labelled U4 5'-SL construct in free and bound form are being carried out in the group of Dr. Teresa Carlomagno. These studies will bring more insights into this issue.

## 4.2 The ternary complex of hPrp31-15.5K-U4 5'-SL

### 4.2.1 NMR findings demonstrate that Nop domain of hPrp31 is the bona fide RNP recognition motif

**Nop domain of hPrp31 is a composite RNA-protein binding domain**

One of the main foci of this project is to investigate the binding property of hPrp31 in the context of the U4 snRNP. Hydroxyl radical footprinting and UV-induced crosslinking have revealed direct interactions between hPrp31 and the U4 5'-SL ([95]; [136]) and, together with mutational analyses ([134]), have delineated RNA secondary structure requirements for hPrp31 binding. The question was whether hPrp31 also engages in direct contacts with 15.5K in the context of the U4 snRNP. The TROSY-HSQC experiments on hPrp31-15.5K-U4 5'-SL complex showed that the largest chemical shift changes with respect to binary 15.5K-U4 5'-SL complex are located on the helices $\alpha 2$ and $\alpha 3$ of 15.5K. With the help of the cross-saturation experiments, these changes were confirmed to be caused by direct protein-protein contacts between hPrp31 and 15.5K. With the same experiments carried out on the hPrp31$^{78-333}$-15.5K-U4 5'-SL complex, we have demonstrated that hPrp31$^{78-333}$ contacts 15.5K in the same manner as the full length, primarily on helices $\alpha 2$ and $\alpha 3$ and that hPrp31$^{78-333}$ contains the entire surface required for interaction with 15.5K in the ternary complex. hPrp31$^{78-333}$ encompasses the Nop domain (residues 215-333) and lacks the N-terminus and the C-terminus. The result indicates that the Nop domain alone is likely

to be capable of forming the observed protein-protein contacts. As H270 in the Nop domain was shown to form direct contacts with the penta-loop of the U4 5'-SL (136), it is reasonable to deduce that Nop domain of hPrp31 is a composite RNA-protein binding domain.

**The docking model demonstrates that Nop domain is the veritable RNP recognition motif**

Results from cross-saturation experiments provided us with information about the interaction surface on the 15.5K in the ternary complex. The interaction surface on hPrp31 was oriented by using the confirmed UV-crosslink between H270 and A39 of U4 5'-SL in combination with plausible contacts from C247 and R293 to the RNA. 3D docking models were generated using HADDOCK2.0 (see Results chapter). The best ranking structure from the best ranking cluster in run2 is presented here. In this docking model, the strongly electropositive flat surface of the Nop domain formed by helices $\alpha 8$, $\alpha 9$, $\alpha 11$ and $\alpha 12$ (see Figure 3.20) faces towards the electronegative flat surface mainly formed by the stem II and the penta-loop of the U4 5'-SL as well as helices $\alpha 2$ and $\alpha 3$ of 15.5K. The lower part of the interaction surface consists of helices $\alpha 8$, $\alpha 9$ and the C-terminal end of the $\alpha 11$ from the Nop domain contacting helices $\alpha 2$ and $\alpha 3$ of 15.5K (regions a, b and c in Figure 4.2). The upper part of the interaction surface is formed by the N-terminal end of $\alpha 9$ and the N-terminal end of $\alpha 11$ to the side of the K-turn which is not associated with 15.5K as well as to the major groove of the stem II (region 1 in Figure 4.2). Additionally, the assigned contact between H270 and A39 results in helix $\alpha 10$ contacting the penta-loop (region 2 in Figure 4.2).
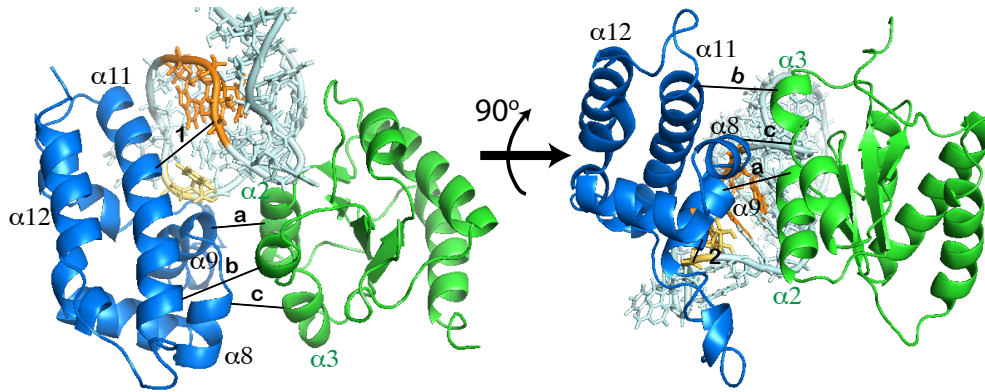
**Figure 4.2:** Contacts between the Nop domain of hPrp31 and 15.5K -U4 5'-SL are shown. hPrp31 Nop domain is indicated in blue, 15.5K in green, RNA in blueish grey. The contacts on the stem II of the RNA are indicated in orange and that on the penta-loop in yellow.

Our docking model demonstrates that the Nop domain is able to form confluent interaction surfaces with both the 15.5K and the U4 5'-SL of comparable size. The Nop domain can, therefore, be viewed as a veritable RNP recognition motif, as opposed to pure RNA interaction domains found in other proteins (184).

**A molecular ruler in the Nop domain discriminates against Box C/D-Like RNAs**

As stated in the introduction, stem II of the K-turn in the box C/D snoRNAs is extended by a U-U base pair (Figure 4.13A) (88) and a single additional base pair in stem II of U4 5'-SL is known to interfere with hPrp31 binding (134), whereas the sequence variation in stem II naturally occurs between U4 5'-SL and U4atac 5'-SL. The variation in sequence is tolerated as hPrp31 binds to both RNAs equally well, but the structural variation namely the

elongation of stem II, is not accepted. In the docking model, all hPrp31 contacts with the stem II of the RNA are mediated by the Nop domain, suggesting that this motif alone is able to discriminate against an extended stem II. From the docking model, we could suggest that the origin of this discrimination is due to the physical barrier imposed on the RNAs by the Nop domain. As shown in Figure 4.3, the penta-loop of the U4 5'-SL orients towards the helix $\alpha$10 of the Nop domain. An elongation of the stem II would lead to structural crowding and unfavourable steric hindrance. Therefore, the Nop domain could most likely act as a molecular ruler to discriminate against the box C/D- like RNAs, which naturally have stem II exceeding the ideal length.

**The additional contacts between hPrp31 and stem I of the U4 5'-SL are likely to be formed by C-terminus of hPrp31 alone**

As mentioned in the introduction, stem I containing eight rather than four base pairs has previously been found necessary in the 15.5K-U4 5'-SL complex (but not in the 15.5K-U4atac 5'-SL) for high affinity binding of hPrp31 (134). However, unlike the full-length hPrp31, hPrp31$^{78-333}$ binds equally well to 15.5K-U4 5'-SL complexes comprising either long or short stems I. To elucidate whether the N-terminal part (residues 1-77) or the C-terminal portion (residue 334-499) of hPrp31, which are missing in the hPrp31$^{78-333}$ fragment, are responsible for the different response to the length of stem I, the protein construct hPrp31$^{78-499}$ was produced by Dr. Sunbin Liu, in which only the N-terminal residues are missing. hPrp31$^{78-499}$ behaves like the full-length protein with respect to stem I, i.e. it weakly binds to a 15.5K-U4
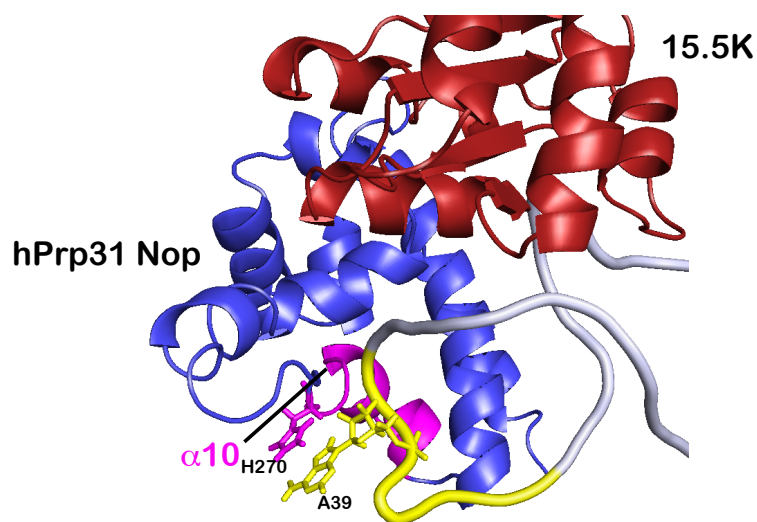
**Figure 4.3:** The Nop domain most likely to act as molecular ruler, which discriminates box C/D snoRNAs by impinging a physical barrier formed by the helix $\alpha$10. The Nop domain of the docking model and the RNA are coloured in blue and grey repectively. The penta-loop is shown in yellow. A physical barrier can be seen to be presented by helix $\alpha$10 (magenta), which blocks the stem II from exceeding 2 base pairs. The stacking interaction between H270 and A39 is shown.

5'-SL complex containing a shortened stem I with only four base pairs, while it binds well to a complex with eight base pairs in stem I. Therefore, most likely the C-terminal part (residues 334-499) of hPrp31 engages in additional contacts to stem I of the U4 5'-SL. In agreement with this hypothesis, hydroxyl radical foot-printing experiments performed by Dr. Olexandr Dybkov, show protection of residues G26-C28 in stem I upon binding of hPrp31 (134) but not of hPrp31$^{78-333}$ (Figure 4.4).

These findings are all in agreement with the high similarity of the NMR saturation transfer data from either the full-length hPrp31 or hPrp31$^{78-333}$ to the 15.5K, which confirm that the contact surface between the two proteins is completely contained in the hPrp31$^{78-333}$ fragment and suggests that the C-terminal part of hPrp31 only binds the RNA and not the 15.5K protein. Our docking model also strongly supports this hypothesis. As shown in Figure 4.5, the stretch of positively charged residues at the C-termius of the hPrp31 could easily form favourable electrostatic interactions with the RNA and subsequently could cause the additional contacts in stem I.

## 4.2.2   The crystal structure of the hPrp31$^{78-333}$-15.5K U4 5'-SL complex

In the crystallization trials carried out by our collaboration partners Dr. Sunbin Liu and Dr. Markus Wahl, flexible regions in full-length hPrp31 presumably prevented crystallization of the ternary hPrp31-15.5K-U4 5'-SL complex. The crystal structure of hPrp31$^{78-333}$-15.5K-U4 5'-SL complex, however, was solved to 2.6 Å resolution by Dr. Sunbin Liu and Dr. Markus Wahl (137). In the crystal structure, hPrp31$^{78-333}$ consists of 13 helices (Figure 4.6
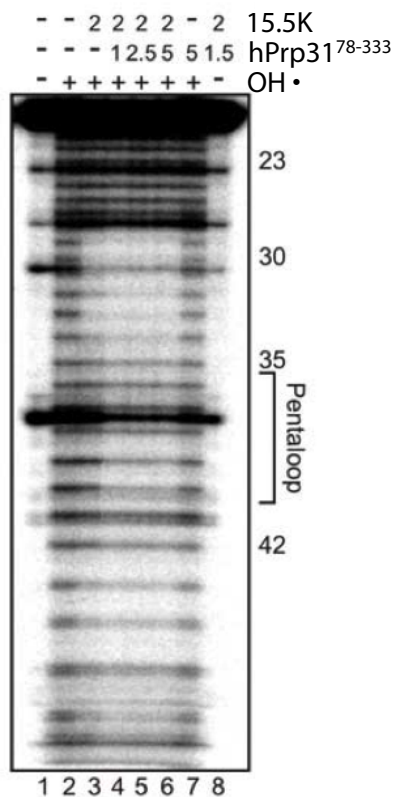
**Figure 4.4:** Hydroxyl radical footprinting of the U4 5'-SL in the absence of protein (lane 2), in the presence of only 15.5K (lane 3) and in the presence of 15.5K and increasing amounts of hPrp31$^{78-333}$ (lanes 4-6). Numbers indicate the protein concentration in $\mu$M. Numbers on the right indicate positions in the U4 5'-SL. The location of the pentaloop is indicated (figure made by Dr. Markus Wahl).
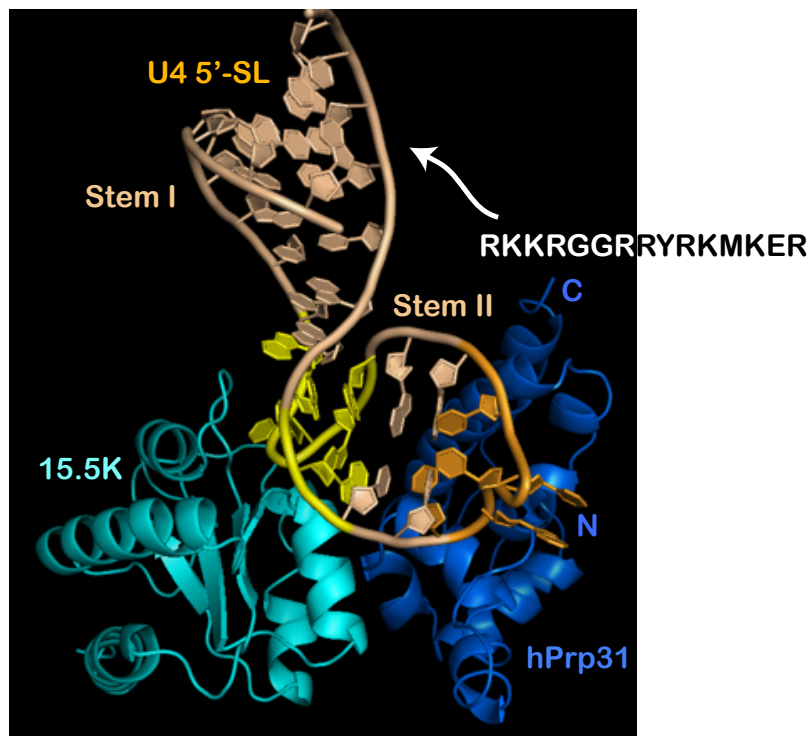
**Figure 4.5:** In this presentation of the best scoring docking model, the 15.5K is coloured in cyan, hPrp31 in marine blue, the U4 5'-SL RNA in peach with the penta-loop in orange and the K-turn in yellow. The C and N- terminus of the Nop domain are indicated. The arrow indicates that the stretch of positively charged residues at the C-termius of hPrp31 could account for the additional contacts between hPrp31 and the stem I of the U4 5'-SL.

A). The numbering of the helices in the crystal structure exceeds that of the hPrp31$^{188-332}$ by 1 due to additional helices in hPrp31 as the N-terminal region (outside the Nop domain) in comparison to Nop5p. From now on, for easy comparison the helices of the hPrp31$^{188-332}$ model will be named as the corresponding helices in the crystal structure. As in our docking model, the Nop domain of hPrp31$^{78-333}$ in the crystal structure exhibits a flat surface formed by helices $\alpha$9, $\alpha$10, $\alpha$12, and $\alpha$13. In the displayed orientation in Figure 4.6, the lower part of this surface is composed of helix $\alpha$9 and the C-terminal half of helix $\alpha$12 and interacts with the $\alpha$2/$\alpha$3 region of 15.5K in excellent agreement with the NMR analysis in solution. The upper portion of the surface, consisting of helix $\alpha$10 and the N-terminal half of helix $\alpha$12, contacts the RNA on the side of the K-turn that is not associated with 15.5K (contact region 1 in Figure 4.6B) and in the major groove of stem II (region 2). In addition, a loop following helix $\alpha$10 interacts with the capping pentaloop (region 3). The surfaces of 15.5K and of the RNA that are buried by binding of the Nop domain are of comparable size (550-650 Å$^2$ each) and confluent. All contacting regions on both proteins are in total agreement with the our docking model. Our docking model and the crystal structure confirm that the Nop domain is the *bona fide* RNP binding domain.
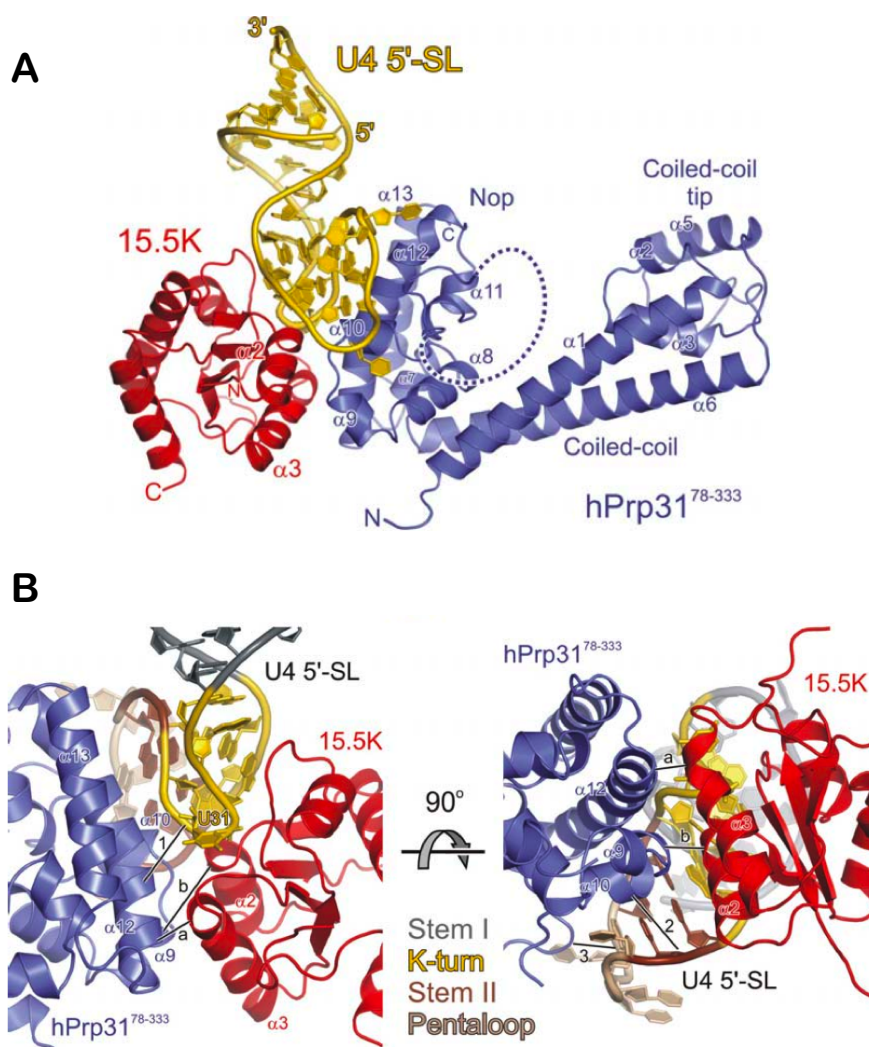
**Figure 4.6:** (A) Overview of the hPrp31$^{78-333}$-15.5K-U4 5'-SL complex. hPrp31$^{78-333}$ blue; 15.5K red; RNA gold. The color-coding is maintained in the following figures. Dashed line disordered loop. (B) Close-up views of the complex from the back (left) and from the bottom (right). Main contact regions between hPrp31$^{78-333}$ and 15.5K and between hPrp31$^{78-333}$ and the RNA are indicated by connecting lines and are labelled by letters and numbers, respectively. Regions of the RNA are color-coded: distal portion of stem I gray; immediate K-turn gold; distal portion of stem II brown; pentaloop beige. The bulged-out U31 denotes the tip of the K-turn and is shown in sticks (figure made by Dr. Markus Wahl).

### 4.2.3 Comparison between the docking model of hPrp31$^{188-331}$-15.5K-U4 5'-SL complex and the crystal structure of hPrp31$^{78-333}$-15.5K U4 5'-SL complex

**The lock-and-key type interaction and triangular overall shape of the ternary complexes**

For all existing docking methods, it is at the present not possible to successfully generate a 3D model for the complex, when each individual component of the complex undergoes dramatic structural rearrangement during complex formation. The model of hPrp31$^{188-331}$ was docked onto the 15.5K-U4 5'-SL complex model under the assumption that no dramatic structural changes occur in either component. From the TROSY-HSQC experiments of hPrp31-15.5K-U4 5'-SL and hPrp31$^{78-333}$-15.5K-U4 5'-SL complexes, refolding of the 15.5K was not expected as the overall backbone chemical shift patterns did not change dramatically after ternary complex formations. In this case, the assumption that hPrp31$^{188-331}$ keeps its structural integrity upon complex formation was also later demonstrated by the crystal structure of the hPrp31$^{78-333}$-15.5K-U4 5'-SL complex to be correct, as the Nop domain of hPrp31$^{188-332}$ model can be closely superimposed on the corresponding domain in the crystal structures of hPrp31$^{78-333}$ with an RMSD of 1.49 Å.

In the crystal structure consistent with our chemical shift observations and our docking model, the folding of both 15.5K and the RNA in the presence of hPrp31$^{78-333}$, as well as the interaction between the two molecules, are virtually unchanged compared to the binary 15.5K-U4 5'-SL complex. The
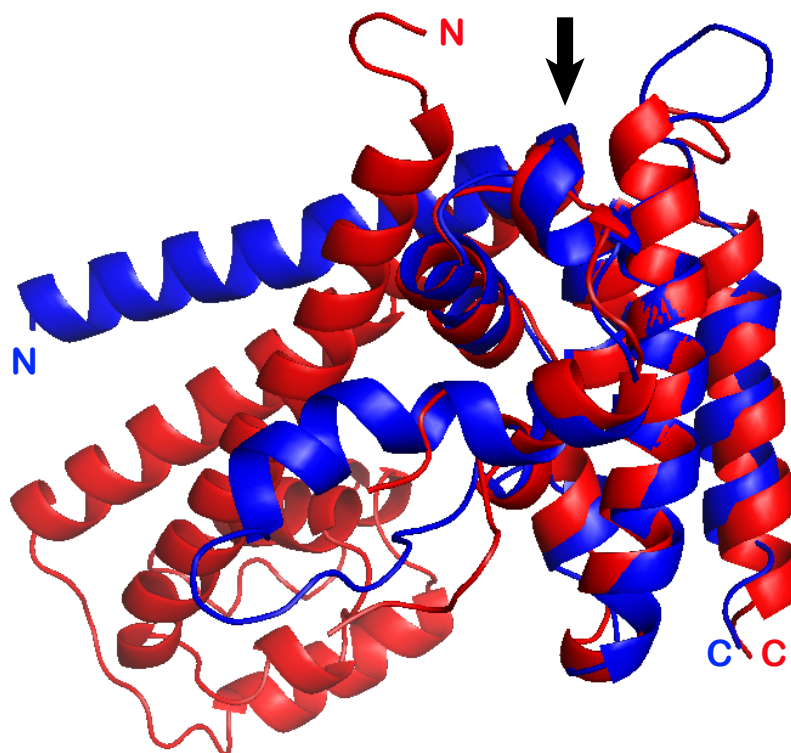
**Figure 4.7:** The crystal structure of hPrp31[78−333] is coloured in red and the model of hPrp31[188−332] obtained from comparative modelling is coloured in blue. The fitting was carried out exclusively on the Nop domain, which correspond to residues 215-334. The black arrow indicates the start of the Nop domain.

RMSD between the binary 15.5K-U4 5'-SL complex and 15.5K-U4 5'-SL in the ternary complex in both the docking model and the crystal structure is 0.4-0.5 Å. The Nop domain of hPrp31$^{78-333}$ can be closely superimposed with the corresponding domain of our model and it also closely fits to the Nop domain of (AF) Nop5p with an RMSD of 1.31Å. As the structure of Nop5p was determined in the absence of a primary RNP (100), it indicates that the structure of the Nop domain does not vary significantly upon RNP assembly. Therefore, binding of the hPrp31 Nop domain to 15.5K and the immediate K-turn resembles a lock-and-key type interaction.

Looking at the overall topology of our docking model of hPrp31$^{188-331}$-15.5K-U4 5'-SL and the crystal structure of the hPrp31$^{78-333}$-15.5K-U4 5'-SL complex, both exhibit a triangular shape with one subunit at each vertex of the triangle and each subunit contacting the other two (Figure 4.8).

**The docking model of hPrp31$^{188-331}$-15.5K-U4 5'-SL complex demonstrates the same interaction surface as the crystal structure of hPrp31$^{78-333}$-15.5K U4 5'-SL complex**

To evaluate the calculated docking structures, analysis was carried out for the 15 clusters in run1 and the 16 clusters in run2 using ana_cluster.csh script provided by HADDOCK2.0, which reports most importantly the combined HADDOCK2.0 scores for all the structures. The HADDOCK scores were calculated as described in the Results chapter and were used to rank the quality of the clusters and the individual structures. As the crystal structure of hPrp31$^{78-333}$-15.5K U4 5'-SL complex was solved in the course of our studies, it was used to cross-validate our docking models. All the structures
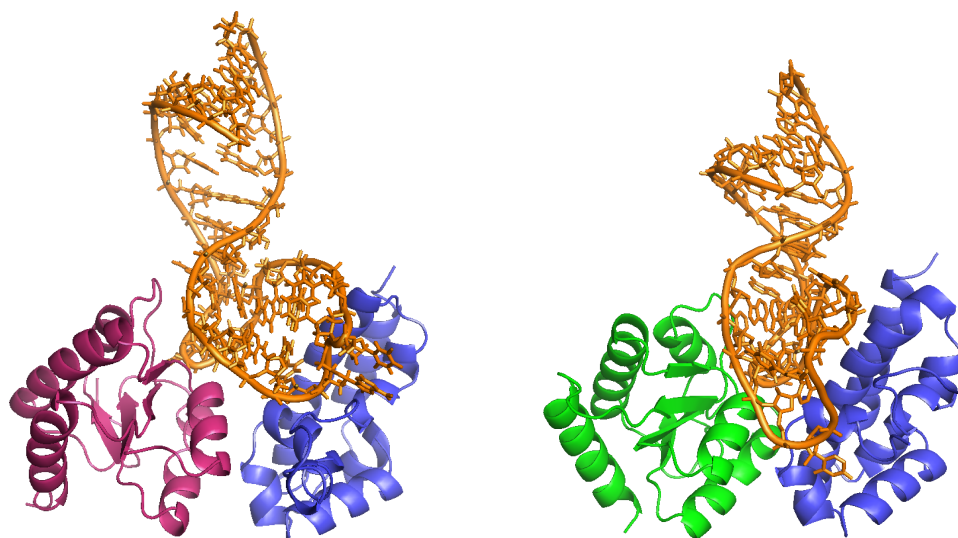
**Figure 4.8:** The docking model (left) and the crystal structure (right) are oriented in terms of hPrp31 Nop domain. In the docking model 15.5K is coloured in raspberry red, hPrp31$^{188-332}$ in blue (only residues 217-332 are shown) and RNA in dark orange. In the crystal structure of hPrp31$^{78-333}$-15.5K-U4 5'-SL complex 15.5K is indicated in green, hPrp31$^{78-333}$ in blue (only residues 217-332 are shown) and RNA in dark orange. In both the docking model and the crystal structure, 15.5K, Nop domain and RNA place themselves into a triangular shape. It can be seen that the structure of the RNA in these two models do not correspond very well, due to very limited structural information available on the RNA during docking. Although the general position of the RNA stays the same in both the docking model and the crystal structure, the fold of the penta-loop of the RNA differs.

included in the clusters from both runs were superimposed with the crystal structure of the hPrp31$^{78-333}$-15.5K-U4 5'-SL complex at the protein-protein interface. The interface on hPrp31 used in the superposition consisted of residues 237-250 from the C-terminal end of $\alpha 8$ to the N-terminal end of the helix $\alpha 10$, residues 291-307 of the helix $\alpha 12$ and the interface on 15.5K included residues 38-48 of the helix $\alpha 2$, residues 62-74 of helices 3-10 and $\alpha 3$. The interface RMSD between each docking structure and the crystal structure was plotted against the HADDOCK2.0 score of each docking structure. The plots of both docking runs are shown in Figure 4.9. Both of the plots demonstrate that the HADDOCK2.0 scores are excellently correlated with the interface RMSDs. The structures with better HADDOCK2.0 scores (lower score values) are in general closer to the crystal structure, which confirms that HADDOCK2.0 was a very suitable choice for this docking task. Using the information that Cys247 forms direct contact with the RNA, the high ranking structures in run2 resulted in a closer fit to the crystal structure at the interaction surface than those in run1, in which C247 was only given ambiguous contacts to the 15.5K protein. The water refined structure with the best HADDOCK2.0 score in the best HADDOCK2.0 scoring cluster of run2 fits to the crystal structure at the protein-protein interface with an RMSD of 2.87 Å. According to the CAPRI (<u>c</u>ritical <u>a</u>ssessment of <u>pr</u>edicted <u>i</u>nteractions) quality criteria, this is an acceptable fit. However, in our case, information about the interaction surface was only obtainable on the 15.5K side, whereas information about hPrp31 was very scarce. Furthermore, docking attempts on such complex systems involving proteins and RNA are very uncommon and are very difficult to achieve. With rational deduction and

assumption, we successfully utilized the low resolution structural information from NMR studies and generated complex docking models in astonishingly good agreement with the crystal structure. It can, therefore, be concluded that our NMR studies and our docking attempts have been very successful.

In Figure 4.10, the VDW radii of the residues involved in the protein-protein interface in the docking model and the crystal structure are shown. In the docking model, the residues of 15.5K in close contact to hPrp31$^{188-332}$ are exclusively located on helices $\alpha$2 (N40, T43, K44 and N47), 3-10 (E64) and $\alpha$3 (H68, L71 and L72). In the crystal structure of hPrp31$^{78-333}$-15.5K-U4 5'-SL complex, the same residues of 15.5K are found to interact with hPrp31$^{78-333}$. There is only one additional contact in the crystal structure namely from K9 at the very N-terminus of 15.5K to F308 on the helix $\alpha$12 of hPrp31$^{78-333}$. The regions on hPrp31$^{188-332}$ contributing to the interaction surface in the docking model include helix $\alpha$9 (K243, M244, P245, C247 and N248), the C-terminal end of the helix $\alpha$10 (Q255) and the C-terminal end of the helix $\alpha$12 (R304). These regions are also found in the crystal structure to be the interaction surface. Additional contacts from T300, F308 and E310 located on the helix $\alpha$12 of hPrp31$^{78-333}$ in crystal structure are absent in the docking model.

The extensive hydrophobic interactions between the helix $\alpha$12 of hPrp31$^{78-333}$ and the helix $\alpha$3 of 15.5K in the crystal structure are underrepresented in the docking model. Overlapping the docking model and the crystal structure only on 15.5K protein results in a rotation of the Nop domain of about 10 degrees. Despite the rotation of the interaction surface in the model compared to the crystal structure, the important regions of both proteins involved in
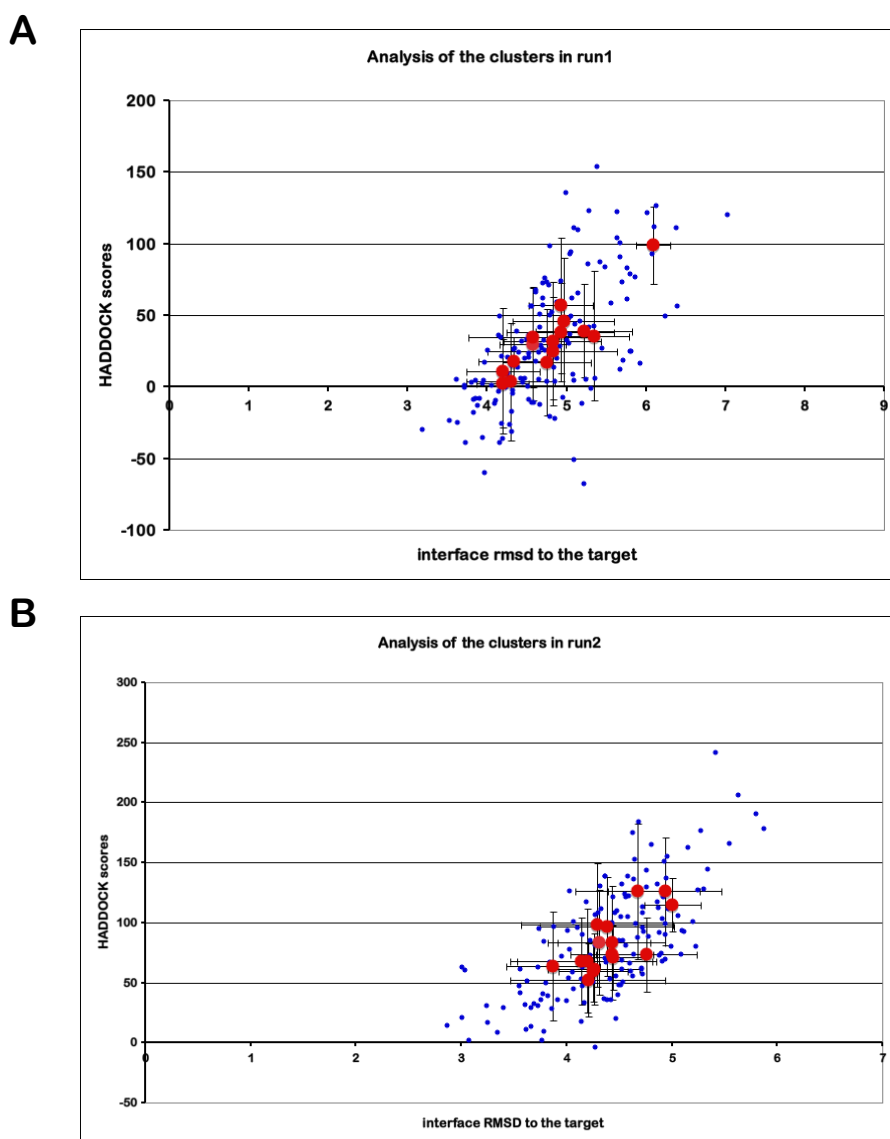
**Figure 4.9:** The RMSDs value between each HADDOCK2.0 generated docking structure and the crystal structure at the interface are plotted against the HADDOCK2.0 score for each docking structure. The structures are indicated in dark blue dots. The average HADDOCK2.0 scores of the clusters are indicated with large red dots. The horizontal error bars represent the standard deviations in the interface RMSDs of each cluster in the clusters. The vertical error bars are the standard deviations in the HADDOCK2.0 scores of each structure in the clusters. (A) Cluster analysis of run1, where Cys247 was not given ambiguous contacts to the RNA, but only to the 15.5K protein. (B) Cluster analysis of run2, where Cys247 was given ambiguous contacts to the stem II of the U4 5'-SL and the 15.5K protein.

**A**



**B**
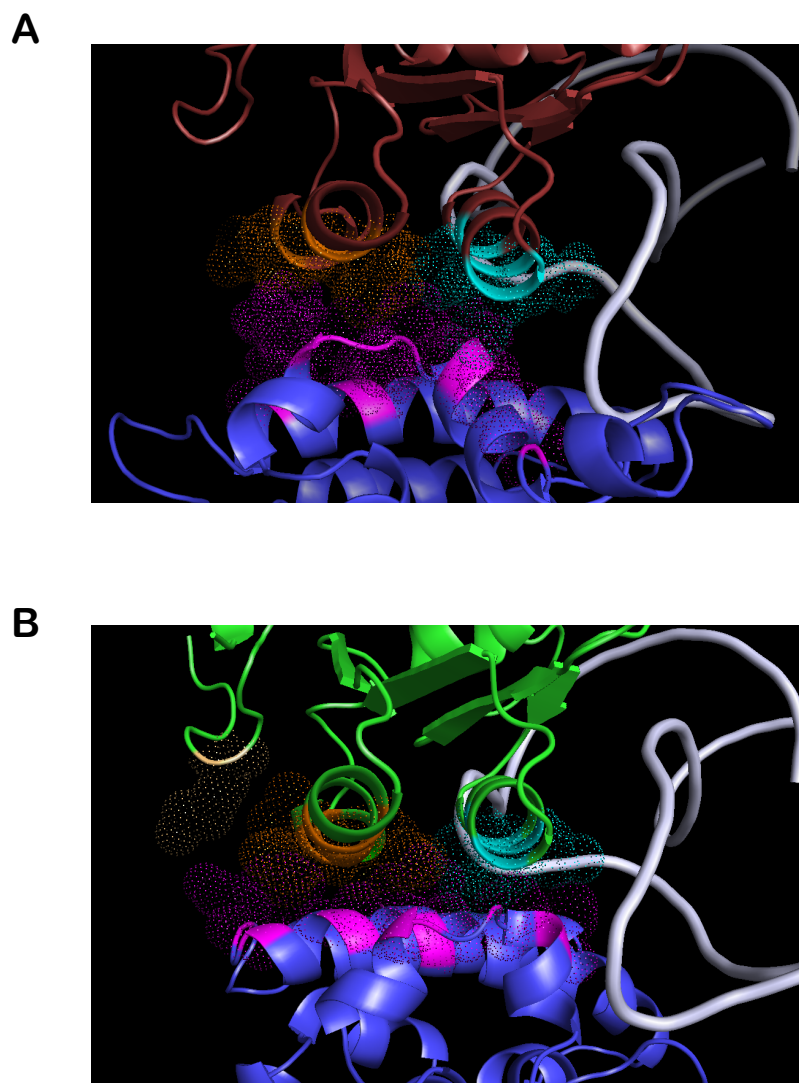


**Figure 4.10:** (A) Protein-protein interface in the docking model. 15.5K protein is coloured in brownish red, hPrp31$^{188-332}$ is coloured in blue and RNA in light grey. The VDW radii of the contacting residues in helix $\alpha 2$ of 15.5K are coloured in orange and those on helix $\alpha 3$ are in cyan. The contacting residues in hPrp31 are coloured in magenta. (B) Protein-protein interface in the crystal structure. 15.5K protein is coloured in green. Other colour schemes are the same as in (A). The additional contact from K9 at the very N-terminus of 15.5K is indicated in light orange.
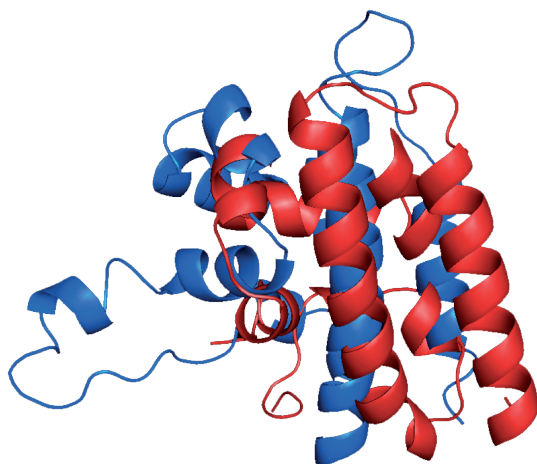
**Figure 4.11:** Superimposing the docking model and the crystal structure on 15.5K protein results in a rotation of the Nop domain to about 10 degrees. The Nop domain of the model is coloured in blue and that of the crystal structure in red.

the direct contacts are consistent between the docking model and the crystal structure (Figure 4.11).

**The charge complementary interaction surface**

In the docking model, the negatively charged RNA is sandwiched between extended positively charged patches on hPrp31$^{188-331}$ and on 15.5K consistent with the crystal structure (Figure 4.12). In addition, the regions of hPrp31$^{78-333}$ in the crystal structure interacting with 15.5K exhibits alternating positively and negatively charged surface patches, a pattern that is matched by a complementary set of patches on the 15.5K protein. In the docking model hPrp31$^{188-331}$, the interaction surface also shows clear charge complementarity. The interaction mainly involves negatively charged patches

on 15.5K with positive charged patches in hPrp31$^{188-331}$ . This electrostatic surface complementarity supports a favourable orientation of hPrp31$^{188-331}$ and hPrp31$^{78-333}$ with respect to the binary 15.5K-RNA complex in both the docking model and the crystal structure. This property of the interaction surface also supports the lock-and-key interaction between hPrp31 and the 15.5K-U4 5'-SL primary RNP, as the complex formation is favoured by long range electrostatic interaction in solution and the components can reorient themselves to match the complementarily charged surface patches without restructuring of the interaction surface.

### The Nop domain also acts as a molecular ruler in the crystal structure

Consistent with the docking model, all hPrp31 contacts with the stem II of the RNA in the crystal structure are mediated by the Nop domain, again suggesting that this motif alone is able to discriminate against an extended stem II. This was further verified by bandshift assays performed by Dr. Olexandr Dybkov by using wild-type (wt) and mutant U4 5'- SLs and a Nop-domain fusion protein (MBP-hPrp31$^{215-333}$). Like full-length hPrp31, the Nop domain did not bind to RNPs, in which stem II of the RNA was extended by a non- canonical U-U (Figure 4.13B, lanes 7-9) or by a canonical C-G base pair (lanes 10-12). These data confirm that the Nop domain is both required and sufficient for binding to the primary RNP and for decoding its structural specificity determinants.

An elongated stem II could reposition the pentaloop and thus disrupt the stacking interaction between H270 and A39 in both the model and the
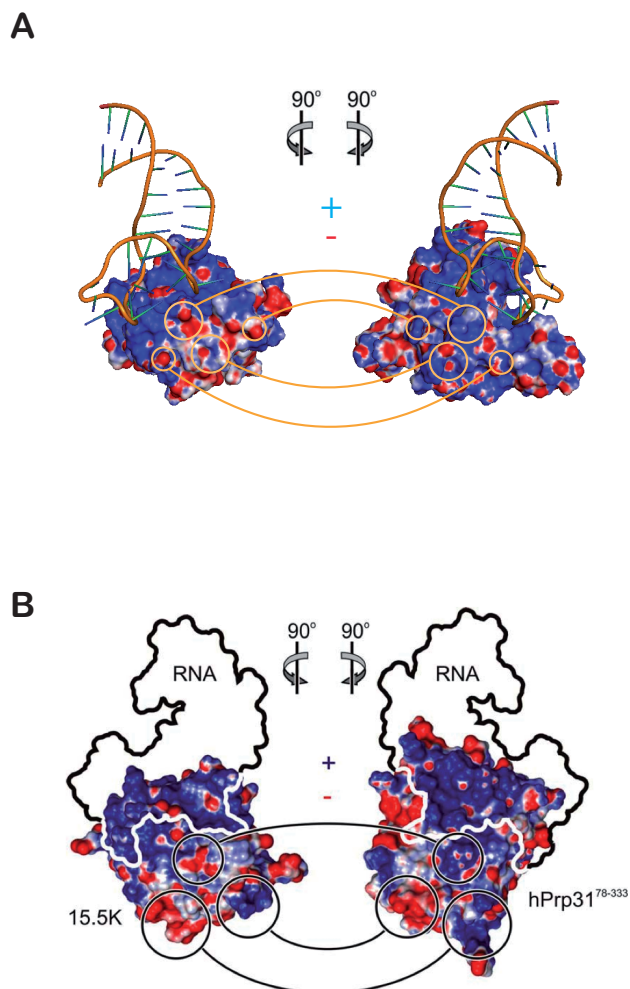
**Figure 4.12:** (A) View onto the interaction surface of 15.5K-RNA (left) and hPrp31$^{188-333}$ in the docking model (right; rotations relative to Figure 1.6 are indicated by the arrows) showing the electrostatic surface potentials of the protein components (blue - positive; red - negative). Potentials were calculated with PDB2PQR server (http://pdb2pqr.sourceforge.net/). Complementary surface patches are indicated by connected circles. RNA is indicated in orange. (B) View onto the interacting surfaces of 15.5K-RNA (left) and hPrp31$^{78-333}$ in the crystal structure (right). Potentials were calculated with APBS (http://apbs.sourceforge.net/). The position of the RNA is indicated by the outlines ((B) made by Dr. Markus Wahl).
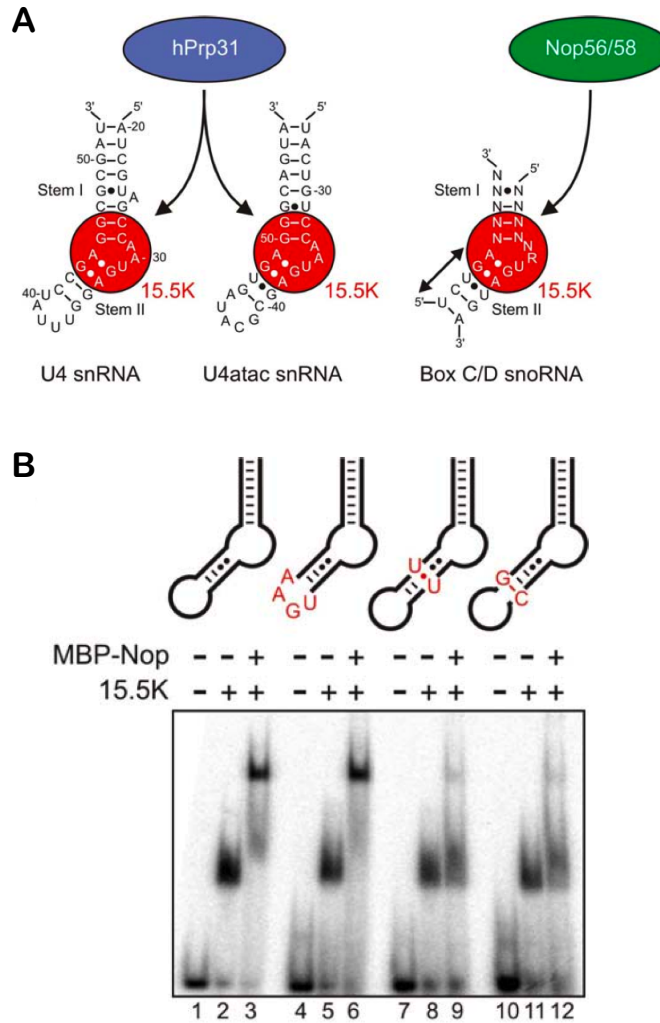
**Figure 4.13:** (A) Schematics of the 5'-SLs of U4 snRNA (left), U4atac snRNA (middle) and the K-turn region of box C/D snoRNAs (right). N any nucleotide; R purine. Binding of 15.5K and the secondary binding proteins are indicated. (B) Gel mobility shift assays monitoring the binding of a maltose binding protein-Nop domain fusion protein (MBP-hPrp31$^{215-333}$) to U4 5'-SL constructs. Lanes 1-3 wt RNA sequence; lanes 4-6 replacement of the pentaloop by a (UGAA)-tetraloop; lanes 7-9 addition of a U-U base pair to stem II following the sheared G-A pairs; lanes 10-12 addition of a C-G base pair at the terminus of stem II.
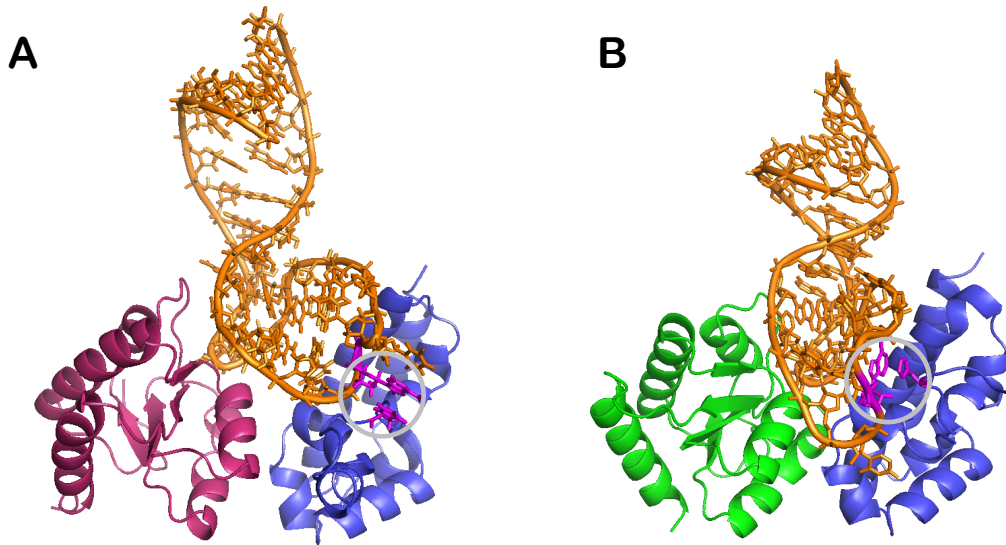
**Figure 4.14:** (A) In the docking model, 15.5K is coloured in raspberry red, the Nop domain of hPrp31$^{188-332}$ in blue and RNA in dark orange. H270 and A39 are indicated in magenta and included in the grey circle. (B) In the crystal structure, 15.5K is coloured in green, hPrp31$^{78-333}$ in blue and RNA in dark yellow. H270 and A39 are also indicated in magenta and included in the grey circle.

crystal structure (Figure 4.14). Loss of hPrp31 affinity to 15.5K complexes with elongated stem II RNAs could, therefore, arise due to the disruption of H270-A39 stacking.

This possibility was tested by converting H270 to an alanine or to a lysine (as found in Nop56/58). Loss or alteration of the H270 side chain resulted in a reduced affinity of hPrp31$^{H270A}$ and hPrp31$^{H270K}$ to the 15.5K-RNA 5'-SL complexes (Figure 4.15A, lanes 1-5). However, the mutants retained significant binding activity and discriminated strongly against long stem II constructs (Figure 4.15A, lanes 6-10), indicating that H270 is not required

for measuring the length of stem II. An extended stem II A-form duplex was modeled into the crystal structure by Dr. Markus Wahl. The additional base pair would lie in the stacking level occupied by A39 in the wt complex (Figure 4.15B). While A39 snugly fits next to helix $\alpha$10 of the Nop domain (Figure 4.15B), the helical twist would lead to a severe clash between an additional Watson-Crick base pair and helix $\alpha$10 (Figure 4.15C). In contrast, in a tetraloop-RNA, to which hPrp31 (134) and hPrp31$^{215-333}$ (Figure 4.13B, lanes 4-6) still bind, the nucleotides would be turned away from helix $\alpha$10 (data not shown). Thus, in the crystal structure the helix $\alpha$10 acts as a ruler for measuring the length of stem II by presenting a physical barrier to additional base pairs. As stated above, in the docking model this ruler effect is also clearly present. However, as the interface on the hPrp31$^{188-332}$ in the model is rotated with respect to the crystal structure, the physical barrier to additional base pairs is formed by the helix corresponding to helix $\alpha$11 in the crystal structure (Figure 4.15D).

This finding explains the stringent requirement on the structure of the RNA, namely the length of stem II, and the relaxed sequence requirements found in the U4 and U4atac RNAs. It could be deduced that a different mechanism must be at work in Nop56/58, in which the Nop domain is compatible with elongated stem II RNAs. Resolution of this issue will have to await the structural analysis of a box C/D snoRNP complex.
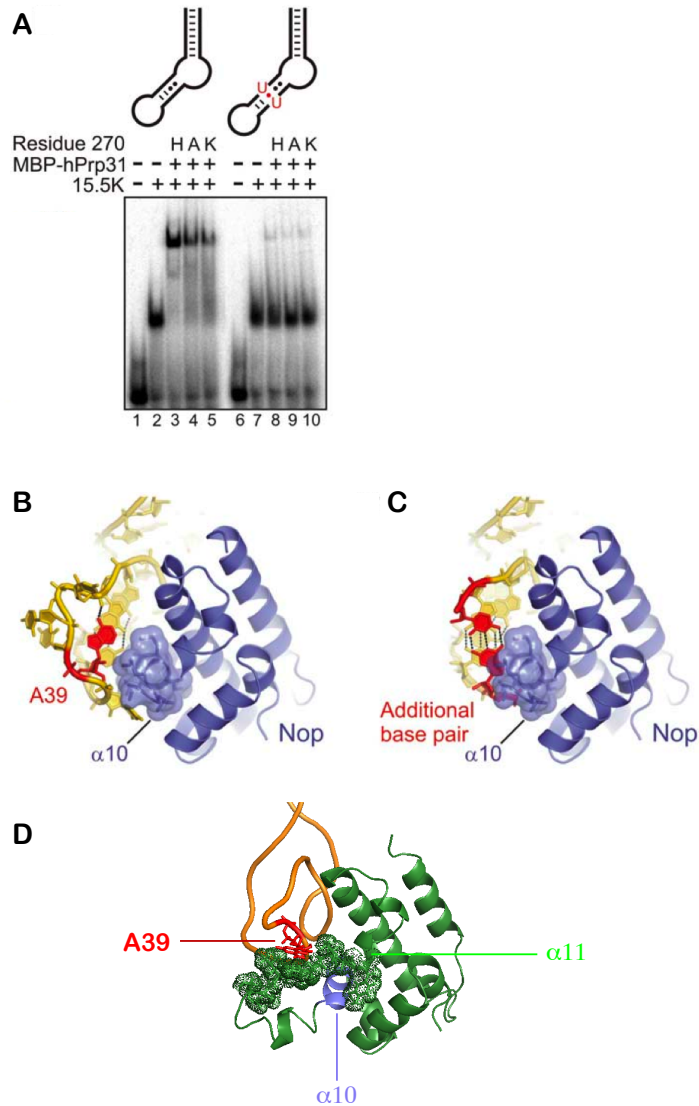
**Figure 4.15:** (A) Gel mobility shift assays monitoring the effects of converting H270 into an alanine or a lysine. Mutant hPrp31 proteins bind less strongly to a wt U4 5'-SL (lanes 1-5) but still discriminate against RNAs with a longer stem II is unaffected (lanes 6-10). (B,C) Ribbon plot of the Nop domain in the crystal structure in complex with the RNA pentaloop (B) and an RNA model with an extended stem II (C; pentaloop omitted). Helix $\alpha 10$ of the Nop domain is shown with a semi- transparent van-der-Waals surface. The lone A39 of the pentaloop (B) and the additional, modeled base pair (C) are in red ((A,B,C) made by Dr. Markus Wahl. (D) The Nop domain of the docking model and the RNA are coloured in green and orange repectively. The van-der-Waals surface of the corresponding helix $\alpha 11$ and the loop immediately before containing H270 is shown. A physical barrier can be seen to be presented by helix $\alpha 11$, which blocks the stem II from exceeding 2 base pairs. The corresponding helix $\alpha 10$ is coloured in blue.

**The structure of the Nop domain is highly conserved among the homolog proteins**

As mentioned above, the close fit between the Nop domains from hPrp31 and (AF) Nop5p demonstrates that the structure of the Nop domain is highly conserved among the homologous proteins of the Nop family. The sequence alignment shown in Figure 4.16 demonstrates high sequence similarity among hPrp31, Nop56/Nop58 and (AF) Nop5p proteins. Although at the present Nop56 and Nop58 proteins are not available in large amounts using conventional expression schemes for NMR or crystallography studies, the Nop domains of Nop56 and Nop58 proteins likely adopt very similar structures as the known Nop domain structures.

**Implications for RNP architecture and spliceosome activation**

The central domain of the 30S ribosomal subunit represents a paradigm for hierarchically assembled RNPs. There, a primary binding protein, S15, induces a conformational change in the RNA and thereby creates a novel RNA binding site for a secondary binding protein complex, S6/S18 (185). The present mode of hPrp31 interaction with a 15.5K-RNA complex in both our docking model and the crystal structure represents a fundamentally different hierarchical assembly strategy. The role of the assembly-initiating 15.5K is not restricted to inducing or stabilizing an hPrp31 binding site in the RNA; rather 15.5K itself provides approximately half of the contact surface for the Nop domain of hPrp31. By concomitantly interacting with both 15.5K and the RNA, the Nop domain reinforces the 15.5K-RNA interaction. The latter interaction is crucial for the transition from the spliceosomal B complex to

```
hPrp31    78   SEVMGPV.EAAPEYRVIVDANNLTVEIENELNIIHKFIRDKYSKRFPELESLVPNALDYI
xlPrp31   77   SEVMGPV.EAAPEYKVIVDANNLTVEIENELNIIHKFIRDKYSKRFPELESLVPNALDYI
dmPrp31   86   AEMLGSV.ESDPEYCLIVDANAIAVDIDNEISIVHKFTKEKYQKRFPELDSLIVGEIEYL
atPrp31   80   GAEKGTVLEDDPEYKLIVDCNQLSVDIENEIVIVHNFIKDKYKLKFQELESLVHHPIDYA
spPrp31  105   QAITGNI.EDDLEYHLIVDSNSIAMEIDDEILRLHRLVKEWYHDRFPELSSLVLNAFDYC
hNop58   147   YSRAKVKFNVNRVDNMIIQSISLLDDLDKELNNYIMRCREWYGWHFPELGKIISDNLTYC
hNop56   159   ............EDRYVVALVKALEEIDESINMLNEKLEDIRAVKESEIT..........
afNop5p   73
                                                           α1                        α2           α3
```

```
hPrp31   137   RTVKELGN..SLDKCKNNENLQQILTNATIMVVSVTASTTQGQQLSEEELERLDEACDMA
xlPrp31  136   RTVKELGN..NLDKCKNNENLQQILTNATIMVVSVTASTTQGQQLTDEELERIEEACDMA
dmPrp31  145   LAVKELGN..DLDQVKNNEKLQAILTQATIMIVSVTASTTQGTMLTPAEKAKIDEACEMA
atPrp31  140   CVVKKIGN..ETDLAL..VDLADLLPSAIIMVVSVTALTTKGSALPEDVLQKVLEACDRA
spPrp31  164   KTVSSLLN..DLDNSK...TKLSFLPSATVMVIATTATTTVGKPLPDEMIKNVKNCCEAI
hNop58   206   RLAQFIGNRRELNEDKLEKLEELTMDGAKAKAILDASRSSMGMDISAIDLINIESFSSRV
hNop56   218   ..................................................EKFEKKI
afNop5p  111
                                                    α4               α5
```

```
hPrp31   195   LELNASKHRIYEYVESRMSFIAPNLSIIICASTAAKIMGVAGGLTNLSKMPACNIMLLGA
xlPrp31  194   LELNQSKHRIYEYVESRMSFIAPNLSIIVGASTAAKIMGIAGGLTNLSKMPACNVMLLGA
dmPrp31  203   IELNNFKSKIYEYVESRMTFIAPNLSMIVGASTAAKLLGIAGGLSKLSKMPACNVQVLGA
atPrp31  196   LDLDSARKKVLEFVESKMGSIAPNLSAIVGSAVAAKLMGTAGGLSALAKMPACNVQVLGH
spPrp31  219   IEISEYRTQLYEYLQNRMMAIAPNVTVMVGELVGARLIAHAGSLLNLAKHAASTVQILGA
hNop58   262   VSLSEYRQSLHTYLRSKMSQVAPSLSALICEAVGARLIAHAGSLTNLAKYPASTVQILGA
hNop56   278   RELRELRRDVEREIEEVMEKIAPNMTELVGCAKVAAKLLERAGSMERLVRLPASKIQVIGA
afNop5p  118
                                     α6                    α7             α8               α9            α10
```

```
hPrp31   255   QRKTLSGFS.....STSVLPHTGYIYHSDIVQSLPPDLRRKAARLVAAKCTLAARVDSFH
xlPrp31  254   QRKTLTGFS.....STSVLPHTGYIYHSEIVQSLPSDLHRKAARLVSAKCTLASRVDSFH
dmPrp31  263   QKKTLSGFS.....QTQMLPHTGYVYYSQIVQDTAPDLRRKAARLVAAKSVLAARVDACH
atPrp31  256   KRKNLAGFS.....SATSQSRVGYLEQTEIYQSTPPGLQARAGRLVAAKSTLAARVDATR
spPrp31  279   EKALFRALK.....SRRDTPKYGLIYHASLVGQTSPKHKGKISRMLAAKTVLAIRYDAFG
hNop58   322   EKALFRALK.....TRGNTPKYGLIFHSTFIGRAAAKNKGRISRYLANKCSIASRIDCFS
hNop56   338   EKSLYKAFARMKKGKKAKIPKHGIIFLHPFIRTLPKAKRGKMARFLAAKLAIAAKIDYFR
afNop5p  178
                                           α11       3₁₀1         α12
```

```
hPrp31   310   ESTEGKVGYELKDEIERKFDKWQE
xlPrp31  309   ENPEGKIGYDLKEEIERKFDKWQE
dmPrp31  318   ESVHGEIGLRFKEDVEKKLDKLQE
atPrp31  311   GDPLGISGKAFREEIRKKIEKWQE
spPrp31  334   EYPDGSFGISARKEVERKIEKLLE
hNop58   377   EDSSSAMGVENRAKLEARLRTLED
hNop56   393   EVPTSVFGEKLREQVEERLSFYET
afNop5p  238   GEIDESLYESIRRRYEELRRK...
                           α13
```

**Figure 4.16:** Multiple sequence alignment of Prp31 and Nop proteins covering the region 78-333 of hPrp31 using Clustal method. h human; xl Xenopus laevis; dm Drosphila melanogaster; at Arabidopsis thaliana; sp Schizosaccharomyces pombe; af Archaeglobus fulgidus. Secondary structure elements of the hPrp31$^{78-333}$ fragment crystallized are shown schematically below the alignment. Lines correspond to disordered regions. The secondary structure elements are color-coded according to the domain structure of hPrp31$^{78-333}$ shown below the alignment. The Nop domain encompasses the region from C-terminus of $\alpha 6$ to $\alpha 13$. Red triangles above the alignment indicate residues of hPrp31$^{78-333}$ that are in contact with 15.5K in the present structure; yellow triangles indicate residues in contact with the RNA. Cyan triangles pinpoint the residues mutated in some Retinitis pigmentosa patients (figure made by Dr. Markus Wahl).

the C complex (19), during which spliceosome activation occurs. Thus, on the one side, hPrp31 may regulate the RNA-protein network at the U4 5'-SL and thereby facilitates disruption of the U4/U6 snRNA duplex.

## 4.3   Mutations in 15.5K-2 mutant do not directly disturb the ternary complex formation

In the TROSY-HSQC and the cross-saturation experiments carried out on hPrp31-15.5K-U4 5'-SL and hPrp31$^{78-333}$-15.5K-U4 5'-SL complexes, the C-terminal region of helix $\alpha 3$ of 15.5K showed neither chemical shift changes nor signal intensity changes. This is in excellent agreement with the crystal structure, in which a network of hydrophobic contacts ensues between the N-terminal end of helix $\alpha 3$ of 15.5K and the C-terminal end of helix $\alpha 12$. However, the mutant 15.5K-2, in which 4 mutations namely E74R/D75K/K76M/N77T were present at the C-terminal of helix $\alpha 3$, reported 4 fold reduction in binding affinity to hPrp31 in comparison to the wild type 15.5K in the previous pulldown assay (135). These residues were neither observed in our NMR studies nor in the crystal structure to be involved in direct contact with hPrp31. Therefore, we tested the possibility that the change in binding affinity of 15.5K-2 to hPrp31 could be caused by unfavourable changes in 15.5K-2 3D structure especially at the protein-protein interface. The structure of the 15.5K-2 converged after one round of RDC refinement. The resulting structure yielded a high quality (Q) factor of 0.063 between the back-calculated

and measured RDCs. The refined structure fitted to the crystal structure of the wild type with an RMSD of 1.26 Å for the backbone atoms, indicating that the 3D structure of the mutant 15.5K-2 is not significantly altered in comparison to the wild type 15.5K. This result is consistent with the finding from the bandshift assays, in which the effect of individual amino acid mutation within the 4 mutations of 15.5K-2 was tested. The bandshift assays demonstrated only slight reduction in binding affinity of the E74R, D75K, K76M and E74R/D75K mutants in comparison to the wild type 15.5K (see Figure 3.32 in the Results chapter). Surprisingly, even 15.5K-2 showed only weakly reduced binding affinity compared to the wild type 15.5K. The discrepancy between the results from the bandshift assays and the previous finding on 15.5K-2 using pulldown assay (135) could be due to the different salt concentration and experimental procedures employed in these two kinds of assays. In the previous pulldown assay, GST-hPrp31 fusion protein was coupled to glutathione-Sepharose and incubated with pre-formed 15.5K-U4 primary RNP in a buffer with NaCl concentration of 150 mM. The glutathione-Sepharose was then washed for 4 times at 150 mM NaCl concentration and once at 500 mM NaCl concentration. In this case, if the formation of hPrp31-15.5K-2-U4 5'-SLternary complex is mainly governed by kinetics, the washes could remove the formed hPrp31-15.5K-2-U4 5'-SL ternary complex and could drive the equilibrium to the free components namely hPrp31 and 15.5K-U4 5'-SL. This effect, however, is not present in the bandshift assays, where the ternary complexes were allowed to form at a salt concentration of 120 mM and were separated on a native 6% RNA gel.

Slight reduction in the binding affinity of the 15.5K mutants could, however, also be observed by bandshift assays. In the 4 mutants namely 15.5K-E74R, 15.5K-D75K, 15.5K-K76M and 15.5K-E74R/D75K, D75K mutation demonstrated the most noticeable reduction in binding affinity to MBP-hPrp31 and could account for all the reduction in affinity found in 15.5K-2. In the crystal structure, the D75 of 15.5K is within 4.56 Å to K243 of hPrp31 and is, therefore, able to form electrostatic interaction with K243. A mutant of D75K could change this favourable interaction with K243 to a charge repulsion and consequently lower the affinity between the two proteins.

Interestingly in the bandshift assays, the migration behavior appeared to be slightly different for each mutant of 15.5K as well as its complex with U4 5'-SL in agreement with the previous finding in the pulldown assay (135). This is likely to be caused by the charge differences between the mutants and the wild type 15.5K proteins.  The mutant protein with the highest positive charge, in this case mutant 15.5K-E74R/D75K, migrated most slowly (highest bands) and the wild type protein, which contains the lowest positive charge among these proteins, migrated fastest (lowest band).  The charge difference of the protein caused by changing of a single amino acid seems to be significant enough for the small 15.5K protein and could be detected in these bandshift assays.

## 4.4 Different regions of 15.5K are affected upon its binding to the cognate RNAs

As described in the introduction chapter, 15.5K protein binds to U14 and U3 box C/D snoRNAs at K-turns similar to those in U4 5'-SL, yet it recruits secondary binding partners other than hPrp31 in snoRNPs. In the case of U14 box C/D snoRNA (U14CD) and the box C'/D part of U3 box C/D snoRNA (U3CD), Nop56 and Nop58 protein are recruited, whereas in the primary RNP formed by 15.5K and the box B/C part of U3 box C/D snoRNA (U3BC) is recognized by hU3-55K protein (Figure 3.34 and Figure 3.35). HSQC experiments on 15.5K in complexes with U14CD, U3BC and U3CD snoRNA constructs showed an overall similar chemical shift pattern in comparison with the 15.5K-U4 5'-SL complex, but detailed chemical shift changes at different regions of 15.5K (Figure 3.38). As 15.5K interacts extensively with the K-turn, the difference in the sequence of the K-turn could be a main cause of the changes in the chemical shift patterns. The sequences of the K-turns in U4 snRNA, U14CD snoRNA as well as in the U3CD and U3BC snoRNA are shown in Figure 4.17. In U14CD RNA the nucleotides corresponding to A29 and A30 in U4 snRNA are changed to U and G. In the three snoRNA constructs, the sequence of the K-turn in U3BC snoRNA shows the highest sequence similarity to that in U4 snRNA, with one nucleotide corresponding to A29 changed to a G. In the case of U3CD snoRNA, a G resides the position of A29 in U4 snRNA and more importantly the position of the protruded out U31 in U4 snRNA, which forms the most extensive interactions with 15.5K in the crystal structure of 15.5K-U4 5'-SL,

is also changed to a G. The differences in the sequences of these RNAs are well reflected on the observations from the HSQC experiments. Helices $\alpha2$ and $\alpha4$ of 15.5K showed most chemical shift changes upon binding to U14 snoRNA in comparison to the HSQC of the 15.5K-U4 5'-SL complex. The changes could be mediated by a different bending angle of the K-turn in U14 snoRNA with respect to the K-turn in U4 snRNA or simply due to the differences in the RNA sequences. 15.5K in complex with the U3BC RNA showed very similar chemical shift pattern as that of the 15.5K-U4 5'-SL complex. The regions which showed slight changes in chemical shifts are the same regions as in the case of 15.5K-U14CD complex, which indicates that similar changes as in U14 snoRNA may be responsible for these chemical shift differences. The complex of 15.5K with the U3CD RNA showed the most changes in chemical shift in comparison to the 15.5K-U4 5'-SL complex. The changes are most pronounced at the helix 3-10 region. This is clearly correlated to the replacement of the U31 in U4 5'-SL to a G in the U3CD RNA. RDC refinement of 15.5K in complex with these snoRNA constructs are currently being carried out. The overall 3D structure of 15.5K in these complexes is not expected to change dramatically in comparison to the structure of 15.5K-U4 5'-SL, which further confirms that the selectivity on secondary binding proteins arises from the RNAs. In all these snoRNAs, the length of the stem II exceeds 2 base-pairs and are, therefore, unfavoured by the Nop domain of hPrp31 due to steric hindrance of the 'molecular ruler' (helix $\alpha10$). This should be the main criterion invoked by these different primary RNPs to discriminate among hPrp31, Nop56 and Nop58. In the case of 15.5K-U3CD complex, the exchange of the protruded uridine to a G could also differently
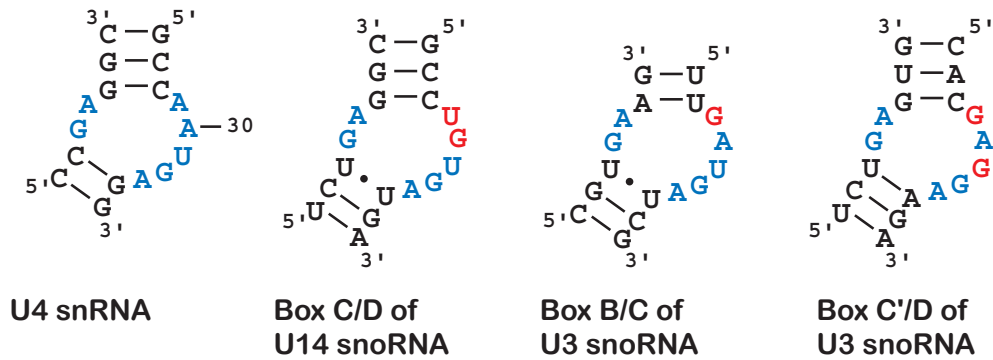
**Figure 4.17:** The sequence of the [5+2] loop region of the K-turns in the RNAs are coloured in blue. The deviations in the sequences of box C/D snoRNAs from U4 snRNA are indicated in red.

structure the 3-10 helix and consequently disturb the hydrophobic interaction with the helix $\alpha12$ of hPrp31. These changes may, however, be in favour of the binding of Nop56 and Nop58 proteins. 15.5K-U3BC complex recruits a very different secondary protein namely the hU3-55K, which contains a WD 40 repeat. The interactions of hU3-55K to the primary RNP is likely to be on a different surface of 15.5K. According to the previous finding (135), helix $\alpha4$ of 15.5K seems to be essential for the binding of hU3-55K. This suggests that a very different selection criterion could be employed in this RNP.

# Chapter 5

# Conclusion and outlook

In this thesis work, NMR spectroscopy was employed to carry out investigations on 15.5K associated RNPs. The goals of these investigations include: (1) mapping the interaction surface between 15.5K and hPrp31 in the hPrp31-15.5K-U4 5'-SL ternary complex; (2) gathering detailed information on the binding property of the Nop domain of hPrp31 and finding the criteria used for selecting the secondary binding partners in different 15.5K associated RNPs; (3) creating a 3D model for the hPrp31-15.5K-U4 5'-SL ternary complex; (4) finding the cause of the reduction in binding affinity of 15.5K-2 mutant to hPrp31 and (5) comparing the structure of 15.5K in complexes with U14, U3 box C/D snoRNAs and U4 snRNA to further address the origin of the selectivity on secondary binding partners in these RNPs. The outcomes of this work have met all these goals and have given insights into fundamental questions related to the incorporation of hPrp31 into the spliceosome, which is known to play an important role in the spliceosome activation. The outcomes also provided a general understanding of RNP

assembly and architecture.

The following conclusions can be made from the results obtained through this work:

(1) Free 15.5K protein undergoes structural rearrangement at the RNA binding site upon association with the U4 5'-SL demonstrating an induced-fit in the protein.

(2) hPrp31 and hPrp31$^{78-333}$ proteins form direct contacts with helices $\alpha 2$ and $\alpha 3$ of 15.5K in the ternary complexes.

(3) Nop domain of hPrp31 forms interaction surfaces with 15.5K and U4 5'-SL RNA of comparable size and is, therefore, a *bona fide* RNP recognition motif. It is both required and sufficient for binding to the primary RNP.

(4) The 3D model of hPrp31$^{188-332}$-15.5K-U4 5'-SL complex generated by comparative modelling and HADDOCK2.0 demonstrated the key features of the crystal structure of hPrp31$^{188-332}$-15.5K-U4 5'-SL. The interaction surfaces of 15.5K and hPrp31 proteins in the docking model are consistent with those in the crystal structure. The docking model hPrp31$^{188-332}$ contacts the primary RNP exclusively via its Nop domain, suggesting that this domain is the most crucial RNP interacting module in hPrp31. Binding of the hPrp31 Nop domain to 15.5K and the immediate K-turn resembles a lock-and-key type interaction in both the docking model and the crystal structure. In hPrp31, the Nop domain acts as a molecular ruler in the docking model as well as in the crystal structure to discriminate the primary RNPs. This explains the stringent structural requirements of the RNA for hPrp31 binding.

(5) The structure of the Nop domain in the Nop family of proteins is highly conserved.

(6) The additional observed interactions between hPrp31 and the upstream region of stem I of U4 snRNA are exclusively mediated by the C-terminus of hPrp31.

(7) C-terminal end of helix $\alpha2$ of 15.5K protein is not involved in direct contact with the hPrp31. The 15.5K-2 mutant, therefore, does not affect the thermodynamics of the binding of hPrp31 to the primary RNP directly. According to RDC refinement, the structure of 15.5K-2 mutant does not differ significantly from the wild type protein.

(8) Compared with 15.5K-U4 5'-SL complex, various regions of 15.5K may adopt different structures upon binding to U14 and U3 box C/D snoRNAs. However, the overall 3D structure of 15.5K is found not to differ in different RNPs. The selectivity on secondary binding partners mainly arises from the RNA and the secondary binding proteins themselves.

The hPrp31-15.5K-U4 5'-SL ($\sim 81$ kDa) ternary complex, which was only available in a very dilut concentration ($\sim 0.1$ mM) is among the largest and most challenging systems of its kind ever studied using NMR spectroscopy. This thesis work has demonstrated that NMR spectroscopy is capable of studying large protein-RNA complexes. In combination with biochemical findings, novel comparative modeling and structural calculation methods, the low resolution structural information obtained from NMR experiments lead to the generation of a reasonable 3D model for the whole protein-RNA complex. With small size complexes, high resolution structural information could be obtained by NMR, whereas large size complexes could also be studied by NMR spectroscopy with sacrifices on the resolution of the information.

**Outlook**    In the frame of this thesis work, only 15.5K was labeled and observed by NMR. Therefore, the information obtained from the investigations is mostly confined to the 15.5K protein. It is of great importance to carry out similar NMR investigations on U4 5'-SL to obtain information on the structural changes and dynamics of the RNA upon binding to 15.5K and hPrp31. Labeling of a U4 5'-SL construct and subsequent NMR experiments on this RNA with and without the presence of unlabeled 15.5K are being carried out in the group of Dr. Teresa Carlomagno. The outcome of these experiments will give more insights into the proposed induced-fit mechanism in the RNA. Unlabeled hPrp31 will be titrated to the 15.5K-U4 5'-SL primary RNP and the structural change of the labeled U4 5'-SL RNA will be monitored. To further understand the selectivity of the primary RNPs for secondary binding proteins, the structures of box C/D snoRNPs will be studied using NMR spectroscopy. Currently human Nop58 and Nop56 proteins are not obtainable in reasonable quantities for NMR studies. However, the homologous box C/D snoRNP from the archea species *Pyrococus furiosus* is being investigated in the group of Dr. Teresa Carlomagno. In this RNP, which carries out 2'-O-methylation of pre-rRNAs, L7ae plays the role of 15.5K, Nop5p is the homolog of hPrp31 and fibrillarin is the putative methyltransferase. The crystal structure of Nop5p in complex with fibrillarin is available (100), but detailed protein-protein interactions between Nop5p and L7ae as well as protein-RNA interactions between Nop5p and the box C/D RNA are unknown. The current NMR studies aim to define the interaction surface between L7ae and Nop5p as well as Nop5p and the snoRNA using the same experiments (HSQC and cross-saturation experiments) employed for investi-

gating the hPrp31-15.5K-U4 5'-SL system. Combining the knowledge from the Nop5p-fibrillarin crystal structure, it would then be very plausible to create a 3D model using HADDOCK2.0 for the entire snoRNP.

RDC refinement of the structure of 15.5K-2 mutant is still ongoing. More sets of RDCs from backbone N-C and C-H$_N$ will be used to validate the refinement structure obtain from only one set of N-H RDCs. RDC refinement is also being carried out on 15.5K in complexes with U14 and U3 box C/D snoRNA constructs to reveal the structural changes of 15.5K in its various associated RNPs.

# Bibliography

[1] E. M. Purcell, H. C. Torrey, and R. V. Pound, Resonance Absorption by Nuclear Magnetic Moments in a Solid, *Physical Review* **69**: 37–38 (1946).

[2] F. Bloch, W. W. Hansen, and M. Packard, Nuclear Induction, *Physical Review* **69**: 127 (1946).

[3] P. Mansfield, Snapshot Magnetic Resonance Imaging (Nobel Lecture), *Angewandte Chemie International Edition* **43**: 5456–5464 (2004).

[4] J. Cavanagh, W. J. Fairbrother, A. G. Palmer III, and N. J. Skelton, Protein NMR Spectroscopy: principles and practice, Academic press, San Diego, USA (1996).

[5] R. R. Ernst, Nuclear Magnetic Double Resonance with an Incoherent Radio-Frequency Field, *Journal of Chemical Physics* **45**: 3845–47 (1966).

[6] J. Jeener, in Ampere International Summer School II, Basko Polje, Yugoslavia (1971).

[7] H. Stark and R. Luhrmann, Cryo-electron microscopy of spliceosomal components, *Annu Rev Biophys Biomol Struct* **35**: 435–57 (2006).

[8] M. A. Kennedy, G. T. Montelione, C. H. Arrowsmith, and J. L. Markley, Role for NMR in structural genomics, *Journal of structural and functional genomics* **2**: 155–169 (2002).

[9] K. Pervushin, R. Riek, G. Wider, and K. Wuthrich, Attenuated T-2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution, *Proceedings of the National Academy of Sciences of the United States of America* **94**: 12366–12371 (1997).

[10] C. Fernandez and G. Wider, TROSY in NMR studies of the structure and function of large biological macromolecules, *Current Opinion in Structural Biology* **13**: 570–580 (2003).

[11] D. M. LeMaster, Deuteration in protein proton magnetic resonance, *Methods Enzymol* **177**: 23–43 (1989).

[12] A. P. Hansen, A. M. Petros, A. P. Mazar, T. M. Pederson, A. Rueter, and S. W. Fesik, A practical method for uniform isotopic labeling of recombinant proteins in mammalian cells, *Biochemistry* **31**: 12713–8 (1992).

[13] S. Grzesiek, J. Anglister, H. Ren, and A. Bax, 13C line narrowing by 2H decoupling in 2H/13C/15N- enriched proteins. Application to triple

resonance 4D J connectivity of sequential amides, *J. Am. Chem. Soc.* **115**: 4369–4370 (1993).

[14] T. Kigawa, Y. Muto, and S. Yokoyama, Cell-free synthesis and amino acid-selective stable isotope labeling of proteins for NMR analysis, *J Biomol NMR* **6**: 129–34 (1995).

[15] K. H. Gardner and L. E. Kay, The use of 2H, 13C, 15N multidimensional NMR to study the structure and dynamics of proteins, *Annu Rev Biophys Biomol Struct* **27**: 357–406 (1998).

[16] M. Kainosho, T. Torizawa, Y. Iwashita, T. Terauchi, A. Mei Ono, and P. Guntert, Optimal isotope labelling for NMR protein structure determinations, *Nature* **440**: 52–7 (2006).

[17] G. M. Clore and C. D. Schwieters, Theoretical and computational advances in biomolecular NMR spectroscopy, *Curr Opin Struct Biol* **12**: 146–53 (2002).

[18] R. R. Ernst, Nobel Lecture. Nuclear magnetic resonance Fourier transform spectroscopy, *Biosci Rep* **12**: 143–87 (1992).

[19] R. B. Russell, F. Alber, P. Aloy, F. P. Davis, D. Korkin, M. Pichaud, M. Topf, and A. Sali, A structural perspective on protein-protein interactions, *Curr Opin Struct Biol* **14**: 313–24 (2004).

[20] M. J. Moore, Gene expression. When the junk isn't junk, *Nature* **379**: 402–3 (1996).

[21] R. Reed, The organization of 3' splice-site sequences in mammalian introns, *Genes Dev* **3**: 2113–23 (1989).

[22] R. Reed and T. Maniatis, The role of the mammalian branchpoint sequence in pre-mRNA splicing, *Genes Dev* **2**: 1268–76 (1988).

[23] M. J. Moore and P. A. Sharp, Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing, *Nature* **365**: 364–8 (1993).

[24] M. R. Green, Biochemical mechanisms of constitutive and regulated pre-mRNA splicing, *Annu Rev Cell Biol* **7**: 559–99 (1991).

[25] T. W. Nilsen, RNA-RNA interactions in nuclear pre-mRNA splicing, in R. W. Simons, Grunberg-Manago, and M., editors, RNA structure and function, 279–307, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (1998).

[26] C. Burge, T. Tuschl, and P. Sharp, Splicing of Precursors to mRNAs by the Spliceosomes, in C. T. . A. J. Gesteland, R.F., editor, The RNA world II, 525–560, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. (1999).

[27] D. A. Brow and C. Guthrie, Spliceosomal RNA U6 is remarkably conserved from yeast to mammals, *Nature* **334**: 213–8 (1988).

[28] C. Guthrie and B. Patterson, Spliceosomal snRNAs, *Annu Rev Genet* **22**: 387–419 (1988).

[29] W. Y. Tarn and J. A. Steitz, A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro, *Cell* **84**: 801–11 (1996).

[30] I. W. Mattaj, Cap trimethylation of U snRNA is cytoplasmic and dependent on U snRNP protein binding, *Cell* **46**: 905–11 (1986).

[31] K. T. Tycowski, Z. H. You, P. J. Graham, and J. A. Steitz, Modification of U6 spliceosomal RNA is guided by other small RNAs, *Mol Cell* **2**: 629–38 (1998).

[32] J. Rinke, B. Appel, M. Digweed, and R. Lhrmann, Localization of a base-paired interaction between small nuclear RNAs U4 and U6 in intact U4/U6 ribonucleoprotein particles by psoralen cross-linking, *J. Mol. Biol.* **185**: 721–31 (1985).

[33] P. Bringmann, B. Appel, J. Rinke, R. Reuter, H. Theissen, and R. Lhrmann, Evidence for the existence of snRNAs U4 and U6 in a single ribonucleoprotein complex and for their association by intermolecular base pairing, *EMBO J.* **3**: 1357–63 (1984).

[34] N. M. Fast, A. J. Roger, C. A. Richardson, and W. F. Doolittle, U2 and U6 snRNA genes in the microsporidian Nosema locustae: evidence for a functional spliceosome, *Nucleic Acids Res* **26**: 3202–7 (1998).

[35] H. D. Madhani and C. Guthrie, A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome, *Cell* **71**: 803–17 (1992).

[36] A. A. Patel and J. A. Steitz, Splicing double: insights from the second spliceosome, *Nat Rev Mol Cell Biol* **4**: 960–70 (2003).

[37] T. Achsel, H. Brahms, B. Kastner, A. Bachi, M. Wilm, and R. Luhrmann, A doughnut-shaped heteromer of human Sm-like proteins binds to the 3'-end of U6 snRNA, thereby facilitating U4/U6 duplex formation in vitro, *Embo J* **18**: 5789–802 (1999).

[38] T. Achsel, H. Stark, and R. Lhrmann, The Sm domain is an ancient RNA-binding motif with oligo(U) specificity, *Proc Natl Acad Sci U S A* **98**: 3685–9 (2001).

[39] C. Kambach, S. Walke, R. Young, J. M. Avis, E. de la Fortelle, V. A. Raker, R. Lhrmann, J. Li, and K. Nagai, Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs, *Cell* **96**: 375–87 (1999).

[40] B. Sraphin, Sm and Sm-like proteins belong to a large family: identification of proteins of the U6 as well as the U1, U2, U4 and U5 snRNPs, *EMBO J.* **14**: 2089–98 (1995).

[41] J. D. Beggs, Lsm proteins and RNA processing, *Biochem Soc Trans* **33**: 433–8 (2005).

[42] V. P. Vidal, L. Verdone, A. E. Mayes, and J. D. Beggs, Characterization of U6 snRNA-protein interactions, *RNA* **5**: 1470–81 (1999).

[43] C. L. Will and R. Luhrmann, Protein functions in pre-mRNA splicing, *Curr Opin Cell Biol* **9**: 320–8 (1997).

[44] C. L. Will and R. Luhrmann, Spliceosome structure and function, in R. F. Gesteland, T. R. Cech, Atkins, and J. F., editors, *The RNA world*, 369–400, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 3rd edition (2006).

[45] B. Kastner and R. Lhrmann, Purification of U small nuclear ribonucleoprotein particles, *Methods Mol Biol* **118**: 289–98 (1999).

[46] J. R. Patton and T. Pederson, The Mr 70,000 protein of the U1 small nuclear ribonucleoprotein particle binds to the 5' stem-loop of U1 RNA and interacts with Sm domain proteins, *Proc Natl Acad Sci U S A* **85**: 747–51 (1988).

[47] H. Urlaub, K. Hartmuth, S. Kostka, G. Grelle, and R. Luhrmann, A general approach for identification of RNA-protein cross-linking sites within native human spliceosomal small nuclear ribonucleoproteins (snRNPs). Analysis of RNA-protein contacts in native U1 and U4/U6.U5 snRNPs, *J Biol Chem* **275**: 41458–68 (2000).

[48] G. Varani and K. Nagai, RNA recognition by RNP proteins during RNA processing, *Annu Rev Biophys Biomol Struct* **27**: 407–45 (1998).

[49] S. E. Behrens, K. Tyc, B. Kastner, J. Reichelt, and R. Lhrmann, Small nuclear ribonucleoprotein (RNP) U2 contains numerous additional proteins and has a bipartite RNP structure under splicing conditions, *Mol Cell Biol* **13**: 307–19 (1993).

[50] R. Brosi, H. P. Hauri, and A. Kramer, Separation of splicing factor SF3

into two components and purification of SF3a activity, *J Biol Chem* **268**: 17640–6 (1993).

[51] C. L. Will and R. Luhrmann, Spliceosomal UsnRNP biogenesis, structure and function, *Curr Opin Cell Biol* **13**: 290–301 (2001).

[52] M. Bach, G. Winkelmann, and R. Luhrmann, 20S small nuclear ribonucleoprotein U5 shows a surprisingly complex protein composition, *Proc Natl Acad Sci U S A* **86**: 6038–42 (1989).

[53] S. E. Behrens and R. Luhrmann, Immunoaffinity purification of a [U4/U6.U5] tri-snRNP from human cells, *Genes Dev* **5**: 1439–52 (1991).

[54] J. P. Staley and C. Guthrie, Mechanical devices of the spliceosome: motors, clocks, springs, and things, *Cell* **92**: 315–26 (1998).

[55] B. Schwer, A new twist on RNA helicases: DExH/D box proteins as RNPases, *Nat Struct Biol* **8**: 113–6 (2001).

[56] C. Bartels, C. Klatt, R. Luhrmann, and P. Fabrizio, The ribosomal translocase homologue Snu114p is involved in unwinding U4/U6 RNA during activation of the spliceosome, *EMBO Rep* **3**: 875–80 (2002).

[57] J. F. Kuhn, E. J. Tran, and E. S. Maxwell, Archaeal ribosomal protein L7 is a functional homolog of the eukaryotic 15.5kD/Snu13p snoRNP core protein, *Nucleic Acids Res* **30**: 931–41 (2002).

[58] P. L. Raghunathan and C. Guthrie, RNA unwinding in U4/U6 snRNPs

requires ATP hydrolysis and the DEIH-box splicing factor Brr2, *Curr Biol* **8**: 847–55 (1998).

[59] M. Bell, S. Schreiner, A. Damianov, R. Reddy, and A. Bindereif, p110, a novel human U6 snRNP protein and U4/U6 snRNP recycling factor, *EMBO J* **21**: 2724–35 (2002).

[60] R. Reed, Initial splice-site recognition and pairing during pre-mRNA splicing, *Curr Opin Genet Dev* **6**: 215–20 (1996).

[61] C. L. Will, S. Rumpler, J. Klein Gunnewiek, W. J. van Venrooij, and R. Luhrmann, In vitro reconstitution of mammalian U1 snRNPs active in splicing: the U1-C protein enhances the formation of early (E) spliceosomal complexes, *Nucleic Acids Res* **24**: 4614–23 (1996).

[62] J. A. Berglund, M. Rosbash, and S. C. Schultz, Crystal structure of a model branchpoint-U2 snRNA duplex containing bulged adenosines, *RNA* **7**: 682–91 (2001).

[63] R. Reed, Mechanisms of fidelity in pre-mRNA splicing, *Curr Opin Cell Biol* **12**: 340–5 (2000).

[64] M. Bennett, S. Michaud, J. Kingston, and R. Reed, Protein components specifically associated with prespliceosome and spliceosome complexes, *Genes Dev* **6**: 1986–2000 (1992).

[65] J. Valcarcel, R. K. Gaur, R. Singh, and M. R. Green, Interaction of U2AF65 RS region with pre-mRNA branch point and promotion of base pairing with U2 snRNA [corrected], *Science* **273**: 1706–9 (1996).

[66] W. Hong, M. Bennett, Y. Xiao, R. Feld Kramer, C. Wang, and R. Reed, Association of U2 snRNP with the spliceosomal complex E, *Nucleic Acids Res* **25**: 354–61 (1997).

[67] O. Dybkov, C. L. Will, J. Deckert, N. Behzadnia, K. Hartmuth, and R. Luhrmann, U2 snRNA-protein contacts in purified human 17S U2 snRNPs and in spliceosomal A and B complexes, *Mol Cell Biol* **26**: 2803–16 (2006).

[68] O. V. Makarova, E. M. Makarov, and R. Luhrmann, The 65 and 110 kDa SR-related proteins of the U4/U6.U5 tri-snRNP are essential for the assembly of mature spliceosomes, *Embo J* **20**: 2553–63 (2001).

[69] A. I. Lamond, M. M. Konarska, P. J. Grabowski, and P. A. Sharp, Spliceosome assembly involves the binding and release of U4 small nuclear ribonucleoprotein, *Proc Natl Acad Sci U S A* **85**: 411–5 (1988).

[70] S. L. Yean and R. J. Lin, U4 small nuclear RNA dissociates from a yeast spliceosome and does not participate in the subsequent splicing reaction, *Mol. Cell Biol.* **11**: 5571–7 (1991).

[71] J. A. Wu and J. L. Manley, Base pairing between U2 and U6 snRNAs is necessary for splicing of a mammalian pre-mRNA, *Nature* **352**: 818–21 (1991).

[72] D. Xu, S. Nouraini, D. Field, S. J. Tang, and J. D. Friesen, An RNA-dependent ATPase associated with U2/U6 snRNAs in pre-mRNA splicing, *Nature* **381**: 709–13 (1996).

[73] J. S. Sun and J. L. Manley, A novel U2-U6 snRNA structure is necessary for mammalian mRNA splicing, *Genes Dev* **9**: 843–54 (1995).

[74] R. Reed and L. Palandjian, Spliceosome assembly, in I. A. K. (Ed.), editor, Eukaryotic pre-mRNA processing, 103–129, Oxford IRL Press, Oxford (1997).

[75] B. Ruskin and M. R. Green, An RNA processing activity that debranches RNA lariats, *Science* **229**: 135–40 (1985).

[76] H. D. Madhani and C. Guthrie, Dynamic RNA-RNA interactions in the spliceosome, *Annu Rev Genet* **28**: 1–26 (1994).

[77] T. W. Nilsen, The spliceosome: the most complex macromolecular machine in the cell?, *Bioessays* **25**: 1147–9 (2003).

[78] C. Hashimoto and J. A. Steitz, U4 and U6 RNAs coexist in a single small nuclear ribonucleoprotein particle, *Nucleic Acids Res* **12**: 3283–93 (1984).

[79] H. Stark, P. Dube, R. Luhrmann, and B. Kastner, Arrangement of RNA and proteins in the spliceosomal U1 small nuclear ribonucleoprotein particle, *Nature* **409**: 539–42 (2001).

[80] M. M. Golas, B. Sander, C. L. Will, R. Luhrmann, and H. Stark, Molecular architecture of the multiprotein splicing factor SF3b, *Science* **300**: 980–4 (2003).

[81] C. L. Will, H. Urlaub, T. Achsel, M. Gentzel, M. Wilm, and R. Luhrmann, Characterization of novel SF3b and 17S U2 snRNP pro-

teins, including a human Prp5p homologue and an SF3b DEAD-box protein, *Embo J* **21**: 4978–88 (2002).

[82] P. Vankan, C. McGuigan, and I. W. Mattaj, Roles of U4 and U6 snR-NAs in the assembly of splicing complexes, *Embo J* **11**: 335–43 (1992).

[83] C. Wersig and A. Bindereif, Reconstitution of functional mammalian U4 small nuclear ribonucleoprotein: Sm protein binding is not essential for splicing in vitro, *Mol. Cell Biol.* **12**: 1460–8 (1992).

[84] C. Wersig and A. Bindereif, Conserved domains of human U4 snRNA required for snRNP and spliceosome assembly, *Nucleic Acids Res.* **18**: 6223–9 (1990).

[85] S. Nottrott, K. Hartmuth, P. Fabrizio, H. Urlaub, I. Vidovic, R. Ficner, and R. Luhrmann, Functional interaction of a novel 15.5kD [U4/U6.U5] tri-snRNP protein with the 5' stem-loop of U4 snRNA, *Embo J* **18**: 6119–33 (1999).

[86] I. Vidovic, S. Nottrott, K. Hartmuth, R. Lhrmann, and R. Ficner, Crystal structure of the spliceosomal 15.5kD protein bound to a U4 snRNA fragment, *Mol Cell* **6**: 1331–42 (2000).

[87] E. V. Koonin, P. Bork, and C. Sander, A novel RNA-binding motif in omnipotent suppressors of translation termination, ribosomal proteins and a ribosome modification enzyme?, *Nucleic Acids Res* **22**: 2166–7 (1994).

[88] T. Moore, Y. Zhang, M. O. Fenley, and H. Li, Molecular basis of box

C/D RNA-protein interactions; cocrystal structure of archaeal L7Ae and a box C/D RNA, *Structure* **12**: 807–18 (2004).

[89] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz, The complete atomic structure of the large ribosomal subunit at 2.4 A resolution, *Science* **289**: 905–20 (2000).

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

[91] B. T. Wimberly, D. E. Brodersen, J. Clemons, W. M., R. J. Morgan-Warren, A. P. Carter, C. Vonrhein, T. Hartsch, and V. Ramakrishnan, Structure of the 30S ribosomal subunit, *Nature* **407**: 327–39 (2000).

[92] D. J. Klein, T. M. Schmeing, P. B. Moore, and T. A. Steitz, The kink-turn: a new RNA secondary structure motif, *Embo J* **20**: 4214–21 (2001).

[93] A. K. Wozniak, S. Nottrott, E. Kuhn-Holsken, G. F. Schroder, H. Grub-muller, R. Luhrmann, C. A. Seidel, and F. Oesterhelt, Detecting protein-induced folding of the U4 snRNA kink-turn by single-molecule multiparameter FRET measurements, *Rna* **11**: 1545–54 (2005).

[94] V. Cojocaru, S. Nottrott, R. Klement, and T. M. Jovin, The snRNP 15.5K protein folds its cognate K-turn RNA: a combined theoretical and biochemical study, *Rna* **11**: 197–209 (2005).

[95] S. Nottrott, H. Urlaub, and R. Lhrmann, Hierarchical, clustered protein interactions with U4/U6 snRNA: a biochemical role for U4/U6 proteins, *EMBO J* **21**: 5527–38 (2002).

[96] J. Lauber, G. Plessel, S. Prehn, C. L. Will, P. Fabrizio, K. Groning, W. S. Lane, and R. Luhrmann, The human U4/U6 snRNP contains 60 and 90kD proteins that are structurally homologous to the yeast splicing factors Prp4p and Prp3p, *Rna* **3**: 926–41 (1997).

[97] T. Gautier, T. Berges, D. Tollervey, and E. Hurt, Nucleolar KKE/D repeat proteins Nop56p and Nop58p interact with Nop1p and are required for ribosome biogenesis, *Mol Cell Biol* **17**: 7088–98 (1997).

[98] E. N. Vithana, L. Abu-Safieh, M. J. Allen, A. Carey, M. Papaioannou, C. Chakarova, M. Al-Maghtheh, N. D. Ebenezer, C. Willis, A. T. Moore, A. C. Bird, D. M. Hunt, and S. S. Bhattacharya, A human homolog of yeast pre-mRNA splicing gene, PRP31, underlies autosomal dominant retinitis pigmentosa on chromosome 19q13.4 (RP11), *Mol Cell* **8**: 375–81 (2001).

[99] O. V. Makarova, E. M. Makarov, S. Liu, H. P. Vornlocher, and R. Lhrmann, Protein 61K, encoded by a gene (PRPF31) linked to autosomal dominant retinitis pigmentosa, is required for U4/U6.U5 tri-snRNP formation and pre-mRNA splicing, *EMBO J* **21**: 1148–57 (2002).

[100] M. Aittaleb, R. Rashid, Q. Chen, J. R. Palmer, C. J. Daniels, and H. Li,

Structure and function of archaeal box C/D sRNP core proteins, *Nat Struct Biol* **10**: 256–63 (2003).

[101] E. M. Makarov, O. V. Makarova, T. Achsel, and R. Luhrmann, The human homologue of the yeast splicing factor prp6p contains multiple TPR elements and is stably associated with the U5 snRNP via protein-protein interactions, *J Mol Biol* **298**: 567–75 (2000).

[102] N. Schaffert, M. Hossbach, R. Heintzmann, T. Achsel, and R. Lhrmann, RNAi knockdown of hPrp31 leads to an accumulation of U4/U6 di-snRNPs in Cajal bodies, *EMBO J* **23**: 3000–9 (2004).

[103] E. C. Deery, E. N. Vithana, R. J. Newbold, V. A. Gallon, S. S. Bhattacharya, M. J. Warren, D. M. Hunt, and S. E. Wilkie, Disease mechanism for retinitis pigmentosa (RP11) caused by mutations in the splicing factor gene PRPF31, *Hum Mol Genet* **11**: 3209–19 (2002).

[104] S. E. Wilkie, K. J. Morris, S. S. Bhattacharya, M. J. Warren, and D. M. Hunt, A study of the nuclear trafficking of the splicing factor protein PRPF31 linked to autosomal dominant retinitis pigmentosa (ADRP), *Biochim Biophys Acta* **1762**: 304–11 (2006).

[105] N. A. Faustino and T. A. Cooper, Pre-mRNA splicing and human disease, *Genes Dev* **17**: 419–37 (2003).

[106] A. Kennan, A. Aherne, and P. Humphries, Light in retinitis pigmentosa, *Trends Genet* **21**: 103–10 (2005).

[107] J. Venema and D. Tollervey, Ribosome synthesis in Saccharomyces cerevisiae, *Annu Rev Genet* **33**: 261–311 (1999).

[108] S. Granneman and S. J. Baserga, Ribosome biogenesis: of knobs and RNA processing, *Exp Cell Res* **296**: 43–50 (2004).

[109] J. Rouquette, V. Choesmel, and P. E. Gleizes, Nuclear export and cytoplasmic processing of precursors to the 40S ribosomal subunits in mammalian cells, *Embo J* **24**: 2862–72 (2005).

[110] K. V. Hadjiolova, M. Nicoloso, S. Mazan, A. A. Hadjiolov, and J. P. Bachellerie, Alternative pre-rRNA processing pathways in human cells and their alteration by cycloheximide inhibition of protein synthesis, *Eur J Biochem* **212**: 211–5 (1993).

[111] W. Decatur and M. J. Fournier, RNA-guided nucleotide modification of ribosomal and other RNAs, *J Biol Chem* **278**: 665–698 (2003).

[112] Z. Kiss-Laszlo, Y. Henry, J. P. Bachellerie, M. Caizergues-Ferrer, and T. Kiss, Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs, *Cell* **85**: 1077–88 (1996).

[113] T. Kiss, Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions, *Cell* **109**: 145–148 (2002).

[114] K. T. Tycowski, M. D. Shu, and J. A. Steitz, Requirement for intron-encoded U22 small nucleolar RNA in 18S ribosomal RNA maturation, *Science* **266**: 1558–61 (1994).

[115] W. Q. Liang and M. J. Fournier, U14 base-pairs with 18S rRNA: a novel snoRNA interaction required for rRNA processing, *Genes Dev* **9**: 2433–43 (1995).

[116] B. A. Peculis, The sequence of the 5' end of the U8 small nucleolar RNA is critical for 5.8S and 28S rRNA maturation, *Mol Cell Biol* **17**: 3702–13 (1997).

[117] A. V. Borovjagin and S. A. Gerbi, U3 small nucleolar RNA is essential for cleavage at sites 1, 2 and 3 in pre-rRNA and determines which rRNA processing pathway is taken in Xenopus oocytes, *J Mol Biol* **286**: 1347–63 (1999).

[118] Y. T. Yu, E. C. Scharl, C. M. Smith, and J. A. Steitz, The RNA World, Cold Spring Harbor Laboratory Press, New York, 2nd edition (1999).

[119] P. Ganot, B. E. Jady, M. L. Bortolin, X. Darzacq, and T. Kiss, Nucleolar factors direct the 2'-O-ribose methylation and pseudouridylation of U6 spliceosomal RNA, *Mol Cell Biol* **19**: 6906–17 (1999).

[120] T. Schimmang, D. Tollervey, H. Kern, R. Frank, and E. C. Hurt, A yeast nucleolar protein related to mammalian fibrillarin is associated with small nucleolar RNA and is essential for viability, *Embo J* **8**: 4015–24 (1989).

[121] P. Wu, J. S. Brockenbrough, A. C. Metcalfe, S. Chen, and J. P. Aris, Nop5p is a small nucleolar ribonucleoprotein component required for pre-18 S rRNA processing in yeast, *J Biol Chem* **273**: 16453–63 (1998).

[122] D. L. Lafontaine and D. Tollervey, Nop58p is a common component of the box C+D snoRNPs that is required for snoRNA stability, *Rna* **5**: 455–67 (1999).

[123] D. L. Lafontaine and D. Tollervey, Synthesis and assembly of the box
      C+D small nucleolar RNPs, *Mol Cell Biol* **20**: 2650–9 (2000).

[124] N. J. Watkins, V. Segault, B. Charpentier, S. Nottrott, P. Fabrizio,
      A. Bachi, M. Wilm, M. Rosbash, C. Branlant, and R. Lhrmann, A
      common core RNP structure shared between the small nucleoar box
      C/D RNPs and the spliceosomal U4 snRNP [In Process Citation], *Cell*
      **103**: 457–66 (2000).

[125] A. Niewmierzycka and S. Clarke, S-Adenosylmethionine-dependent
      methylation in Saccharomyces cerevisiae. Identification of a novel pro-
      tein arginine methyltransferase, *J Biol Chem* **274**: 814–24 (1999).

[126] A. D. Omer, T. M. Lowe, A. G. Russell, H. Ebhardt, S. R. Eddy, and
      P. P. Dennis, Homologs of small nucleolar RNAs in Archaea, *Science*
      **288**: 517–22 (2000).

[127] N. J. Watkins, A. Dickmanns, and R. Luhrmann, Conserved stem II of
      the box C/D motif is essential for nucleolar localization and is required,
      along with the 15.5K protein, for the hierarchical assembly of the box
      C/D snoRNP, *Mol Cell Biol* **22**: 8342–52 (2002).

[128] N. M. Cahill, K. Friend, W. Speckmann, Z. H. Li, R. M. Terns, M. P.
      Terns, and J. A. Steitz, Site-specific cross-linking analyses reveal an
      asymmetric protein distribution for a box C/D snoRNP, *Embo J* **21**:
      3816–28 (2002).

[129] J. Venema, H. R. Vos, A. W. Faber, W. J. van Venrooij, and H. A.
      Raue, Yeast Rrp9p is an evolutionarily conserved U3 snoRNP protein

essential for early pre-rRNA processing cleavages and requires box C for its association, *Rna* **6**: 1660–71 (2000).

[130] A. A. Lukowiak, S. Granneman, S. A. Mattox, W. A. Speckmann, K. Jones, H. Pluk, W. J. Venrooij, R. M. Terns, and M. P. Terns, Interaction of the U3-55k protein with U3 snoRNA is mediated by the box B/C motif of U3 and the WD repeats of U3-55k, *Nucleic Acids Res* **28**: 3462–71 (2000).

[131] S. Granneman, G. J. Pruijn, W. Horstman, W. J. van Venrooij, R. Lhrmann, and N. J. Watkins, The hU3-55K protein requires 15.5K binding to the box B/C motif as well as flanking RNA elements for its association with the U3 small nucleolar RNA in Vitro, *J Biol Chem* **277**: 48490–500 (2002).

[132] L. B. Szewczak, J. S. Gabrielsen, S. J. Degregorio, S. A. Strobel, and J. A. Steitz, Molecular basis for RNA kink-turn recognition by the h15.5K small RNP protein, *Rna* **11**: 1407–19 (2005).

[133] C. Schneider, C. L. Will, O. V. Makarova, E. M. Makarov, and R. Luhrmann, Human U4/U6.U5 and U4atac/U6atac.U5 tri-snRNPs exhibit similar protein compositions, *Mol Cell Biol* **22**: 3219–29 (2002).

[134] A. Schultz, S. Nottrott, K. Hartmuth, and R. Luhrmann, RNA structural requirements for the association of the spliceosomal hPrp31 protein with the U4 and U4atac small nuclear ribonucleoproteins, *J Biol Chem* **281**: 28278–86 (2006).

[135] A. Schultz, S. Nottrott, N. J. Watkins, and R. Luhrmann, Protein-protein and protein-RNA contacts both contribute to the 15.5K-mediated assembly of the U4/U6 snRNP and the box C/D snoRNPs, *Mol Cell Biol* **26**: 5146–54 (2006).

[136] E. Khn-Hlsken, C. Lenz, B. Sander, R. Lhrmann, and H. Urlaub, Complete MALDI-ToF MS analysis of cross-linked peptide-RNA oligonucleotides derived from nonlabeled UV-irradiated ribonucleoprotein particles, *RNA* 1915–1930 (2006).

[137] S. Liu, P. Li, O. Dybkov, S. Nottrott, K. Hartmuth, R. Luhrmann, T. Carlomagno, and M. C. Wahl, Binding of the human Prp31 Nop domain to a composite RNA-protein platform in U4 snRNP, *Science* **316**: 115–20 (2007).

[138] F. W. Studier, Protein production by auto-induction in high density shaking cultures, *Protein Expr Purif* **41**: 207–34 (2005).

[139] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, UCSF Chimera–a visualization system for exploratory research and analysis, *J Comput Chem* **25**: 1605–12 (2004).

[140] J. D. Thompson, D. G. Higgins, and T. J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res* **22**: 4673–80 (1994).

[141] A. D. van Dijk and A. M. J. J. Bonvin, Solvated docking: introducing water into the modelling of biomolecular complexes, *Bioinformatics* **22**: 2340–2347 (2006).

[142] R. Koradi, M. Billeter, and K. Wuthrich, MOLMOL: a program for display and analysis of macromolecular structures, *J Mol Graph* **14**: 51–55 (1996).

[143] M. Zweckstetter and A. Bax, Prediction of sterically induced alignment in a dilute liquid crystalline phase: Aid to protein structure determination by NMR, *Journal of the American Chemical Society* **122**: 3791–3792 (2000).

[144] T. J. Dolinsky, J. E. Nielsen, J. A. McCammon, and N. A. Baker, PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations, *Nucleic Acids Res* **32**: W665–7 (2004).

[145] B. Rost, G. Yachdav, and J. Liu, The PredictProtein server, *Nucleic Acids Res* **32**: W321–6 (2004).

[146] N. Guex and M. C. Peitsch, SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling, *Electrophoresis* **18**: 2714–23 (1997).

[147] T. Schwede, J. Kopp, N. Guex, and M. C. Peitsch, SWISS-MODEL: An automated protein homology-modeling server, *Nucleic Acids Res* **31**: 3381–5 (2003).

[148] C. D. Schwieters, J. J. Kuszewski, N. Tjandra, and G. M. Clore, The

Xplor-NIH NMR molecular structure determination package, *Journal of Magnetic Resonance* **160**: 65–73 (2003).

[149] D. J. Craik and J. A. Wilce, Studies of protein-ligand interactions by NMR, *Methods Mol Biol* **60**: 195–232 (1997).

[150] H. Oschkinat, C. Griesinger, P. J. Kraulis, O. W. Sorensen, R. R. Ernst, A. M. Gronenborn, and G. M. Clore, Three-dimensional NMR spectroscopy of a protein in solution, *Nature* **332**: 374–6 (1988).

[151] G. W. Vuister and R. Boelens, 3-Dimensional J-Resolved Nmr-Spectroscopy, *Journal of Magnetic Resonance* **73**: 328–333 (1987).

[152] L. E. Kay, G. M. Clore, A. Bax, and A. M. Gronenborn, Four-dimensional heteronuclear triple-resonance NMR spectroscopy of interleukin-1 beta in solution, *Science* **249**: 411–4 (1990).

[153] M. Sattler, J. Schleucher, and C. Griesinger, Heteronuclear multidimensional NMR experiments for the structure determination of protein in solution emplying pulsed field gradients, *Progress in Nuclear Magnetic Resonance Spectroscopy* **34**: 93–158 (1999).

[154] J. Keeler, Understanding NMR Spectroscopy, Wiley, Chichester, England (2005).

[155] S. Grzesiek and A. BAX, Improved 3d Triple-Resonance Nmr Techniques Applied to a 31-Kda Protein, *Journal of Magnetic Resonance* **96**: 432–440 (1992).

[156] S. Grzesiek and A. BAX, Correlating Backbone Amide and Side-Chain Resonances in Larger Proteins by Multiple Relayed Triple Resonance Nmr, *Journal of the American Chemical Society* **114**: 6291–6293 (1992).

[157] A. Bax and M. Ikura, An efficient 3D NMR technique for correlating the proton and 15N backbone amide resonances with the alpha-carbon of the preceding residue in uniformly 15N/13C enriched proteins, *J Biomol NMR* **1**: 99–104 (1991).

[158] M. Wittekind and L. Mueller, Hncacb, a High-Sensitivity 3d Nmr Experiment to Correlate Amide-Proton and Nitrogen Resonances with the Alpha-Carbon and Beta-Carbon Resonances in Proteins, *Journal of Magnetic Resonance Series B* **101**: 201–205 (1993).

[159] M. Czisch and R. Boelens, Sensitivity enhancement in the TROSY experiment, *Journal of Magnetic Resonance* **134**: 158–160 (1998).

[160] H. Takahashi, T. Nakanishi, K. Kami, Y. Arata, and I. Shimada, A novel NMR method for determining the interfaces of large protein-protein complexes, *Nat Struct Biol* **7**: 220–3 (2000).

[161] E. Kupce and G. Wagner, Wideband homonuclear decoupling in protein spectra, *Journal of magnetic Resonance B* **109**: 329–333 (1995).

[162] M. A. Marti-Renom, A. C. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali, Comparative protein structure modeling of genes and genomes, *Annu Rev Biophys Biomol Struct* **29**: 291–325 (2000).

[163] J. Bajorath, R. Stenkamp, and A. Aruffo, Knowledge-based model building of proteins: Concepts and examples, *Protein Sci.* **2**: 1798–1810 (1994).

[164] T. L. Blundell, B. L. Sibanda, M. J. Sternberg, and J. M. Thornton, Knowledge-based prediction of protein structures and the design of novel molecules, *Nature* **326**: 347–52 (1987).

[165] P. Koehl and M. Levitt, A brighter future for protein structure prediction, *Nature* **6**: 108–111 (1999).

[166] A. M. Lesk and C. Chothia, How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins, *J Mol Biol* **136**: 225–70 (1980).

[167] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res* **25**: 3389–402 (1997).

[168] C. D. Schwieters, J. J. Kuszewski, and G. M. Clore, Using Xplor-NIH for NMR molecular structure determination, *Progress in Nuclear Magnetic Resonance Spectroscopy* **48**: 47–62 (2006).

[169] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, Electrostatics of nanosystems: application to microtubules and the ribosome, *Proc Natl Acad Sci U S A* **98**: 10037–41 (2001).

[170] C. Dominguez, R. Boelens, and A. M. J. J. Bonvin, HADDOCK. A

protein-protein docking approach based on biochemical or biophysical information, *J. Am. Chem. Soc.* **125**: 1731–1737 (2003).

[171] A. D. J. van Dijk, R. Boelens, and A. M. J. J. Bonvin, Data-driven docking for the study of biomolecular complexes, *FEBS Journal* **272**: 293–312 (2005).

[172] A. D. J. van Dijk, S. J. de Vries, C. Dominguez, H. Chen, H.-X. Zhou, and A. M. J. J. Bonvin, Data-driven docking: HADDOCK's adventures in CAPRI, *Proteins: Structure, Function, and Bioinformatics* **60**: 232–238 (2005).

[173] A. D. J. van Dijk, R. Kaptein, R. Boelens, and A. M. J. J. Bonvin, Combining NMR relaxation with chemical shift perturbation data to drive protein-protein docking, *Journal of Biomolecular NMR* **34**: 237–244 (2006).

[174] M. van Dijk, A. D. van Dijk, V. Hsu, R. Boelens, and A. M. Bonvin, Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility, *Nucleic Acids Res* **34**: 3317–25 (2006).

[175] E. Brunner, Residual Dipolar Couplings in Protein NMR, *Concepts in magnetic resonance* **13**: 238–259 (2001).

[176] F. Kramer, M. V. Deshmukh, H. Kessler, and S. J. Glaser, Residual dipolar coupling constants: An elementary derivation of key equations, *Concepts in Magnetic Resonance Part A* **21A**: 10–21 (2004).

[177] A. Bax, Weak alignment offers new opportunities in NMR structure

determination of proteins and nucleic acids, *Abstracts of Papers of the American Chemical Society* **226**: U241–U241 (2003).

[178] J. H. Prestegard, H. M. al Hashimi, and J. R. Tolman, NMR structures of biomolecules using field oriented media and residual dipolar couplings, *Q Rev Biophys* **33**: 371–424 (2000).

[179] J. R. Tolman, J. M. Flanagan, M. A. Kennedy, and J. H. Prestegard, Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution, *Proc Natl Acad Sci U S A* **92**: 9279–83 (1995).

[180] N. Tjandra and A. Bax, Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium, *Science* **278**: 1111–4 (1997).

[181] M. Ottiger, F. Delaglio, and A. Bax, Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra, *Journal of Magnetic Resonance* **131**: 373–378 (1998).

[182] M. Zweckstetter and A. Bax, Characterization of molecular alignment in aqueous suspensions of Pf1 bacteriophage, *Journal of Biomolecular NMR* **20**: 365–377 (2001).

[183] J. R. Williamson, Induced fit in RNA-protein recognition, *Nature Structural Biology* **7**: 834–837 (2000).

[184] C. G. Burd and G. Dreyfuss, Conserved structures and diversity of functions of RNA-binding proteins, *Science* **265**: 615–21 (1994).

[185] S. C. Agalarov, G. Sridhar Prasad, P. M. Funke, C. D. Stout, and J. R. Williamson, Structure of the S15,S6,S18-rRNA complex: assembly of the 30S ribosome central domain, *Science* **288**: 107–13 (2000).

# Appendix A

# NMR Data

In this appendix, the chemical shifts of the assigned backbone resonances of the 70%, $^{15}$N, $^{13}C$- labelled 15.5K in complex with U4 5'-SL-24nt and the intensity changes of the amide protons of 15.5K observed using cross-satruation experiment for complexes hPrp31-15.5K-U4 5'-SL-33nt are tabulated in Table A.1 and Table A.2. The chemical shifts of the assigned backbone resonances of the $^{15}$N, $^{13}C$- labelled 15.5K-2 mutant in complex with U4 5'-SL-33nt and the measured N-H RDCs, which were used as 'sani' restraints in the RDC refinement protocol are documented in Table A.3 and Table A.4.

## A.1   The backbone assignment of 15.5K

**Table A.1:** The backbone resonances of 70%, $^{15}$N, $^{13}C$- labelled 15.5 protein were assigned with the help from TROSY version 3D experiments including HNCA, HNCACB, HNCOCACB, HNCOCA and HNCO as well as the 2D HSQC experiment. These experiments were all measured at 308 K at pH 7.6 on an Bruker 900 MHz spectrometer equipped with a cryo-probe.

| Number | Residue | H | N | $C_\alpha$ | $C_\beta$ | CO |
|--------|---------|-------|---------|--------|--------|---------|
| 1 | M | - | - | - | - | |
| 2 | T | 8.081 | 114.631 | 58.864 | 66.805 | 181.292 |
| 3 | E | 8.246 | 122.832 | 53.376 | 27.349 | 173.225 |

<div align="center">continued on next page</div>

| Number | Residue | H | N | $C_\alpha$ | $C_\beta$ | CO |
|---|---|---|---|---|---|---|
| 4 | A | 8.124 | 124.245 | 49.421 | 16.267 | 174.689 |
| 5 | D | 8.185 | 119.167 | 51.377 | 37.945 | 173.074 |
| 6 | V | 7.745 | 118.430 | 57.938 | 30.265 | 172.421 |
| 7 | N | 8.487 | 125.806 | 48.814 | 36.247 | 172.360 |
| 8 | P | - | - | - | - | - |
| 9 | K | 8.249 | 115.699 | 53.741 | 29.095 | 173.414 |
| 10 | A | 7.507 | 122.968 | 46.577 | 71.653 | 170.018 |
| 11 | Y | 6.615 | 123.421 | 50.915 | 38.778 | 169.738 |
| 12 | P | - | - | - | - | - |
| 13 | L | 8.389 | 124.096 | 50.041 | 40.982 | 173.278 |
| 14 | A | 8.604 | 129.279 | 49.433 | 16.940 | 173.946 |
| 15 | D | 8.172 | 119.853 | 49.444 | 37.470 | 175.014 |
| 16 | A | - | - | - | - | - |
| 17 | H | 7.837 | 118.443 | 55.995 | 27.043 | 175.653 |
| 18 | L | 8.776 | 121.902 | 54.190 | 36.719 | 175.776 |
| 19 | T | 8.619 | 115.335 | 66.255 | 64.758 | - |
| 20 | K | 7.118 | 120.118 | 56.807 | 28.406 | 175.604 |
| 21 | K | 7.083 | 116.806 | 56.032 | 29.430 | 177.165 |
| 22 | L | 8.832 | 120.433 | 54.872 | 39.643 | 175.392 |
| 23 | L | 8.612 | 117.500 | 54.921 | 35.786 | 177.166 |
| 24 | D | 7.952 | 120.046 | 54.350 | 37.476 | 175.659 |
| 25 | L | 7.673 | 121.303 | 54.283 | 37.302 | 177.158 |
| 26 | V | 8.938 | 123.538 | 64.280 | 28.507 | 175.062 |
| 27 | Q | 7.860 | 120.328 | 56.339 | 25.223 | 176.786 |
| 28 | Q | 8.088 | 118.601 | 56.326 | 25.916 | 176.457 |
| 29 | S | 8.962 | 116.785 | 60.179 | 59.062 | 173.619 |
| 30 | C | 7.940 | 122.892 | 59.353 | 23.074 | 175.622 |
| 31 | N | 7.590 | 119.507 | 52.765 | 34.883 | 173.560 |
| 32 | Y | 7.663 | 116.565 | 55.085 | 35.214 | 171.090 |
| 33 | K | 7.647 | 114.955 | 54.085 | 25.053 | 173.436 |
| 34 | Q | 8.334 | 117.159 | 51.074 | 27.521 | 170.655 |
| 35 | L | 7.155 | 119.493 | 50.570 | 42.471 | 173.606 |
| 36 | R | 9.015 | 124.932 | 49.864 | 28.502 | 171.771 |
| 37 | K | 9.232 | 123.816 | 52.986 | 30.680 | 170.103 |
| 38 | G | 7.456 | 107.911 | 40.346 | 0.000 | 171.924 |
| 39 | A | 10.115 | 127.457 | 52.773 | 14.700 | 178.088 |
| 40 | N | 10.193 | 122.587 | 54.378 | 34.627 | 176.087 |
| 41 | E | 8.151 | 119.595 | 58.472 | 26.962 | 176.836 |
| 42 | A | 9.307 | 122.250 | 52.451 | 72.241 | 178.239 |
| 43 | T | 8.441 | 117.710 | 64.919 | 65.817 | 173.204 |
| 44 | K | 7.624 | 120.587 | 57.672 | 28.630 | 176.804 |
| 45 | T | 7.994 | 107.452 | 61.925 | 66.014 | 172.954 |
| 46 | L | 7.180 | 122.816 | 53.882 | 38.962 | 177.278 |
| 47 | N | 8.028 | 118.621 | 52.948 | 34.855 | 174.831 |
| 48 | R | 7.870 | 115.647 | 53.209 | 27.944 | 173.663 |
| 49 | G | 7.468 | 108.161 | 43.917 | 0.000 | 173.206 |
| 50 | I | 7.817 | 109.149 | 57.212 | 35.551 | 171.973 |
| 51 | S | 7.957 | 116.529 | 55.233 | 61.788 | 173.764 |
| 52 | E | 9.773 | 126.429 | 55.088 | 27.371 | 175.695 |
| 53 | F | 7.037 | 109.814 | 54.122 | 36.131 | 169.122 |
| 54 | I | 7.851 | 118.443 | 54.053 | 36.446 | 172.003 |
| 55 | V | 7.871 | 122.960 | 56.981 | 30.251 | 171.876 |
| 56 | M | 8.248 | 123.565 | 50.210 | 34.544 | 169.506 |
| 57 | A | 7.941 | 121.558 | 46.904 | 17.963 | 173.993 |
| 58 | A | 9.314 | 121.064 | 50.416 | 17.052 | 173.719 |

| Number | Residue | H | N | $C_\alpha$ | $C_\beta$ | CO |
|--------|---------|-------|---------|--------|--------|---------|
| 59 | D | 8.033 | 113.005 | 49.366 | 35.243 | 174.063 |
| 60 | A | 6.917 | 122.118 | 50.774 | 16.246 | 172.730 |
| 61 | E | 8.464 | 122.186 | 49.439 | 30.007 | 171.498 |
| 62 | P | - | - | - | - | - |
| 63 | L | 8.528 | 124.198 | 54.477 | 38.294 | 174.888 |
| 64 | E | 8.703 | 112.871 | 57.387 | 25.977 | 174.872 |
| 65 | I | 7.249 | 114.902 | 61.842 | 34.614 | 173.635 |
| 66 | I | 7.707 | 107.568 | 57.468 | 36.576 | 174.793 |
| 67 | L | 7.095 | 118.776 | 54.445 | 36.870 | 173.990 |
| 68 | H | 8.245 | 117.201 | 50.834 | 38.490 | 174.108 |
| 69 | L | 6.998 | 119.200 | 56.680 | 29.596 | 174.356 |
| 70 | P | - | - | - | - | - |
| 71 | L | 5.993 | 112.515 | 54.338 | 38.389 | 176.705 |
| 72 | L | 7.439 | 121.278 | 54.351 | 39.116 | 176.521 |
| 73 | C | 8.466 | 116.635 | 62.086 | 24.228 | 175.293 |
| 74 | E | 8.015 | 119.192 | 56.309 | 25.632 | 177.080 |
| 75 | D | 7.588 | 119.129 | 54.244 | 38.477 | 175.007 |
| 76 | K | 7.676 | 114.173 | 51.460 | 28.777 | 172.864 |
| 77 | N | 7.872 | 117.693 | 51.380 | 34.149 | 171.094 |
| 78 | V | 8.275 | 121.373 | 56.439 | 31.952 | 171.509 |
| 79 | P | - | - | - | - | - |
| 80 | Y | 6.791 | 114.410 | 51.441 | 38.848 | 169.098 |
| 81 | V | 7.716 | 114.084 | 56.616 | 31.762 | 181.971 |
| 82 | F | 8.520 | 121.873 | 53.397 | 37.806 | 173.424 |
| 83 | V | 8.943 | 110.941 | 55.478 | 30.548 | 174.065 |
| 84 | R | 8.638 | 118.355 | 57.364 | 27.984 | 173.828 |
| 85 | S | 7.490 | 106.881 | 51.984 | 61.335 | 172.766 |
| 86 | K | 11.321 | 132.941 | 54.663 | 28.640 | 176.202 |
| 87 | Q | 8.461 | 124.220 | 55.905 | 25.033 | 175.225 |
| 88 | A | 7.796 | 123.245 | 51.361 | 15.015 | 178.079 |
| 89 | L | 8.371 | 119.997 | 54.401 | 37.718 | 175.877 |
| 90 | G | 7.995 | 106.639 | 45.046 | 0.000 | 172.218 |
| 91 | R | 7.420 | 119.315 | 55.664 | 25.862 | 178.510 |
| 92 | A | 7.693 | 123.510 | 52.035 | 15.970 | 174.622 |
| 93 | C | 7.224 | 110.795 | 57.218 | 24.630 | 170.859 |
| 94 | G | 7.558 | 106.072 | 42.853 | 0.000 | 171.413 |
| 95 | V | 7.406 | 113.708 | 55.953 | 30.803 | 172.535 |
| 96 | S | 8.517 | 116.821 | | 59.967 | 170.652 |
| 97 | R | 6.276 | 118.594 | 47.395 | 26.939 | 169.293 |
| 98 | P | - | - | - | - | - |
| 99 | V | 6.969 | 120.357 | 57.860 | 32.950 | 172.321 |
| 100 | I | 7.797 | 113.035 | 58.109 | 34.534 | 172.293 |
| 101 | A | 6.730 | 118.179 | 47.865 | 18.417 | 171.140 |
| 102 | C | 8.899 | 113.261 | 52.951 | 28.523 | 181.381 |
| 103 | S | 8.607 | 112.318 | 51.406 | 63.475 | 171.168 |
| 104 | V | 7.372 | 129.798 | 58.503 | 28.475 | 173.352 |
| 105 | T | 8.275 | 118.194 | 57.394 | 67.667 | 170.957 |
| 106 | I | 7.930 | 118.700 | 58.470 | 36.584 | 174.118 |
| 107 | K | 8.725 | 130.607 | 52.355 | 32.203 | 172.860 |
| 108 | E | - | - | - | - | - |
| 109 | G | - | - | - | - | - |
| 110 | S | - | - | - | - | - |
| 111 | Q | - | - | - | - | - |
| 112 | L | - | - | - | - | - |
| 113 | K | - | - | - | - | - |

| Number | Residue | H | N | $C_\alpha$ | $C_\beta$ | CO |
|--------|---------|-------|---------|--------|--------|---------|
| 114 | Q | - | - | - | - | - |
| 115 | Q | 7.898 | 120.025 | 56.168 | 26.462 | 176.363 |
| 116 | I | 8.508 | 120.914 | 63.010 | 34.786 | 174.580 |
| 117 | Q | 8.460 | 118.664 | 55.780 | 25.073 | 176.320 |
| 118 | S | 7.683 | 113.540 | 58.320 | 60.092 | 175.112 |
| 119 | I | 7.735 | 123.206 | 59.560 | 32.187 | 175.090 |
| 120 | Q | 8.606 | 120.260 | 57.076 | 24.235 | 175.488 |
| 121 | Q | 7.839 | 116.680 | 55.820 | 25.027 | 175.976 |
| 122 | S | 7.769 | 115.923 | 59.654 | 60.457 | 173.613 |
| 123 | I | 8.197 | 121.290 | 62.409 | 34.532 | 175.221 |
| 124 | E | 7.926 | 118.530 | 56.638 | 26.066 | 176.686 |
| 125 | R | 7.200 | 115.075 | 54.611 | 26.722 | 174.745 |
| 126 | L | 7.565 | 117.103 | 52.692 | 38.401 | 174.970 |
| 127 | L | 7.116 | 118.334 | 51.779 | 38.539 | 173.698 |
| 128 | V | 7.165 | 122.471 | 60.592 | 29.779 | 178.480 |

## A.2   Cross-saturation experiment on the hPrp31-15.5K-U4 5'-SL-33nt complex

**Table A.2:** The cross-saturation experiment on hPrp31-15.5K-U4 5'-SL-33nt complex was carried out at 308 K at pH 7.6 on an Bruker 900 MHz spectrometer equipped with a cryo-probe. The normalized intensity changes and errors, which were calculated as described in 3.2.3 are listed here.

| Number | Residue | Normalized intensity Changes | Errors |
|--------|---------|------------------------------|--------|
| 1 | M | - | - |
| 2 | T | -0.780 | 0.017 |
| 3 | E | -0.843 | 0.010 |
| 4 | A | -0.942 | 0.008 |
| 5 | D | -0.874 | 0.008 |
| 6 | V | -0.770 | 0.015 |
| 7 | N | -0.865 | 0.053 |
| 8 | P | - | - |
| 9 | K | 0.251 | 0.080 |
| 10 | A | -0.288 | 0.141 |
| 11 | Y | 0.754 | 0.143 |
| 12 | P | - | - |
| 13 | L | -0.120 | 0.071 |
| 14 | A | -0.175 | 0.163 |
| 15 | D | - | - |
| 16 | A | - | - |
| 17 | H | 0.259 | 0.091 |
| 18 | L | -0.687 | 0.066 |
| 19 | T | 0.435 | 0.210 |
| 20 | K | -0.678 | 0.080 |
| 21 | K | -0.574 | 0.099 |
| 22 | L | -0.218 | 0.158 |
| 23 | L | -0.018 | 0.146 |
| | | continued on next page | |

| Number | Residue | Normalized intensity Changes | Errors |
|--------|---------|------------------------------|--------|
| 24 | D | -0.181 | 0.047 |
| 25 | L | -0.190 | 0.095 |
| 26 | V | -0.399 | 0.174 |
| 27 | Q | -0.570 | 0.055 |
| 28 | Q | -0.006 | 0.057 |
| 29 | S | 0.086 | 0.131 |
| 30 | C | -0.170 | 0.080 |
| 31 | N | -0.238 | 0.052 |
| 32 | Y | -0.338 | 0.090 |
| 33 | K | -0.346 | 0.065 |
| 34 | Q | -0.670 | 0.085 |
| 35 | L | -0.811 | 0.074 |
| 36 | R | 1.239 | 0.598 |
| 37 | K | -0.578 | 0.407 |
| 38 | G | 0.710 | 0.153 |
| 39 | A | 2.267 | 1.660 |
| 40 | N | -0.006 | 0.551 |
| 41 | E | - | - |
| 42 | A | 2.521 | 1.715 |
| 43 | T | -0.654 | 0.225 |
| 44 | K | 3.350 | 1.377 |
| 45 | T | -0.801 | 0.318 |
| 46 | L | -0.792 | 0.114 |
| 47 | N | 0.162 | 0.239 |
| 48 | R | 0.409 | 0.317 |
| 49 | G | -0.259 | 0.137 |
| 50 | I | 1.751 | 0.272 |
| 51 | S | -0.039 | 0.118 |
| 52 | E | 0.049 | 0.270 |
| 53 | F | 2.433 | 0.396 |
| 54 | I | -0.465 | 0.076 |
| 55 | V | -0.601 | 0.091 |
| 56 | M | -0.715 | 0.052 |
| 57 | A | -0.289 | 0.079 |
| 58 | A | -0.403 | 0.519 |
| 59 | D | 0.071 | 0.158 |
| 60 | A | -0.370 | 0.128 |
| 61 | E | -0.818 | 0.077 |
| 62 | P | - | - |
| 63 | L | 1.100 | 0.424 |
| 64 | E | -0.279 | 0.171 |
| 65 | I | 0.595 | 0.220 |
| 66 | I | 0.882 | 0.407 |
| 67 | L | -0.217 | 0.116 |
| 68 | H | -0.965 | 0.053 |
| 69 | L | -0.393 | 0.053 |
| 70 | P | - | - |
| 71 | L | 2.483 | 0.343 |
| 72 | L | 1.106 | 0.334 |
| 73 | C | -0.442 | 0.063 |
| 74 | E | -0.166 | 0.114 |
| 75 | D | -0.046 | 0.093 |
| 76 | K | -0.144 | 0.144 |
| 77 | N | 0.890 | 0.233 |
| 78 | V | -0.181 | 0.130 |

| Number | Residue | Normalized intensity Changes | Errors |
|--------|---------|------------------------------|--------|
| 79 | P | - | - |
| 80 | Y | 0.096 | 0.092 |
| 81 | V | 0.204 | 0.124 |
| 82 | F | -1.033 | 0.075 |
| 83 | V | 0.177 | 0.459 |
| 84 | R | 0.366 | 0.126 |
| 85 | S | -0.885 | 0.080 |
| 86 | K | - | - |
| 87 | Q | -0.029 | 0.077 |
| 88 | A | -0.129 | 0.092 |
| 89 | L | -0.457 | 0.074 |
| 90 | G | -0.156 | 0.254 |
| 91 | R | 0.501 | 0.173 |
| 92 | A | -0.586 | 0.090 |
| 93 | C | -1.055 | 0.057 |
| 94 | G | 0.674 | 0.256 |
| 95 | V | 0.506 | 0.257 |
| 96 | S | -0.394 | 0.215 |
| 97 | R | 0.253 | 0.120 |
| 98 | P | - | - |
| 99 | V | 0.032 | 0.322 |
| 100 | I | -0.257 | 0.193 |
| 101 | A | -0.327 | 0.144 |
| 102 | C | 2.271 | 0.540 |
| 103 | S | 0.405 | 0.629 |
| 104 | V | 0.690 | 0.410 |
| 105 | T | 0.999 | 0.206 |
| 106 | I | 0.458 | 0.109 |
| 107 | K | -0.526 | 0.082 |
| 108 | E | - | - |
| 109 | G | - | - |
| 110 | S | - | - |
| 111 | Q | - | - |
| 112 | L | - | - |
| 113 | K | - | - |
| 114 | Q | - | - |
| 115 | Q | -0.440 | 0.025 |
| 116 | I | -0.382 | 0.117 |
| 117 | Q | 0.172 | 0.136 |
| 118 | S | -0.284 | 0.095 |
| 119 | I | -0.349 | 0.048 |
| 120 | Q | -0.698 | 0.079 |
| 121 | Q | -0.466 | 0.053 |
| 122 | S | 0.150 | 0.071 |
| 123 | I | -0.578 | 0.035 |
| 124 | E | -0.403 | 0.060 |
| 125 | R | -0.100 | 0.079 |
| 126 | L | -0.621 | 0.085 |
| 127 | L | 0.192 | 0.056 |
| 128 | V | -0.682 | 0.020 |

## A.3 The backbone assignment of 15.5K-2 mutant

**Table A.3:** The backbone resonances of $^{15}$N, $^{13}C$- labelled 15.5 protein were assigned with the help from non-TROSY version 3D experiments including HNCA, HN(CO)CA and the 2D HSQC experiment. The previous information from the assignment of the wild type 15.5K was used in combination with the information obtained from these spectra. These experiments were all measured at 308 K at pH 7.6 on an Bruker 600 MHz spectrometer equipped with a cryo-probe.

| Number | Residue | H | N | $C_\alpha$ |
|--------|---------|------|---------|--------|
| 1 | M | - | - | - |
| 2 | T | - | - | - |
| 3 | E | 8.351 | 123.707 | 56.394 |
| 4 | A | 8.230 | 124.609 | 49.683 |
| 5 | D | 8.285 | 119.580 | 51.550 |
| 6 | V | 7.856 | 118.753 | 58.207 |
| 7 | N | 8.594 | 126.048 | 49.028 |
| 8 | P | - | - | - |
| 9 | K | 8.355 | 116.269 | 54.303 |
| 10 | A | 7.662 | 123.474 | 46.900 |
| 11 | Y | 6.736 | 122.888 | 51.097 |
| 12 | P | - | - | - |
| 13 | L | 8.560 | 124.632 | 50.206 |
| 14 | A | 8.729 | 129.892 | 49.765 |
| 15 | D | 8.381 | 120.509 | 49.645 |
| 16 | A | - | - | - |
| 17 | H | - | - | - |
| 18 | L | 8.914 | 122.297 | 54.504 |
| 19 | T | 8.745 | 115.673 | 65.866 |
| 20 | K | 7.232 | 120.518 | 57.130 |
| 21 | K | 7.202 | 117.143 | 56.477 |
| 22 | L | 8.937 | 120.760 | 55.314 |
| 23 | L | 8.733 | 117.856 | 55.073 |
| 24 | D | 8.076 | 120.451 | 54.740 |
| 25 | L | 7.780 | 121.656 | 54.611 |
| 26 | V | 9.064 | 123.961 | 64.664 |
| 27 | Q | 7.969 | 120.668 | 56.625 |
| 28 | Q | 8.195 | 118.954 | 56.665 |
| 29 | S | 9.063 | 117.179 | 60.378 |
| 30 | C | 8.059 | 123.279 | 59.494 |
| 31 | N | 7.690 | 119.725 | 53.063 |
| 32 | Y | 7.779 | 116.916 | 55.341 |
| 33 | K | 7.760 | 115.249 | 54.386 |
| 34 | Q | 8.443 | 117.484 | 51.291 |
| 35 | L | 7.258 | 119.797 | 50.812 |
| 36 | R | 9.118 | 125.404 | 50.287 |
| 37 | K | 9.360 | 124.202 | 53.249 |
| 38 | G | 7.563 | 108.328 | 41.040 |

| Number | Residue | H | N | $C_\alpha$ |
|--------|---------|--------|---------|--------|
| 39 | A | 10.236 | 127.824 | 53.179 |
| 40 | N | 10.305 | 123.050 | 54.567 |
| 41 | E | 8.264 | 119.812 | 58.898 |
| 42 | A | 9.423 | 122.601 | 52.670 |
| 43 | T | 8.527 | 117.802 | 65.255 |
| 44 | K | 7.703 | 120.892 | 58.010 |
| 45 | T | 8.078 | 107.671 | 62.092 |
| 46 | L | 7.289 | 123.004 | 54.438 |
| 47 | N | 8.049 | 118.041 | 53.215 |
| 48 | R | 7.890 | 115.753 | 53.370 |
| 49 | G | 7.687 | 108.842 | 44.440 |
| 50 | I | 7.915 | 109.433 | 57.450 |
| 51 | S | 8.099 | 116.876 | 55.466 |
| 52 | E | 9.872 | 126.757 | 55.378 |
| 53 | F | 7.142 | 109.955 | 54.372 |
| 54 | I | 7.911 | 118.695 | 54.251 |
| 55 | V | 7.981 | 123.394 | 57.336 |
| 56 | M | 8.347 | 123.858 | 50.280 |
| 57 | A | 8.059 | 122.052 | 47.160 |
| 58 | A | 9.412 | 121.501 | 50.601 |
| 59 | D | 8.149 | 113.374 | 49.579 |
| 60 | A | 7.007 | 122.360 | 50.990 |
| 61 | E | 8.568 | 122.527 | 49.989 |
| 62 | P | - | - | - |
| 63 | L | 8.657 | 124.601 | 54.480 |
| 64 | E | 8.804 | 113.457 | 57.735 |
| 65 | I | 7.350 | 115.159 | 62.076 |
| 66 | I | 7.814 | 107.695 | 57.703 |
| 67 | L | 7.227 | 119.149 | 54.788 |
| 68 | H | 8.355 | 117.539 | 51.021 |
| 69 | L | 7.109 | 119.598 | 57.009 |
| 70 | P | - | - | - |
| 71 | L | 6.171 | 113.204 | 55.042 |
| 72 | L | 7.626 | 122.120 | 54.782 |
| 73 | C | 8.732 | 117.615 | 62.377 |
| 74 | E | 8.242 | 119.516 | 57.018 |
| 75 | D | 7.710 | 120.405 | 56.523 |
| 76 | K | 7.879 | 114.519 | 51.503 |
| 77 | N | 7.845 | 115.938 | 59.951 |
| 78 | V | 8.336 | 121.152 | 62.640 |
| 79 | P | - | - | - |
| 80 | Y | 6.730 | 114.756 | 51.570 |
| 81 | V | 7.788 | 114.035 | 56.704 |
| 82 | F | 8.569 | 122.200 | 53.610 |
| 83 | V | 9.060 | 111.330 | 55.761 |
| 84 | R | 8.743 | 118.872 | 57.586 |
| 85 | S | 7.592 | 107.233 | 52.112 |
| 86 | K | 11.400 | 105.700 | 54.865 |
| 87 | Q | 8.566 | 124.734 | 56.229 |
| 88 | A | 7.898 | 123.611 | 51.743 |
| 89 | L | 8.482 | 120.367 | 54.692 |
| 90 | G | 8.101 | 107.043 | 45.372 |
| 91 | R | 7.524 | 119.665 | 56.038 |
| 92 | A | 7.799 | 123.864 | 52.344 |
| 93 | C | 7.338 | 111.255 | 57.370 |

| Number | Residue | H | N | $C_\alpha$ |
|--------|---------|-------|---------|--------|
| 94 | G | 7.663 | 106.385 | 43.500 |
| 95 | V | 7.498 | 113.884 | 56.110 |
| 96 | S | 8.599 | 117.079 | 55.171 |
| 97 | R | 6.364 | 118.974 | 47.547 |
| 98 | P | - | - | - |
| 99 | V | 7.069 | 120.730 | 57.995 |
| 100 | I | 7.909 | 113.377 | 58.184 |
| 101 | A | 6.853 | 118.655 | 48.102 |
| 102 | C | 8.995 | 113.568 | 53.126 |
| 103 | S | 8.710 | 112.632 | 51.562 |
| 104 | V | 7.447 | 130.229 | 58.646 |
| 105 | T | 8.374 | 118.589 | 57.497 |
| 106 | I | 8.082 | 119.507 | 58.903 |
| 107 | K | 8.838 | 131.058 | 52.625 |
| 108 | E | - | - | - |
| 109 | G | - | - | - |
| 110 | S | - | - | - |
| 111 | Q | - | - | - |
| 112 | L | - | - | - |
| 113 | K | - | - | - |
| 114 | Q | - | - | - |
| 115 | Q | 8.018 | 120.445 | 56.638 |
| 116 | I | 8.640 | 121.198 | 63.297 |
| 117 | Q | 8.599 | 119.100 | 56.062 |
| 118 | S | 7.781 | 113.924 | 58.620 |
| 119 | I | 7.840 | 123.542 | 59.860 |
| 120 | Q | 8.786 | 120.751 | 57.502 |
| 121 | Q | 7.934 | 116.886 | 56.137 |
| 122 | S | 7.831 | 115.800 | 59.935 |
| 123 | I | 8.336 | 121.152 | 62.640 |
| 124 | E | 8.049 | 118.364 | 56.503 |
| 125 | R | 7.265 | 115.411 | 55.074 |
| 126 | L | 7.656 | 117.010 | 52.745 |
| 127 | L | 7.242 | 119.047 | 52.007 |
| 128 | V | 7.262 | 123.084 | 60.932 |

# A.4  N-H dipolar coupling of 15.5K-2 mutant

**Table A.4:** The measured N-H dipolar couplings of the 15.5K-2 mutant are listed here. The 3-interleaved IPAP experiment was carried out at 308 K at pH 7.6 on a Bruker 600 MHz spectrometer equipped with a cryo-probe. The concentration of Pf1 phage in the anisotropic sample was 10mg/mL.

| Number | Residue | Atom | Number | Residue | Atom | D [Hz] |
|--------|---------|------|--------|---------|------|--------|
| 7 | ASN | HN | 7 | ASN | N | -9.677 |
| 9 | LYS | HN | 9 | LYS | N | -10.568 |
| 10 | ALA | HN | 10 | ALA | N | -9.084 |
| 11 | TYR | HN | 11 | TYR | N | 25.826 |
| 14 | ALA | HN | 14 | ALA | N | 22.085 |
| | | | | | | continued on next page |

| Number | Residue | Atom | Number | Residue | Atom | D [Hz] |
|--------|---------|------|--------|---------|------|--------|
| 15 | ASP | HN | 15 | ASP | N | -12.943 |
| 18 | LEU | HN | 18 | LEU | N | -10.983 |
| 19 | THR | HN | 19 | THR | N | -6.353 |
| 20 | LYS | HN | 20 | LYS | N | 3.740 |
| 21 | LYS | HN | 21 | LYS | N | -10.805 |
| 22 | LEU | HN | 22 | LEU | N | -6.293 |
| 23 | LEU | HN | 23 | LEU | N | -0.119 |
| 24 | ASP | HN | 24 | ASP | N | 0.831 |
| 25 | LEU | HN | 25 | LEU | N | -6.590 |
| 26 | VAL | HN | 26 | VAL | N | -3.325 |
| 27 | GLN | HN | 27 | GLN | N | 7.006 |
| 28 | GLN | HN | 28 | GLN | N | -5.462 |
| 29 | SER | HN | 29 | SER | N | -6.649 |
| 30 | CYS | HN | 30 | CYS | N | 5.106 |
| 31 | ASN | HN | 31 | ASN | N | 6.471 |
| 32 | TYR | HN | 32 | TYR | N | -3.147 |
| 33 | LYS | HN | 33 | LYS | N | -6.590 |
| 34 | GLN | HN | 34 | GLN | N | 5.818 |
| 35 | LEU | HN | 35 | LEU | N | -12.824 |
| 37 | LYS | HN | 37 | LYS | N | 0.475 |
| 39 | ALA | HN | 39 | ALA | N | -16.089 |
| 40 | ASN | HN | 40 | ASN | N | -16.327 |
| 41 | GLU | HN | 41 | GLU | N | -22.798 |
| 42 | ALA | HN | 42 | ALA | N | -9.618 |
| 43 | THR | HN | 43 | THR | N | -20.898 |
| 44 | LYS | HN | 44 | LYS | N | -23.570 |
| 46 | LEU | HN | 46 | LEU | N | 0.297 |
| 47 | ASN | HN | 47 | ASN | N | -26.716 |
| 48 | ARG | HN | 48 | ARG | N | -18.939 |
| 49 | GLY | HN | 49 | GLY | N | -7.607 |
| 50 | ILE | HN | 50 | ILE | N | 9.024 |
| 51 | SER | HN | 51 | SER | N | -7.896 |
| 53 | PHE | HN | 53 | PHE | N | -8.312 |
| 55 | VAL | HN | 55 | VAL | N | 4.338 |
| 56 | MET | HN | 56 | MET | N | 15.377 |
| 57 | ALA | HN | 57 | ALA | N | 26.598 |
| 58 | ALA | HN | 58 | ALA | N | 35.800 |
| 59 | ASP | HN | 59 | ASP | N | 5.343 |
| 60 | ALA | HN | 60 | ALA | N | -5.937 |
| 61 | GLU | HN | 61 | GLU | N | -13.833 |
| 63 | LEU | HN | 63 | LEU | N | -5.224 |
| 64 | GLU | HN | 64 | GLU | N | -8.252 |
| 65 | ILE | HN | 65 | ILE | N | -12.586 |
| 66 | ILE | HN | 66 | ILE | N | -9.855 |
| 67 | LEU | HN | 67 | LEU | N | 3.265 |
| 68 | HIS | HN | 68 | HIS | N | 25.232 |
| 69 | LEU | HN | 69 | LEU | N | -9.855 |
| 71 | LEU | HN | 71 | LEU | N | -9.855 |
| 72 | LEU | HN | 72 | LEU | N | -21.076 |
| 73 | CYS | HN | 73 | CYS | N | -7.540 |
| 74 | GLU | HN | 74 | GLU | N | -3.206 |
| 75 | ASP | HN | 75 | ASP | N | -10.805 |
| 76 | LYS | HN | 76 | LYS | N | -16.445 |
| 77 | ASN | HN | 77 | ASN | N | -4.631 |
| 80 | TYR | HN | 80 | TYR | N | -12.289 |

| Number | Residue | Atom | Number | Residue | Atom | D [Hz] |
|--------|---------|------|--------|---------|------|--------|
| 81 | VAL | HN | 81 | VAL | N | -7.480 |
| 82 | PHE | HN | 82 | PHE | N | -2.968 |
| 83 | VAL | HN | 83 | VAL | N | 28.438 |
| 84 | ARG | HN | 84 | ARG | N | 38.946 |
| 85 | SER | HN | 85 | SER | N | 12.349 |
| 87 | GLN | HN | 87 | GLN | N | -4.275 |
| 88 | ALA | HN | 88 | ALA | N | -1.306 |
| 89 | LEU | HN | 89 | LEU | N | -3.206 |
| 90 | GLY | HN | 90 | GLY | N | -10.865 |
| 91 | ARG | HN | 91 | ARG | N | -7.481 |
| 92 | ALA | HN | 92 | ALA | N | -4.928 |
| 94 | GLY | HN | 94 | GLY | N | 2.850 |
| 95 | VAL | HN | 95 | VAL | N | -10.805 |
| 96 | SER | HN | 96 | SER | N | 6.293 |
| 97 | ARG | HN | 97 | ARG | N | -3.443 |
| 99 | VAL | HN | 99 | VAL | N | 25.529 |
| 100 | ILE | HN | 100 | ILE | N | 26.182 |
| 101 | ALA | HN | 101 | ALA | N | -18.764 |
| 102 | CYS | HN | 102 | CYS | N | -4.096 |
| 103 | SER | HN | 103 | SER | N | -17.455 |
| 104 | VAL | HN | 104 | VAL | N | -14.476 |
| 105 | THR | HN | 105 | THR | N | -20.779 |
| 106 | ILE | HN | 106 | ILE | N | 21.373 |
| 107 | LYS | HN | 107 | LYS | N | 25.054 |
| 115 | GLN | HN | 115 | GLN | N | -13.477 |
| 116 | ILE | HN | 116 | ILE | N | -8.134 |
| 117 | GLN | HN | 117 | GLN | N | -1.544 |
| 118 | SER | HN | 118 | SER | N | -14.545 |
| 119 | ILE | HN | 119 | ILE | N | -10.627 |
| 120 | GLN | HN | 120 | GLN | N | -3.918 |
| 121 | GLN | HN | 121 | GLN | N | -2.119 |
| 122 | SER | HN | 122 | SER | N | -12.171 |
| 123 | ILE | HN | 123 | ILE | N | -6.471 |
| 124 | GLU | HN | 124 | GLU | N | -2.434 |
| 125 | ARG | HN | 125 | ARG | N | -9.974 |
| 126 | LEU | HN | 126 | LEU | N | -7.955 |
| 127 | LEU | HN | 127 | LEU | N | 0.238 |
| 128 | VAL | HN | 128 | VAL | N | 1.247 |

# Appendix B

# Protocols

Selected protocols used in the generation of the 15.5K-U4 5'-SL-33nt model and the docking of the hPrp31$^{188-332}$-15.5K-U4 5'-SL using HADDOCK2.0 are shown in this Appendix.

## This protocol below randomizes the penta-loop conformation in the models of 15.5K-U4 5'-SL-33nt primary RNP.

```
{!*******************************************
!reading in planarity restraints for loop bases
@/nmr1/deva/ping/taco/input/restraints/plan_loop.rst

!*******************************************}

!looping three times as in pardis protocol during SA
evaluate ($icount =0)
while ($icount<3) loop imain
evaluate ($icount = $icount + 1)

!*******************************************
!Dynamics at 1000K, 600 K and 400K

!  evaluate ($bath  = 1000)

 for $bath in (1000 600 400) loop anneal

  flags
    exclude *
    include bonds vdw noe angl plan cdih dihe impr

  end

  vector do (vx = maxwell($bath)) (all)
```

```
  vector do (vy = maxwell($bath)) (all)
  vector do (vz = maxwell($bath)) (all)



constraints  interaction = (all) (all)
weights * 1. angles $k_ang improper $k_imp vdw $k_vdw
end end

constraints fix=(not (store1)) end


    parameter
        nbonds
          cutnb=4.5 rcon=$k_vdw nbxmod=3 repel=$radius
        end
    end

    noe scale * $knoe  end

    restraints dihedral
        scale $kcdih
    end

  flags
    exclude *
    include bonds vdw noe angl plan cdih dihe impr

  end

!constraints fix=(segid U4LN and (resid 226:235 or resid 241:247)) end

    dynamics  verlet
        nstep=$nstep  timestep=0.001 iasvel=current
        tcoupling = true  tbath = $bath   nprint=5000  iprfrq=0
        ntrfr = 99999999
    end

    end loop anneal

!***********2000 step powell minimization****************
constraints fix=(not (store1)) end
 flags  exclude *
    include bonds angl vdw noe cdih dihe plan impr
  end
 mini powell nstep= 20000 nprint= 2000 end

!**********end powell minimization***********************
!rigid body minimization

flags  exclude *
    include  angl noe cdih dihe bonds impr vdw
end

!constraints fix=(segid U4LN and (resid 226:235 or resid 241:247)) end

minimize rigid
```

```
nstep = 10000
group = (store 5) group = (store 6)
end
!***********************************************
end loop imain

set message on echo on end
```

# The water refinement protocol below was used for generating the 15.5K-U4 5'-SL-33nt model for the ternary complex docking.

```
!remarks Solvent refinement protocol from ARIA1.2 (Nilges and Linge), modified for XPLOR-NIH


{*========================================================================*}
{*=== READ THE PARAMETER, TOPOLOGY, STRUCTURE AND COORDINATE FILES =========*}
{*========================================================================*}

{* type of non-bonded parameters *}
{* choice: "PROLSQ" "PARMALLH6" "PARALLHDG" "OPLSX" *}
{* The water refinement uses the OPLSX parameters *}

evaluate ( $par_nonbonded = "OPLSX" )

parameter
      @/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/rna_try_2.par
      @/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/parallhdg5.3.pro
end

parameter
  @/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/parallhdg5.3.sol
end

topology
  @/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/topallhdg5.3_iupac.pro
  @/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/rna.top
end

topology
  @/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/topallhdg5.3.sol
end

structure  @/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/rna.psf
        @/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/protein.psf
end

coordinates
@/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/initial_str/dock_rand_loop_153.pdb
set message on echo on end

!vector ident (store5) (segid U4LN and (resid 226:235 or resid 241:247)) ! X-RAY stem in STORE2
```

```
! Set occupancies to 1.00
vector do (Q = 1.00) (all)


{*==============================================================================*}
{*================== SET VALUES FOR NONBONDED PARAMETERS ===================*}
{*==============================================================================*}

parameter

! special solute-solute hydrogen bonding potential parameters
AEXP 4
REXP 6

HAEX 4

AAEX 2

! "all" possible combinations of HB-pairs in nucleic acids:
!  WELL DEPTHS DEEPENED BY 0.5 KCAL TO IMPROVE BASEPAIR ENERGIES /LN
!  AND DISTANCES INCREASED BY 0.05
!                  Emin      Rmin
!               (Kcal/mol)   (A)
hbond  N*   O*      -14.0     2.95
hbond  N*   N*      -14.5     3.05
hbond  O*   O*      -15.75    2.80
hbond  O*   N*      -15.50    2.90
hbond  S*   O*      -14.0     2.95

nbonds
nbxmod=5 atom cdiel shift
cutnb=9.5 ctofnb=8.5 ctonnb=6.5 eps=1.0 e14fac=0.4 inhibit 0.25
wmin=0.5
tolerance  0.5
end
end

! Minimize proton positions to remove problems with wrong stereochemical types
constraints fix=(not hydrogen) end
constraints
  interaction (not resname ANI) (not resname ANI)
  interaction ( resname ANI) ( resname ANI)
end

flags exclude * include  bond angle impr cdih dihe vdw elec noe end
minimize powell nstep=500 nprint=50 end
constraints fix=(not all) end

{*==============================================================================*}
{*============== COPY THE COORDINATES TO THE REFERENCE SET =================*}
{*==============================================================================*}

vector do (refx = x) (all)
vector do (refy = y) (all)
vector do (refz = z) (all)


{*==============================================================================*}
```

```
{*========================= GENERATE THE WATER LAYER =======================*}
{*========================================================================*}
{
vector do (segid = "PROT") (segid "    ")
@/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/generate_water.inp
vector do (segid = "    ") (segid "PROT")
}
{*========================================================================*}
{*========================= READ THE EXPERIMENTAL DATA ===================*}
{*========================================================================*}

!*******************************************
!reading in the artificial aform restraints
@/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/afrm.rst
!*******************************************


noe reset
  nrestraints = 30000
  ceiling     = 100

  class      cryst
  averaging  cryst sum
  potential  cryst soft
  scale      cryst 50
  sqconstant cryst 1.0
  sqexponent cryst 2
  soexponent cryst 1
  rswitch    cryst 1.0
  sqoffset   cryst 0.0
  asymptote  cryst 2.0

@/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/devan_latest.rst

end

noe
  nrestraints = 30000
  ceiling     = 1000

  class      base
  averaging  base sum
  potential  base soft
  scale      base 100
  sqconstant base 1.0
  sqexponent base 2
  soexponent base 1
  rswitch    base 1.0
  sqoffset   base 0.0
  asymptote  base 2.0

!@/nmr1/pili/155U4longmodel_all_new/input/new/restraints/base_stacking.rst

end

restraints dihedral reset nassign=10000
@/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/gen_dihedrals.tbl
```

```
  scale=200
end

@/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/plan_try_2.inp
end

sani reset
  nrestraints = 2000

  class       NHf
  force       1.0
  potential   square
!  coefficients 0.00 3.5 0.55
!  @/nmr1/deva/ping/taco/input/restraints/rdc.rst

end

{*=============================================================================*}
{*========================== SET FLAGS ========================================*}
{*=============================================================================*}

flags exclude *
include bond  angle dihe impr vdw elec
        noe cdih coup oneb carb ncs dani
        vean sani dipo prot harm plan
end


{*=============================================================================*}
{*================== SET PARAMETERS FOR MD-SIMULATION ======================*}
{*=============================================================================*}

{* set number of md steps for the heating stage *}
evaluate ( $mdsteps.heat = 100 )

{* set number of md steps for the hot md stage *}
evaluate ( $mdsteps.hot = 500)

{* set number of md steps for the cooling stage *}
evaluate ( $mdsteps.cool = 100 )

{* seed for velocity generation *}
evaluate ( $seed = 12397 )


{*=============================================================================*}
{*========================= START THE REFINEMENT ==========================*}
{*=============================================================================*}

set seed $seed end

! We loop untill we have an accepted structure, maximum trials=1
evaluate ($end_count = 1)
evaluate ($count = 0)

while ($count < $end_count ) loop main
```

```
evaluate ($accept = 0)

! since we do not use SHAKe, increase the water bond angle energy constant
parameter
angle (resn tip3) (resn tip3) (resn tip3) 500 TOKEN
end

! reduce improper and angle force constant for some atoms
evaluate ($kbonds = 1000)
evaluate ($kangle = 50)
evaluate ($kimpro = 5)
evaluate ($kchira = 5)
evaluate ($komega = 5)
parameter
angle    (not resn tip3)(not resn tip3)(not resn tip3) $kangle  TOKEN
improper (all)(all)(all)(all) $kimpro  TOKEN TOKEN
end

! fix the protein for initial minimization
constraints fix (not resn tip3) end
minimize powell nstep=500 drop=50 nprint=50 end

! release protein and restrain harmonically
constraints fix (not all) end
vector do (refx=x) (all)
vector do (refy=y) (all)
vector do (refz=z) (all)
restraints harmonic
exponent = 2
end
vector do (harmonic = 0)  (all)
vector do (harmonic = 10) (not name h*)
vector do (harmonic = 20.0)(resname ANI and name OO)
vector do (harmonic = 0.0) (resname ANI and name Z )
vector do (harmonic = 0.0) (resname ANI and name X )
vector do (harmonic = 0.0) (resname ANI and name Y )

constraints
interaction (not resname ANI) (not resname ANI)
interaction ( resname ANI) ( resname ANI)
end

minimize powell nstep=5000 drop=500 nprint=500 end
vector do (refx=x) (not resname ANI)
vector do (refy=y) (not resname ANI)
vector do (refz=z) (not resname ANI)
minimize powell nstep=500 drop=50 nprint=50 end
vector do (refx=x) (not resname ANI)
vector do (refy=y) (not resname ANI)
vector do (refz=z) (not resname ANI)

vector do (mass =50) (all)
vector do (mass=1000) (resname ani)
vector do (fbeta = 0) (all)
vector do (fbeta = 20. {1/ps} ) (not resn ani)
evaluate ($kharm = 50)
! heat to 2000 K
```

```
for $bath in (100 200 400 800 1000 1500 2000) loop heat
vector do (harm = $kharm) (not name h* and not resname ANI)
vector do (vx=maxwell($bath)) (all)
vector do (vy=maxwell($bath)) (all)
vector do (vz=maxwell($bath)) (all)

!constraints fix=(segid U4LN and (resid 226:234 or resid 242:247)) end


dynamics verlet
        nstep=$mdsteps.heat timest=0.002{ps}
        tbath=$bath  tcoupling = true
        iasvel=current
        nprint=50
end
evaluate ($kharm = max(0, $kharm - 4))
vector do (refx=x) (not resname ANI)
vector do (refy=y) (not resname ANI)
vector do (refz=z) (not resname ANI)
end loop heat

! refinement at high T:
constraints
interaction (not resname ANI) (not resname ANI) weights * 1 dihed 2 end
interaction ( resname ANI) ( resname ANI) weights * 1 end
end

vector do (harm = 0)  (not resname ANI)
vector do (vx=maxwell($bath)) (all)
vector do (vy=maxwell($bath)) (all)
vector do (vz=maxwell($bath)) (all)

!constraints fix=(segid U4LN and (resid 226:234 or resid 242:247)) end

dynamics verlet
nstep=$mdsteps.hot timest=0.002{ps}
tbath=$bath  tcoupling = true
iasvel=current
nprint=50
end

constraints
interaction (not resname ANI) (not resname ANI) weights * 1 dihed 3 end
interaction ( resname ANI) ( resname ANI) weights * 1  end
end


! cool
evaluate ($bath = 2000)
while ($bath >= 25) loop cool

evaluate ($kbonds   = max(225,$kbonds / 1.1))
evaluate ($kangle   = min(200,$kangle * 1.1))
evaluate ($kimpro   = min(200,$kimpro * 1.4))
evaluate ($kchira   = min(800,$kchira * 1.4))
evaluate ($komega   = min(80,$komega * 1.4))
```

```
parameter
        bond     (not resn tip3 and not name H*)(not resn tip3 and not name H*)
        $kbonds  TOKEN
        angle    (not resn tip3 and not name H*)(not resn tip3 and not name H*)
        (not resn tip3 and not name H*) $kangle  TOKEN
        improper (all)(all)(all)(all) $kimpro  TOKEN TOKEN
        !VAL: stereo CB
        improper (name HB and resn VAL)(name CA and resn VAL)
        (name CG1 and resn VAL)(name CG2 and resn VAL) $kchira TOKEN TOKEN
        !THR: stereo CB
        improper (name HB and resn THR)(name CA and resn THR)
        (name OG1 and resn THR)(name CG2 and resn THR) $kchira TOKEN TOKEN
        !LEU: stereo CG
        improper (name HG and resn LEU)(name CB and resn LEU)
        (name CD1 and resn LEU)(name CD2 and resn LEU) $kchira TOKEN TOKEN
        !ILE: chirality CB
        improper (name HB and resn ILE)(name CA and resn ILE)
        (name CG2 and resn ILE)(name CG1 and resn ILE) $kchira TOKEN TOKEN
        !chirality CA
        improper (name HA)(name N)(name C)(name CB) $kchira TOKEN TOKEN

        improper (name O)  (name C) (name N) (name CA) $komega TOKEN TOKEN
        improper (name HN) (name N) (name C) (name CA) $komega TOKEN TOKEN
        improper (name CA) (name C) (name N) (name CA) $komega TOKEN TOKEN
        improper (name CD) (name N) (name C) (name CA) $komega TOKEN TOKEN
end

vector do (vx=maxwell($bath)) (all)
vector do (vy=maxwell($bath)) (all)
vector do (vz=maxwell($bath)) (all)

!constraints fix=(segid U4LN and (resid 226:234 or resid 242:247)) end

dynamics verlet
        nstep=$mdsteps.cool timest=0.002{ps}
        tbath=$bath  tcoupling = true
        iasvel=current
        nprint=50
end

evaluate ($bath = $bath - 25)
end loop cool


!final minimization:
mini powell nstep 200 end

{*==========================================================================*}
{*====================== CHECK RESTRAINT VIOLATIONS ======================*}
{*==========================================================================*}

constraints interaction
(not resname TIP* and not resname ANI)
(not resname TIP* and not resname ANI)
end

energy end
```

```
! NOES overall analysis
print threshold=0.5 noe
evaluate ( $viol.noe.viol05 = $violations )
print threshold=0.3 noe
evaluate ( $viol.noe.viol03 = $violations )
print threshold=0.1 noe
evaluate ( $viol.noe.viol01 = $violations )
evaluate ( $rms.noe = $result )

! NOES
evaluate ( $viol.noe.total = 0 )

evaluate ( $accept.noe.1 = 0.5 )

noe reset
  nrestraints = 30000
  ceiling     = 100
    ^M
  class      cryst
  averaging  cryst sum
  potential  cryst soft
  scale      cryst 50
  sqconstant cryst 1.0
  sqexponent cryst 2
  soexponent cryst 1
  rswitch    cryst 1.0
  sqoffset   cryst 0.0
  asymptote  cryst 2.0

@/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/devan_latest.rst

end

noe
  nrestraints = 30000
  ceiling     = 1000

  class      2
  averaging  2 sum
  potential  2 soft
  scale      2 100
  sqconstant 2 1.0
  sqexponent 2 2
  soexponent 2 1
  rswitch    2 1.0
  sqoffset   2 0.0
  asymptote  2 2.0

!@/nmr1/pili/155U4longmodel_all_new/input/new/restraints/base_stacking.rst

end

print threshold=$accept.noe.1
end
evaluate ( $viol.noe.1 = $violations )
evaluate ( $viol.noe.total = $violations + $viol.noe.total )
```

```
! DIHEDRALS
evaluate ( $viol.cdih.total = 0 )

evaluate ( $accept.cdih.1 = 5.0 )
restraints dihedral reset
@/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/gen_dihedrals.tbl
  scale=200
end
@/home/mpg04/MBPC/030/pli/Haddock/haddock1.3/HD_splice_1/extend_aform/plan_try_2.inp
end
print threshold=$accept.cdih.1 cdih
evaluate ( $rms.cdih = $result )
evaluate ( $viol.cdih.1 = $violations )
evaluate ( $viol.cdih.total = $viol.cdih.total + $violations )

! SANI

! SANI class NHf
sani
  print thres = 0.0 class NHf
end

sani
  print thres = 0.0 all
end
evaluate ( $rms.sani = $result )

{*============================================================================*}
{*===================== CHECK ACCEPTANCE CRITERIA =========================*}
{*============================================================================*}

if ( $viol.cdih.total  > 0 ) then  evaluate ( $accept=$accept + 1 ) end if
if ( $viol.noe.total  > 0 ) then  evaluate ( $accept=$accept + 1 ) end if

if ($accept = 0 ) then
evaluate ( $label = "ACCEPTED" )
exit main
else
evaluate ( $label = "NOT ACCEPTED" )
evaluate ( $count = $count + 1 )
end if

end loop main

!****************************************************************************

!lines below this print out threshold and violations

   print threshold=0.5 noe
   evaluate ($rms_noe=$result)
   evaluate ($violations_noe=$violations)

   print threshold=5.0 cdih
   evaluate ($rms_cdih=$result)
   evaluate ($violations_cdih=$violations)
```

```
    print threshold=0.05 bonds
    evaluate ($rms_bonds=$result)
    evaluate ($violations_bonds=$violations)

    print threshold=5.0 angles
    evaluate ($rms_angles=$result)
    evaluate ($violations_angles=$violations)

    print threshold=5.0 impropers
    evaluate ($rms_impropers=$result)
    evaluate ($violations_imp=$violations)

    set message on echo on end


remarks Structure $label
remarks Energies and RMS deviations:
remarks ener.total= $ener
remarks ener.cdih= $cdih ener.noe= $noe ener.sani= $sani
remarks rms.cdih= $rms.cdih rms.noe= $rms.noe rms.sani= $rms.sani
remarks SANI/DIPO violations are not listed, Q-factors are in file cornilQ_xxx.txt
remarks cdih violations ( classes: ['1'] )
remarks viol.cdih.1=$viol.cdih.1 ( accept.cdih.1=$accept.cdih.1 )
remarks noe violations ( classes: ['1'] )
remarks viol.noe.1=$viol.noe.1 ( accept.noe.1=$accept.noe.1 )
remarks All NOE violations >0.5, 0.3 and 0.1A respectively ( all classes ):
remarks viol.noe.viol05=$viol.noe.viol05 viol.noe.viol03=$viol.noe.viol03 viol.noe.viol01=$viol.noe.viol01
remarks ener.cdih= $cdih ener.noe= $noe
evaluate ($filename="h2o/str/str_gen_model/water_ref_afxt_153.pdb")
write coordinates sele= (not resn TIP3 and not resn ANI) output=$filename end

stop
```

# Below is the dna-rna_restraints file modified from the default file provided by HADDOCK2.0. The file was used to keep the structural integrity of the U4 5'-SL-33nt model during the docking procedure.

```
{+ file: dna-rna_restraints.def  +}
{+ directory: auxiliary +}
{+ description: Creates restraints to maintain conformation of DNA/RNA +}
{+ comment:
          This file is to be read by refinement files that modify
          atom coordinates +}
{+ authors: Axel T. Brunger, and Paul D. Adams, modified by Alexandre Bonvin for NMR use +}
{+ copyright: Yale University +}

{- begin block parameter definition -} define(
```

```
{======================= base planarity =========================}

{* Nucleic acid residues to have base planarity restrained. This selection
   must only include nucleotide residues *}
{===>} bases_planar=((resid 220:252) and segid B);

{======================= sugar puckers =========================}

{* residues with sugar pucker restrained - group 1 *}
{===>} pucker_1=(resid 220:223 and segid B);

{* conformation of group 1 *}
{+ choice: "a-form" "b-form" "other" +}
{===>} form_1="a-form";

{* user defined sugar pucker for group 1 *}

{* dihedral C1'-C2'-C3'-C4' *}
{===>} dihedral_nu2_1=37.053;
{* dihedral C5'-C4'-C3'-C2' *}
{===>} dihedral_nu3_1=-155.59;
{* dihedral C1'-O4'-C4'-C5' *}
{===>} dihedral_nu4_1=144.26;

{* user defined backbone dihedrals for group 1 *}

{* alpha dihedral O3'-P-O5'-C5' *}
{===>} dihedral_alpha_1=-62.0;
{* alpha dihedral range *}
{===>} error_alpha_1=5.0;
{* beta dihedral P-O5'-C5'-C4' *}
{===>} dihedral_beta_1=-180;
{* beta dihedral range *}
{===>} error_beta_1=5.0;
{* gamma dihedral O5'-C5'-C4'-C3' *}
{===>} dihedral_gamma_1=47.0;
{* gamma dihedral range *}
{===>} error_gamma_1=5.0;
{* epsilon dihedral C4'-C3'-O3'-P *}
{===>} dihedral_eps_1=-152.0;
{* epsilon dihedral range *}
{===>} error_eps_1=5.0;
{* zeta dihedral C3'-O3'-P-O5' *}
{===>} dihedral_zeta_1=-74.0;
{* zeta dihedral range *}
{===>} error_zeta_1=5.0;

{* residues with sugar pucker restrained - group 2 *}
{===>} pucker_2=(resid 226:228 and segid B);

{* conformation of group 2 *}
{+ choice: "a-form" "b-form" "other" +}
{===>} form_2="a-form";

{* user defined sugar pucker for group 2 *}

{* dihedral C1'-C2'-C3'-C4' *}
```

```
{===>} dihedral_nu2_1=37.053;
{* dihedral C5'-C4'-C3'-C2' *}
{===>} dihedral_nu3_1=-155.59;
{* dihedral C1'-O4'-C4'-C5' *}
{===>} dihedral_nu4_1=144.26;

{* user defined backbone dihedrals for group 2 *}

{* alpha dihedral O3'-P-O5'-C5' *}
{===>} dihedral_alpha_1=-62.0;
{* alpha dihedral range *}
{===>} error_alpha_1=5.0;
{* beta dihedral P-O5'-C5'-C4' *}
{===>} dihedral_beta_1=-180;
{* beta dihedral range *}
{===>} error_beta_1=5.0;
{* gamma dihedral O5'-C5'-C4'-C3' *}
{===>} dihedral_gamma_1=47.0;
{* gamma dihedral range *}
{===>} error_gamma_1=5.0;
{* epsilon dihedral C4'-C3'-O3'-P *}
{===>} dihedral_eps_1=-152.0;
{* epsilon dihedral range *}
{===>} error_eps_1=5.0;
{* zeta dihedral C3'-O3'-P-O5' *}
{===>} dihedral_zeta_1=-74.0;
{* zeta dihedral range *}
{===>} error_zeta_1=5.0;

{* residues with sugar pucker restrained - group 3 *}
{===>} pucker_3=(resid 234:235 and segid B);

{* conformation of group 3 *}
{+ choice: "a-form" "b-form" "other" +}
{===>} form_3="a-form";

{* user defined sugar pucker for group 3 *}

{* dihedral C1'-C2'-C3'-C4' *}
{===>} dihedral_nu2_1=37.053;
{* dihedral C5'-C4'-C3'-C2' *}
{===>} dihedral_nu3_1=-155.59;
{* dihedral C1'-O4'-C4'-C5' *}
{===>} dihedral_nu4_1=144.26;

{* user defined backbone dihedrals for group 3 *}

{* alpha dihedral O3'-P-O5'-C5' *}
{===>} dihedral_alpha_1=-62.0;
{* alpha dihedral range *}
{===>} error_alpha_1=5.0;
{* beta dihedral P-O5'-C5'-C4' *}
{===>} dihedral_beta_1=-180;
{* beta dihedral range *}
{===>} error_beta_1=5.0;
{* gamma dihedral O5'-C5'-C4'-C3' *}
{===>} dihedral_gamma_1=47.0;
```

```
{* gamma dihedral range *}
{===>} error_gamma_1=5.0;
{* epsilon dihedral C4'-C3'-O3'-P *}
{===>} dihedral_eps_1=-152.0;
{* epsilon dihedral range *}
{===>} error_eps_1=5.0;
{* zeta dihedral C3'-O3'-P-O5' *}
{===>} dihedral_zeta_1=-74.0;
{* zeta dihedral range *}
{===>} error_zeta_1=5.0;

{* residues with sugar pucker restrained - group 4 *}
{===>} pucker_4=(resid 241:242 and segid B);

{* conformation of group 4 *}
{+ choice: "a-form" "b-form" "other" +}
{===>} form_4="a-form";

{* user defined sugar pucker for group 4 *}

{* dihedral C1'-C2'-C3'-C4' *}
{===>} dihedral_nu2_1=37.053;
{* dihedral C5'-C4'-C3'-C2' *}
{===>} dihedral_nu3_1=-155.59;
{* dihedral C1'-O4'-C4'-C5' *}
{===>} dihedral_nu4_1=144.26;

{* user defined backbone dihedrals for group 4 *}

{* alpha dihedral O3'-P-O5'-C5' *}
{===>} dihedral_alpha_1=-62.0;
{* alpha dihedral range *}
{===>} error_alpha_1=5.0;
{* beta dihedral P-O5'-C5'-C4' *}
{===>} dihedral_beta_1=-180;
{* beta dihedral range *}
{===>} error_beta_1=5.0;
{* gamma dihedral O5'-C5'-C4'-C3' *}
{===>} dihedral_gamma_1=47.0;
{* gamma dihedral range *}
{===>} error_gamma_1=5.0;
{* epsilon dihedral C4'-C3'-O3'-P *}
{===>} dihedral_eps_1=-152.0;
{* epsilon dihedral range *}
{===>} error_eps_1=5.0;
{* zeta dihedral C3'-O3'-P-O5' *}
{===>} dihedral_zeta_1=-74.0;
{* zeta dihedral range *}
{===>} error_zeta_1=5.0;

{* residues with sugar pucker restrained - group 5 *}
{===>} pucker_4=(resid 245:247 and segid B);

{* conformation of group 4 *}
{+ choice: "a-form" "b-form" "other" +}
{===>} form_4="a-form";
```

```
{* user defined sugar pucker for group 5 *}

{* dihedral C1'-C2'-C3'-C4' *}
{===>} dihedral_nu2_1=37.053;
{* dihedral C5'-C4'-C3'-C2' *}
{===>} dihedral_nu3_1=-155.59;
{* dihedral C1'-O4'-C4'-C5' *}
{===>} dihedral_nu4_1=144.26;

{* user defined backbone dihedrals for group 5 *}

{* alpha dihedral O3'-P-O5'-C5' *}
{===>} dihedral_alpha_1=-62.0;
{* alpha dihedral range *}
{===>} error_alpha_1=5.0;
{* beta dihedral P-O5'-C5'-C4' *}
{===>} dihedral_beta_1=-180;
{* beta dihedral range *}
{===>} error_beta_1=5.0;
{* gamma dihedral O5'-C5'-C4'-C3' *}
{===>} dihedral_gamma_1=47.0;
{* gamma dihedral range *}
{===>} error_gamma_1=5.0;
{* epsilon dihedral C4'-C3'-O3'-P *}
{===>} dihedral_eps_1=-152.0;
{* epsilon dihedral range *}
{===>} error_eps_1=5.0;
{* zeta dihedral C3'-O3'-P-O5' *}
{===>} dihedral_zeta_1=-74.0;
{* zeta dihedral range *}
{===>} error_zeta_1=5.0;

{* residues with sugar pucker restrained - group 6 *}
{===>} pucker_4=(resid 249:252 and segid B);

{* conformation of group 4 *}
{+ choice: "a-form" "b-form" "other" +}
{===>} form_4="a-form";

{* user defined sugar pucker for group 6 *}

{* dihedral C1'-C2'-C3'-C4' *}
{===>} dihedral_nu2_1=37.053;
{* dihedral C5'-C4'-C3'-C2' *}
{===>} dihedral_nu3_1=-155.59;
{* dihedral C1'-O4'-C4'-C5' *}
{===>} dihedral_nu4_1=144.26;

{* user defined backbone dihedrals for group 6 *}

        {* alpha dihedral O3'-P-O5'-C5' *}
{===>} dihedral_alpha_1=-62.0;
{* alpha dihedral range *}
{===>} error_alpha_1=5.0;
{* beta dihedral P-O5'-C5'-C4' *}
{===>} dihedral_beta_1=-180;
{* beta dihedral range *}
```

```
{===>} error_beta_1=5.0;
{* gamma dihedral O5'-C5'-C4'-C3' *}
{===>} dihedral_gamma_1=47.0;
{* gamma dihedral range *}
{===>} error_gamma_1=5.0;
{* epsilon dihedral C4'-C3'-O3'-P *}
{===>} dihedral_eps_1=-152.0;
{* epsilon dihedral range *}
{===>} error_eps_1=5.0;
{* zeta dihedral C3'-O3'-P-O5' *}
{===>} dihedral_zeta_1=-74.0;
{* zeta dihedral range *}
{===>} error_zeta_1=5.0;


{====================== Watson-Crick base pairs ======================}

{* residues which form Watson-Crick pairs *}

{* selection for pair 1 base A *}
{===>} base_a_1=(resid 220 and segid B);
{* selection for pair 1 base B *}
{===>} base_b_1=(resid 252 and segid B);

{* selection for pair 2 base A *}
{===>} base_a_2=(resid 221 and segid B);
{* selection for pair 2 base B *}
{===>} base_b_2=(resid 251 and segid B);

{* selection for pair 3 base A *}
{===>} base_a_3=(resid 222 and segid B);
{* selection for pair 3 base B *}
{===>} base_b_3=(resid 250 and segid B);

{* selection for pair 4 base A *}
{===>} base_a_4=(resid 223 and segid B);
{* selection for pair 4 base B *}
{===>} base_b_4=(resid 249 and segid B);

{* selection for pair 5 base A *}
{===>} base_a_5=(resid 226 and segid B);
{* selection for pair 5 base B *}
{===>} base_b_5=(resid 247 and segid B);

{* selection for pair 6 base A *}
{===>} base_a_6=(resid 227 and segid B);
{* selection for pair 6 base B *}
{===>} base_b_6=(resid 246 and segid B);

{* selection for pair 7 base A *}
{===>} base_a_7=(resid 228 and segid B);
{* selection for pair 7 base B *}
{===>} base_b_7=(resid 245 and segid B);

{* selection for pair 8 base A *}
{===>} base_a_8=(resid 234 and segid B);
{* selection for pair 8 base B *}
```

```
{===>} base_b_8=(resid 242 and segid B);

{* selection for pair 9 base A *}
{===>} base_a_9=(resid 235 and segid B);
{* selection for pair 9 base B *}
{===>} base_b_9=(resid 241 and segid B);


{==========================================================================}
{          things below this line do not normally need to be changed        }
{==========================================================================}

 ) {- end block parameter definition -}

 set message on echo off end

{- the planarity restraints single bases -}

 for $id in id ( &bases_planar and tag ) loop plan

   show (segid) (id $id)
   evaluate ($segid=$result)
   show (resid) (id $id)
   evaluate ($resid=decode($result))
   evaluate ($plweight = 500)

   restraints plane

     group
       selection=( segid $segid and resid $resid and
                  (resname THY or resname CYT or resname GUA or
                   resname ADE or resname URI) and
                   not (name c#' or name h#' or name h#'' or name o#p or
                        name h7# or name o#' or name p or name h#t or name o#t))

       weight=$plweight
     end

   end

 end loop plan

{- the planarity restraints for Watson-Crick base pairing -}

 evaluate ($pair=1)
 evaluate ($done=false)

 while ( $done = false ) loop plan_paired

   if ( &exist_base_a_$pair = true ) then

     if ( &exist_base_b_$pair = true ) then

       show (segid) ( &base_a_$pair and name C1' )
       evaluate ($Asegid=$result)
       show (resid) ( &base_a_$pair and name C1' )
       evaluate ($Aresid=$result)
```

```
            show (segid) ( &base_b_$pair and name C1' )
            evaluate ($Bsegid=$result)
            show (resid) ( &base_b_$pair and name C1' )
            evaluate ($Bresid=$result)
            evaluate ($plweight = 300)

            restraints plane

              group
                selection=(((segid $Asegid and resid $Aresid) or (segid $Bsegid and resid $Bresid)) and
                            (resname THY or resname CYT or resname GUA or
                             resname ADE or resname URI) and
                            not (name c#' or name h#' or name h#'' or name o#p or
                                 name h7# or name o#' or name p or name h#t or name o#t))

                weight=$plweight
              end

            end

          end if

      else

        evalute ($done = true)

      end if

      evaluate ($pair = $pair + 1)

  end loop plan_paired

  flag include plan end

  {- the dihedral restraints for sugar puckers -}

  evaluate ($group=1)
  evaluate ($done=false)

  while ( $done = false ) loop dihe

    if ( &exist_pucker_$group = true ) then

      show sum(1) ( &pucker_$group )

      if ( $result > 0 ) then

        if ( &form_$group = "a-form" ) then
          evaluate ($dihedral_1=37.053)
          evaluate ($dihedral_2=-155.59)
          evaluate ($dihedral_3=144.26)
        elseif ( &form_$group = "b-form" ) then
          evaluate ($dihedral_1=-34.9)
          evaluate ($dihedral_2=-86.4)
          evaluate ($dihedral_3=106.4)
        elseif ( &form_$group = "other" ) then
          evaluate ($dihedral_1=&dihedral_nu2_$group)
```

```
      evaluate ($dihedral_2=&dihedral_nu3_$group)
      evaluate ($dihedral_3=&dihedral_nu4_$group)
   end if

   evaluate ($min_resid_$group = 99999)
   evaluate ($max_resid_$group = -99999)

   for $id in id ( &pucker_$group and tag ) loop resid

     show (segid) (id $id)
     evaluate ($segid=$result)
     show (resid) ( id $id )
     evaluate ($resid=decode($result))
evaluate ($min_resid_$group = max($min_resid_$group,$resid))
evaluate ($max_resid_$group = max($max_resid_$group,$resid))

     restraints dihedral

       assign  ( segid $segid and resid $resid and name c1' )
               ( segid $segid and resid $resid and name c2' )
               ( segid $segid and resid $resid and name c3' )
               ( segid $segid and resid $resid and name c4' )
                                        20.0 $dihedral_1 0.0 2

       assign  ( segid $segid and resid $resid and name c5' )
               ( segid $segid and resid $resid and name c4' )
               ( segid $segid and resid $resid and name c3' )
               ( segid $segid and resid $resid and name c2' )
                                        20.0 $dihedral_2 0.0 2

       assign  ( segid $segid and resid $resid and name c1' )
               ( segid $segid and resid $resid and name o4' )
               ( segid $segid and resid $resid and name c4' )
               ( segid $segid and resid $resid and name c5' )
                                        20.0 $dihedral_3 0.0 2

       scale=20.0

     end

   end loop resid

  end if

else

  evaluate ($done=true)

end if

evaluate ($group=$group+1)

end loop dihe

{- the dihedral restraints for the backone -}

evaluate ($group=1)
```

```
evaluate ($done=false)

while ( $done = false ) loop bdihe

  if ( &exist_pucker_$group = true ) then

    show sum(1) ( &pucker_$group )

    if ( $result > 0 ) then

        evaluate ($resid=$min_resid_$group)
        evaluate ($nres=$max_resid_$group - $min_resid_$group + 1)

        if ( &form_$group = "a-form" ) then
          evaluate ($dihedral_alpha=-62)
evaluate ($error_alpha=5)
          evaluate ($dihedral_beta=-180)
evaluate ($error_beta=5)
          evaluate ($dihedral_gamma=47)
evaluate ($error_gamma=5)
          evaluate ($dihedral_zeta=-74)
evaluate ($error_zeta=5)
          evaluate ($dihedral_epsilon=-152)
evaluate ($error_epsilon=5)
        elseif ( &form_$group = "b-form" ) then
          evaluate ($dihedral_alpha=-63.6)
evaluate ($error_alpha=6)
          evaluate ($dihedral_beta=176)
evaluate ($error_beta=7)
          evaluate ($dihedral_gamma=51.4)
evaluate ($error_gamma=7)
          evaluate ($dihedral_epsilon=-171.7)
evaluate ($error_epsilon=3.7)
          evaluate ($dihedral_zeta=-103.8)
evaluate ($error_zeta=10)
        elseif ( &form_$group = "other" ) then
          evaluate ($dihedral_alpha=&dihedral_alpha_$group)
evaluate ($error_alpha=&error_alpha_$group)
          evaluate ($dihedral_beta=&dihedral_beta_$group)
evaluate ($error_beta=&error_beta_$group)
          evaluate ($dihedral_gamma=&dihedral_gamma_$group)
evaluate ($error_gamma=&error_gamma_$group)
          evaluate ($dihedral_zeta=&dihedral_zeta_$group)
evaluate ($error_zeta=&error_zeta_$group)
          evaluate ($dihedral_epsilon=&dihedral_eps_$group)
evaluate ($error_epsilon=&error_eps_$group)
        end if

        for $id in id ( &pucker_$group and tag ) loop resid

          show (segid) (id $id)
          evaluate ($segid=$result)
          show (resid) ( id $id )
          evaluate ($resid=decode($result))

          if ($resid > $min_resid_$group) then
            evaluate ($rprec = $resid - 1)
```

```
        restraint dihedral
         ! alpha
           assign  ( segid $segid and resid $rprec and name O3' )
                   ( segid $segid and resid $resid and name P )
                   ( segid $segid and resid $resid and name O5' )
                   ( segid $segid and resid $resid and name C5' )
                                           1.0 $dihedral_alpha $error_alpha 2
          ! beta
           assign  ( segid $segid and resid $resid and name P )
                   ( segid $segid and resid $resid and name O5' )
                   ( segid $segid and resid $resid and name C5' )
                   ( segid $segid and resid $resid and name C4' )
                                           1.0 $dihedral_beta $error_beta 2
          scale 200.0
         end
        end if

        restraints dihedral
          ! gamma
          assign  ( segid $segid and resid $resid and name O5' )
                  ( segid $segid and resid $resid and name C5' )
                  ( segid $segid and resid $resid and name C4' )
                  ( segid $segid and resid $resid and name C3' )
                                          1.0 $dihedral_gamma $error_gamma 2
          scale=200.0
        end

        if ($resid < $max_resid_$group) then
          evaluate ($rfoll = $resid + 1)
          restraint dihedral
            ! epsilon
            assign  ( segid $segid and resid $resid and name C4' )
                    ( segid $segid and resid $resid and name C3' )
                    ( segid $segid and resid $resid and name O3' )
                    ( segid $segid and resid $rfoll and name P )
                                            1.0 $dihedral_epsilon $error_epsilon 2
            ! zeta
            assign  ( segid $segid and resid $resid and name C3' )
                    ( segid $segid and resid $resid and name O3' )
                    ( segid $segid and resid $rfoll and name P )
                    ( segid $segid and resid $rfoll and name O5' )
                                            1.0 $dihedral_zeta $error_zeta 2
          scale 200.0
         end
        end if

      end loop resid

    end if

  else

    evaluate ($done=true)

  end if

  evaluate ($group=$group+1)
```

```
end loop bdihe

flags include cdih end

{- Watson-Crick base pairing -}

noe
  class hres
  averaging hres cent
  potential hres square
  sqconstant hres 1.
  sqexponent hres 2
  scale hres 50.
end

evaluate ($pair=1)
evaluate ($done=false)

while ( $done = false ) loop noe

  if ( &exist_base_a_$pair = true ) then

    if ( &exist_base_b_$pair = true ) then

      show ( resname ) ( &base_a_$pair and name C1' )
      evaluate ($ares=$result)
      show ( resname ) ( &base_b_$pair and name C1' )
      evaluate ($bres=$result)

      noe

        if ( $ares = THY ) then

          assign (&base_a_$pair and name o4)
                 (&base_b_$pair and name n6) 2.89 0.2 0.2
          assign (&base_a_$pair and name n3)
                 (&base_b_$pair and name n1) 2.92 0.2 0.2
          assign (&base_a_$pair and name h3)
                 (&base_b_$pair and name n1) 1.87 0.2 0.2
          assign (&base_a_$pair and name o2)
                 (&base_b_$pair and name h2) 2.94 0.2 0.2
          assign (&base_a_$pair and name o4)
                 (&base_b_$pair and name n1) 3.69 0.2 0.2
          assign (&base_a_$pair and name o2)
                 (&base_b_$pair and name n1) 3.67 0.2 0.2

        elseif ( $ares = URI ) then

          assign (&base_a_$pair and name o4)
                 (&base_b_$pair and name n6) 2.95 0.01 0.01
          assign (&base_a_$pair and name n3)
                 (&base_b_$pair and name n1) 2.82 0.01 0.01
          assign (&base_a_$pair and name o4)
                 (&base_b_$pair and name n1) 3.63 0.01 0.01
          assign (&base_a_$pair and name o2)
                 (&base_b_$pair and name n6) 5.40 0.01 0.01
```

```
      elseif ( $ares = ADE ) then

        assign (&base_b_$pair and name o4)
               (&base_a_$pair and name n6) 2.89 0.2 0.2
        assign (&base_b_$pair and name n3)
               (&base_a_$pair and name n1) 2.92 0.2 0.2
        assign (&base_b_$pair and name h3)
               (&base_a_$pair and name n1) 1.87 0.2 0.2
        assign (&base_b_$pair and name o2)
               (&base_a_$pair and name h2) 2.94 0.2 0.2
        assign (&base_b_$pair and name o4)
               (&base_a_$pair and name n1) 3.69 0.2 0.2
        assign (&base_b_$pair and name o2)
               (&base_a_$pair and name n1) 3.67 0.2 0.2

      elseif ( $ares = CYT ) then

        assign (&base_a_$pair and name n3)
               (&base_b_$pair and name n1) 2.87 0.2 0.2
        assign (&base_a_$pair and name n3)
               (&base_b_$pair and name h1) 1.86 0.2 0.2
        assign (&base_a_$pair and name n4)
               (&base_b_$pair and name o6) 2.81 0.2 0.2
        assign (&base_a_$pair and name o2)
               (&base_b_$pair and name n2) 2.81 0.2 0.2
        assign (&base_a_$pair and name n3)
               (&base_b_$pair and name o6) 3.58 0.2 0.2
        assign (&base_a_$pair and name n3)
               (&base_b_$pair and name n2) 3.63 0.2 0.2

      elseif ( $ares = GUA ) then

        assign (&base_b_$pair and name n3)
               (&base_a_$pair and name n1) 2.87 0.2 0.2
        assign (&base_b_$pair and name n3)
               (&base_a_$pair and name h1) 1.86 0.2 0.2
        assign (&base_b_$pair and name n4)
               (&base_a_$pair and name o6) 2.81 0.2 0.2
        assign (&base_b_$pair and name o2)
               (&base_a_$pair and name n2) 2.81 0.2 0.2
        assign (&base_b_$pair and name n3)
               (&base_a_$pair and name o6) 3.58 0.2 0.2
        assign (&base_b_$pair and name n3)
               (&base_a_$pair and name n2) 3.63 0.2 0.2

      end if

    end

  end if

else

  evaluate ($done=true)

end if
```

```
   evaluate ($pair=$pair+1)

end loop noe

flags include noe end
```

# Lebenslauf

**Name:** Ping Li

**Geschlecht:** weiblich

**Geburtsdatum:** 26.06.1981

**Geburtsort:** Beijing, China

**Staatsangehörigkeit:** Chinesisch

**Familienstand:** ledig

## Schulbildung

**1996-1998**    Beijing No.4 High School, Peking, China
**1998-1999**    Burnside High School, Christchurch, New Zealand
**12/1999**    Abgeschlossene Schulbildung und Zulassung zum Universitätsstudium durch 'Bursary' Prüfung, Note 'A'.

## Hochschulstudium

**2000**    Direkte Zulassung zum zweiten Studienjahr der Universität im Hauptfach Chemie an der Universität University of Canterbury, Christchurch, Neuseeland.
**2001**    Weiterführendes Studium der Chemie an der Universität Uppsala, Schweden, durch ein Austauschprogramm der Universität University of Canterbury, Neuseeland.
**2002**    Abgeschlossenes Studium zum 'Bachelor of Science with Honors' der Universität Canterbury, Neuseeland, Note 'First Class'.
**02/2003**
**-2007**    Doktorarbeit am Max-Planck-Institut für Biophysicalische Chemie in Göttingen, Deutschland, in den Abteilungen von Prof. Dr. Reinhard Lührmann und Prof. Dr. Christian Griesinger, mit dem Titel: *'NMR-spectroscopische Untersuchungen von 15.5K Protein assoziierten Protein-RNA Komplexen'*

# CURRICULUM VITAE

**Name:** Ping Li

**Gender:** Female

**Date of Birth:** 26.06.1981

**Place of Birth:** Beijing, China

**Citizenship:** Chinese

**Family status:** Single, no children

## Education

**1996-1998**   Beijing No.4 High School, Beijing, China

**1998-1999**   Burnside High School, Christchurch, New Zealand

**12/1999**   Completed the high school studies and qualified for the university entrance with grade 'A' in Bursary examinations.

**2000**   Gained direct entry into the second year of university study and studied in the major of chemistry at University of Canterbury, Christchurch, New Zealand.

**2001**   Further study in chemistry at Uppsala University, Sweden, through the exchange program from University of Canterbury, New Zealand.

**2002**   Completed the Bachelor of Science with Honors degree with a first class honor at University of Canterbury, New Zealand.

**02/2003 -2007**   PhD study at Max-Planck-Institute for Biophysical Chemistry in Göttingen, Germany, in the departments of Prof. Dr. Reinhard Lührmann and Prof. Dr. Christian Griesinger. The PhD project title is *NMR spectroscopy investigations of 15.5K protein associated protein-RNA complexes*

**09/2007**   Thesis submitted