

# **Structural Analyses of Spliceosomal Proteins: The Prp19-Complex and the U11/U12 65K Protein**

**Inaugural-Dissertation**

zur Erlangung des Grades eines  
Doktors der Naturwissenschaften

an der  
Fakultät für Medizin



vorgelegt von:  
**Catharina Netter**  
aus Hannover, Deutschland  
Witten 2008

Mentor (Erstgutachter): Prof. Dr. M. C. Wahl  
Fakultätsreferent (Zweitgutachter): Prof. Dr. W. Wintermeyer  
Tag der Disputation: 16.07.2008

**Table of Contents**

Figures	VI
Tables	VII
Declaration	VIII
1 Summary	1
2 Introduction	3
2.1 Chemistry of the Splicing Reaction	4
2.1.1 The Two Classes of Introns	4
2.1.2 The Two-Step Splicing Reaction	5
2.2 Spliceosomes	7
2.3 The Major Spliceosome	7
2.3.1 Components of the Major Spliceosome	7
2.3.2 Assembly and Splicing Cycle of the Major Spliceosome	11
2.3.3 The Prp19-Complex	15
2.4 The Minor Spliceosome	32
2.4.1 Components of the Minor Spliceosome	32
2.4.2 Assembly of the Minor Spliceosome	35
2.4.3 The U11/U12 65K Protein	35
2.4.4 RNA Recognition Motifs and the Mechanism of RNA-Binding	37
2.5 Aim of the Present Study	41
3 Materials and Methods	42
3.1 Materials	42
3.1.1 Chemicals and Fine Chemicals	42
3.1.2 Media and Antibiotics	43
3.1.3 Enzymes and Enzyme Inhibitors	43
3.1.4 Antibodies	44
3.1.5 Nucleotides	44
3.1.6 DNA-Oligonucleotides	45
3.1.7 RNA-Oligonucleotides	45

3.1.8	Plasmids	45
3.1.9	<i>Escherichia coli</i> Strains	45
3.1.10	Common Buffers	46
3.1.11	Commercial Kits	46
3.1.12	Working Materials	47
3.1.13	Technical Devices	47
3.2	Methods	49
3.2.1	Molecular Biology Standard Methods	49
3.2.2	Protein Biochemical Standard Methods	52
3.2.3	Bioinformatic Methods	59
3.2.4	Special Methods	60
3.2.5	Methods for Biophysical Characterization	62
3.2.6	Crystallization	63
3.2.7	Data Collection and Data Processing	66
3.2.8	Structure Solution	68
3.2.9	Model Building and Refinement	69
4	Results	71
4.1	Native Purification and <i>In Vitro</i> -Translation of the Human Prp19-Complex	71
4.2	Recombinant Production of the Human Prp19-Complex Proteins	72
4.2.1	The Human Prp19 Protein	73
4.2.2	The Human Prp19 Coiled-Coil Domain	74
4.2.3	The CDC5 C-Terminal Domain and the PRL1 WD40 Domain	77
4.2.4	Other Proteins from the Human Prp19-Complex	77
4.2.5	Interactions among the Prp19-Complex Proteins	78
4.3	Proteins from the <i>Saccharomyces cerevisiae</i> Prp19p-Complex	80
4.3.1	The Prp19p Protein	81
4.3.2	The Prp19p U-box and Coiled-Coil Domain	85
4.3.3	Other Proteins from the <i>Saccharomyces cerevisiae</i> Prp19p-Complex	86
4.4	Studies on the U11/U12 65K Protein	87
4.4.1	RNA-Binding of Truncated 65K Proteins	87
4.4.2	Crystallization of U11/U12 65K <sup>380-517</sup>	90

4.4.3	Crystal Structure of 65K <sup>380-517</sup>	92
4.4.4	RNA-Binding by the 65K C-Terminal RNA Recognition Motif	95
4.4.5	The Role of the N-Terminus	99
5	Discussion	102
5.1	The Prp19-Complex	102
5.2	The U11/U12 65K C-Terminal RNA Recognition Motif	105
5.2.1	RNA-Binding and the Role of the C-Terminus	105
5.2.2	Functional Stabilization of the 65K C-Terminal RNA Recognition Motif	106
6	Outlook	108
6.1	The Prp19-Complex	108
6.2	The U11/U12 65K C-Terminal RNA Recognition Motif	109
7	References	111
8	Appendices	121
8.1	Details on Protein Expression	121
8.2	DNA-Oligonucleotides	124
8.3	Protein and DNA Sequences	127
8.4	Refinement Statistics for the 65K-cRRM Crystal Structure	128
8.5	Designing Targets for Structural Analyses	129
8.6	List of Abbreviations	133
8.7	Acknowledgements	136
8.8	<i>Curriculum Vitae</i>	138

## Figures

Figure 1: Consensus Sequences of the Two Intron Types	5
Figure 2: Catalytic Steps of the Splicing Reaction	6
Figure 3: Splicing Cycle of the Major Spliceosome	12
Figure 4: Intron Bridging in Early Spliceosomal Complexes	13
Figure 5: Domain Structure of Prp19-Complex Components	16
Figure 6: Proteome Comparison of Spliceosomal Complexes	18
Figure 7: Dynamic Interactions of U5 and U6 with Pre-mRNA and the Role of NTC	21
Figure 8: Model of <i>S. cerevisiae</i> Prp19p	25
Figure 9: Interactions among Prp19-Complex and NTC Components	31
Figure 10: Domain Structure of U11/U12 65K	36
Figure 11: RNA-Oligonucleotides bound by 65K-cRRM, U1A- and U2B''-nRRM	37
Figure 12: Topology and Structure of U1A C-Terminal RRM	39
Figure 13: The Process of Protein Crystallization	64
Figure 14: Human Prp19 Coiled-Coil Analytical Gel Filtration	76
Figure 15: Human Prp19 Coiled-Coil Initial Crystals	76
Figure 16: <i>In Vitro</i> -Translations and Pulldown Assays of Human Prp19 and CDC5	80
Figure 17: Limited Proteolysis of Prp19p	83
Figure 18: Electron Microscopy Negative Stain Images of Prp19p	84
Figure 19: Prp19p U-box and Coiled-Coil Melting Curves	86
Figure 20: Gel Shift Assay with 65K <sup>380-517</sup> and U12-wt RNA	88
Figure 21: Gel Shift Assays of Truncated 65K Proteins	89
Figure 22: Crystals of 65K <sup>380-517</sup>	91
Figure 23: Crystal Structure and Topology of 65K <sup>380-517</sup>	94
Figure 24: Structure Based Sequence Alignment of the Homologous RRMs	96
Figure 25: RNA-Binding of Mutants in Loop 6 and of $\beta$ -Sheet Aromatics	98
Figure 26: Limited Proteolysis of 65K <sup>380-517</sup>	99
Figure 27: Thermal Melting of Truncated 65K Proteins	100

**Tables**

Table 1: Protein Composition of the Human Major Spliceosome	9
Table 2: Proteome Comparison of Major and Minor Spliceosomal Particles	34
Table 3: RNA-Oligonucleotides	45
Table 4: Plasmids	45
Table 5: Commonly Tested Expression Conditions with LB-Medium	54
Table 6: Commonly Tested Expression Conditions with Auto-Inducing Medium	54
Table 7: Conditions for Coexpression	55
Table 8: Analyzed proteins from the human Prp19-complex	73
Table 9: Coexpression of Human Prp19-Complex Proteins	79
Table 10: Analyzed Proteins from the <i>S. cerevisiae</i> Prp19p-complex	81
Table 11: Mass Spectrometric Results of Prp19p Limited Proteolysis	83
Table 12: U11/U12 65K Domains	88
Table 13: Initial Crystallization Trials of 65K <sup>380-517</sup>	91
Table 14: Statistics from the Data Collection	92
Table 15: Insoluble Proteins and Tested Conditions	121
Table 16: Soluble Proteins and Expression Conditions	122
Table 17: DNA-Oligonucleotides	124
Table 18: Identification Numbers for Protein and DNA Sequences	127
Table 19: Statistics for the Refinement of the 65K-cRRM Crystal Structure	128
Table 20: Protein sequence analysis	132

## Declaration

I hereby declare that the following thesis is based on the original results of my own investigations. Work other than my own is clearly indicated in the text by reference to the relevant researchers or the publications. Where the thesis is based on work done in collaboration with others, I have made clear exactly what was done by others and what I have contributed myself.

I wrote the thesis on my own. It was not, in whole or part, previously submitted at another university.

---

Date

---

Catharina Netter



## 1 Summary

The removal of non-coding sequences from precursor-messenger RNA (pre-mRNA) transcripts is an essential step in eukaryotic protein biosynthesis. The reaction, called pre-mRNA splicing, is catalyzed by the spliceosome, a dynamic multi-subunit RNA-protein enzyme. Uridine-rich small nuclear ribonucleoprotein particles (U snRNPs) constitute the main building blocks of the spliceosome. In each round of splicing, the spliceosome assembles anew on the RNA-substrate by the coordinated interaction of the U snRNPs and several non-snRNP components. Two different types of spliceosomes have been discovered: the major and the minor spliceosome. The major spliceosome catalyzes the excision of 99 % of all introns, whereas the minor spliceosome only cleaves a small minority of special introns.

The structure of a major spliceosome-associated complex, the Prp19-complex, from human and from *Saccharomyces cerevisiae* (*S. cerevisiae*) was investigated in the present study. The Prp19-complex is a non-snRNP protein assembly with at least seven components in both organisms. The complex is believed to promote spliceosomal activation prior to catalysis. In this process it may serve as a specificity factor, which defines base pairing interactions of the U5 and the U6 snRNAs with the pre-mRNA. Moreover, the Prp19-complex may be needed during early spliceosomal assembly steps. It is essential for splicing in various organisms. The Prp19p protein in *S. cerevisiae* forms a homo-tetramer and thereby creates a central platform, on which other proteins from the complex assemble.

To obtain the core components from the human and the *Saccharomyces cerevisiae* Prp19-complex, a major focus was laid on recombinant techniques: the full-length proteins and individual domains, which were predicted or identified experimentally, were expressed in *Escherichia coli*. In addition to crystallographic and electron microscopy studies, biophysical characterization and mapping of interaction partners were performed. The results of these investigations were expected to provide more detailed insights into the architecture of this important protein complex.

The present studies revealed that the human Prp19 coiled-coil domain forms a homo-tetramer like its counterpart in yeast. Additionally, an interaction between the human Prp19 coiled-coil domain and the CDC5 protein was detected in a cooperative project. These results demonstrated that the central role, which has been proposed for P1p19p in the yeast complex, is transferable to humans. The human Prp19 coiled-coil domain was subjected to crystallization, but the obtained crystals diffracted poorly and could not be improved.

*S. cerevisiae* Prp19p tetramerization was shown and together with electron microscopic (EM) negative stain images, an existing model for Prp19p homo-oligomerization and domain-

structure was confirmed. The Prp19p coiled-coil domain was mapped in limited proteolysis to be slightly shorter than previously thought. It was demonstrated that this shorter Prp19p coiled-coil domain is still sufficient for tetramer formation. Attempts to crystallize the full-length protein and truncated proteins, which comprised the amino-terminal U-box and the coiled-coil domain, were not successful.

In order to assemble the Prp19-complex from recombinant components and investigate their interaction network, coexpression trials with the human proteins as well as with the yeast proteins were carried out. However, the only detectable interaction was a known interaction between human PRL1 and CDC5. In all other cases, problems to express the proteins in sufficient amounts prevented more in-depth studies on interactions, structure and function of the Prp19-complex proteins.

In a second part of the thesis RNA-binding by the carboxy-terminal RNA recognition motif of the human minor spliceosomal U11/U12 65K protein was analyzed. It is known that this protein builds a molecular bridge between the minor spliceosomal U11 and the U12 snRNPs: the amino-terminus of 65K interacts with the U11-associated 59K protein, whereas the RNA recognition motif (RRM) on its carboxy-terminus is used to bind the U12 snRNA. Bridging the two snRNPs brings the 5' splice site of the pre-mRNA into close proximity of the branch point sequence, as these two sites are bound by U11 and U12 snRNP respectively. The proximity of these pre-mRNA regions is required for the first catalytic step in splicing.

The investigation of the U11/U12 65K carboxy-terminal RRM (65K-cRRM) showed that approximately 30 residues amino-terminally of the RRM core extend this domain and are essential for binding to the U12 snRNA hairpin III. Mutageneses and gel shift assays with an RNA oligomer from the U12 snRNA hairpin III indicated that 65K-cRRM employs a similar RNA-binding mechanism as its close homologs, the N-terminal RRMs of U1A and U2B". However, these two proteins do not require a comparable extension for RNA-binding. The crystal structure of the expanded 65K C-terminal RRM showed that the N-terminal tail adopts an  $\alpha$ -helical conformation and wraps around the protein. Consistent with this architecture, removal of these N-terminal residues significantly decreased the thermal stability of the 65K-cRRM. These results demonstrate that the 65K C-terminal RRM is supplemented by an N-terminal extension, which confers stability on the domain and thereby preserves the RNA-binding capacity.

## 2 Introduction

Eukaryotic genes often exhibit a mosaic structure: sequences that code for amino acids (exons) are interrupted by non-coding sequences (introns). The number of introns in a gene varies extensively from species to species and together with other factors it seems to reflect the complexity of an organism. Only about 4 % of all genes in *Saccharomyces cerevisiae* contain one intron and only six genes have two introns (Lopez and Seraphin 2000). In contrast, human genes have on average 7.8 introns per gene (Lander, Linton *et al.* 2001).

During protein biosynthesis of intron-containing genes a primary transcript, called precursor-messenger RNA, is synthesized by RNA polymerase II. In three coupled, posttranscriptional modification steps the pre-mRNA is processed in the nucleus to a mature mRNA (summarized in (Maniatis and Reed 2002)). These RNA-processing steps are capping, splicing and polyadenylation. During capping, guanosine-5'-triphosphate is added via a 5'-5'-pyrophosphate bond and methylated to an m<sup>7</sup>G-cap. During splicing, introns are removed from the transcript and the exons are ligated to a continuous reading frame. In the last processing step the 3'-end of the pre-mRNA is cut and numerous adenine-nucleotides are added (poly-A tail). Finally the mature mRNA is exported to the cytoplasm and serves as a template for protein biosynthesis.

Pre-mRNA splicing is catalyzed by the spliceosome (reviewed in (Brow 2002)), which is mainly composed of five uridine-rich small nuclear ribonucleoprotein particles (U snRNPs: U1, U2, U4, U5 and U6). Each U snRNP contains a unique snRNA and a set of snRNP proteins. To generate the functionally active spliceosome, the U snRNPs together with numerous non-snRNP factors assemble in an ordered manner newly for each round of splicing on the pre-mRNA substrate. The following precise excision of an intron and ligation of the exons in a two-step transesterification reaction is achieved by a multitude of RNA-RNA, RNA-protein and protein-protein interactions.

The chemistry of the splicing reaction and the roles of the snRNAs are well understood by now. However, structural and functional roles of many protein components remain entirely unclear. As for example more than 200 proteins have been co-purified with human spliceosomal complexes, the major part of the splicing machinery still needs to be explored (Will and Luhrmann 2006). In various macromolecular assemblies, such as RNA polymerases, the exosome or the ribosome, structural studies have provided valuable insights into functional details (Ban, Nissen *et al.* 2000; Cramer, Bushnell *et al.* 2001; Schuwirth, Borovinskaya *et al.* 2005; Shen and Kiledjian 2006). Unfortunately structural investigations of the entire spliceosome and its active site(s) at atomic resolution are presently not possible because of its very dynamic behavior. Therefore characterization and

structures of catalytic subunits, including individual U snRNPs and single proteins, constitute important steps towards a more complete picture of the splicing machinery. In order to contribute to this effort, the structure of a multi-protein complex, which associates with the major spliceosome, as well as the structure of an important protein domain from the minor spliceosome were investigated in the course of this work.

The following chapters will first explain the chemistry behind the splicing reaction. Afterwards the composition of the major spliceosome and its assembly pathway will be introduced. A special focus will be laid on the Prp19-complex, which was studied during this thesis. The second part of this introduction is dedicated to the minor spliceosome with a particular emphasis on the U11/U12 65K protein.

## 2.1 Chemistry of the Splicing Reaction

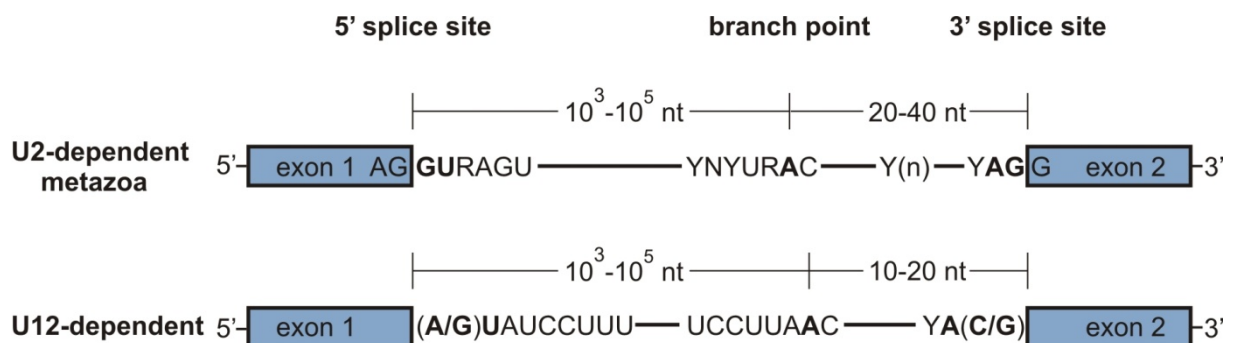
### 2.1.1 The Two Classes of Introns

Recognition of the intron-exon boundaries is a critical step during splicing. Missing the splice-site by only a single nucleotide could already produce a non-functional RNA. To ensure proper identification of the splice-sites, introns have highly conserved sequences at their 5' and 3' ends and at the branch point (summarized in (Burge, Padgett *et al.* 1998; Burge, Tuschl *et al.* 1999)). On the basis of these sequences two different types of introns can be distinguished.

Referring to the first and last two nucleotides of the non-coding sequence, one type of intron is called the "GT-AG"-type. Over 99 % of all non-coding sequences belong to this class. Their branch point sequence is recognized by the RNA component of the U2 snRNP and they are spliced by the so called major or U2-dependent spliceosome. GT-AG (or U2-type) introns have a 5'-splice site (5'-ss), which is composed of eight highly conserved nucleotides (Figure 1). The 3'-splice site (3'-ss) consists of two components: the actual 3'-ss and the branch point (Reed 1989). The branch point consensus sequence is YNYUR**A**C, where the highly conserved adenosine (bold letter) is the most important component. An additional element, the polypyrimidine-tract is found in GT-AG introns of higher eukaryotes between the 3'-ss and the branch point. Most likely these 10-15 pyrimidines compensate for the lower sequence conservation of the branch point sequence in higher eukaryotes.

The second intron class with the terminal di-nucleotides AT-AC was discovered around 15 years ago (Figure 1; (Hall and Padgett 1994; Tarn and Steitz 1997; Wu and Krainer 1999; Patel and Steitz 2003)). They occur in metazoa, but are absent in lower organisms such as yeast or nematodes. Although AT-AC-type introns represent only 0.4 % of all human introns

(Burge, Padgett *et al.* 1998; Levine and Durbin 2001), they are found in genes whose products carry out essential cellular functions (e.g., DNA replication and repair, transcription, RNA processing and translation). AT-AC-introns can co-exist in the same gene with U2-spliced introns. Their branch point sequence is recognized by the RNA-component of the U12 snRNP from the minor or U12-dependent spliceosome. Therefore they are also called U12-type introns. The 5'-ss and the branch point consensus sequences are far more conserved than in the GT-AG introns. There are no polypyrimidine-tracts in U12-type introns. Later it became evident that AT-AC introns can also have GT-AG ends, but the described sequences in the middle of the intron are still different to the ones, which are spliced by the major spliceosome (Dietrich, Inorvaia *et al.* 1997).



**Figure 1: Consensus Sequences of the Two Intron Types**

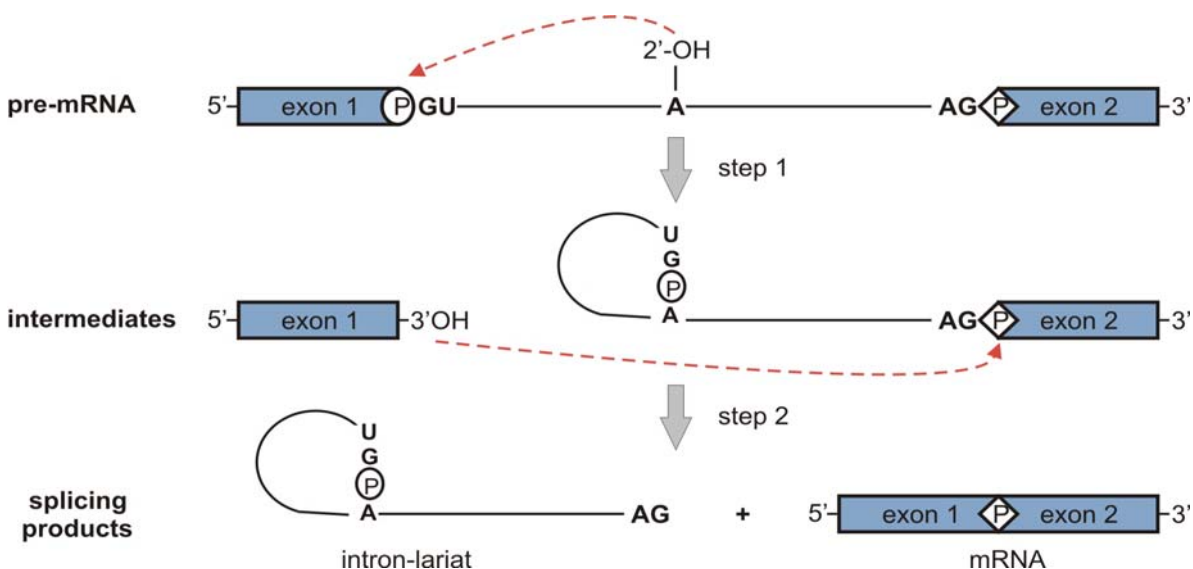
The consensus sequences of the 5'- and 3'-splice sites, the branch point and the polypyrimidine-tract (Y(n)) are shown. Highly conserved nucleotides are shown in bold letters. Y: pyrimidine, R: purine, N: any nucleotide. Exons are shown as boxes, introns are represented by a line. The numerical values above the introns inform about the average length in nucleotides (nt).

### 2.1.2 The Two-Step Splicing Reaction

Chemically, the splicing reaction consists of two subsequent transesterification reactions (Figure 2) (Sharp 1987; Moore, Query *et al.* 1993). In the first reaction the pre-mRNA is cleaved at the 5'-exon/intron boundary by a nucleophilic attack of the branch point adenosine's 2'-hydroxyl group (OH) on the phosphodiester bond at the 5'-splice site. This reaction produces a lariat structure of the intron, which is still connected to the 3'-exon, and a 5'-exon with a free 3'-OH group. In the second catalytic step this free 3'-OH group of the 5'-exon nucleophilically attacks the phosphodiester bond at the 3'-exon/intron boundary. The latter reaction leads to ligation of the exons and release of the intron in form of a lariat. While the mRNA is transported to the cytoplasm, the intron is debranched by a nuclear

2'-5'-phosphodiesterase. Normally the debranched intron is entirely degraded. However, sometimes pieces code for small functional non-coding RNAs like small nuclear RNAs.

The basic chemistry of pre-mRNA splicing is similar to that of the group II or self-splicing introns, which catalyze their own removal and require no protein cofactors (Michel and Jacquier 1987; Cech 1990; Cech 1990). Self-splicing introns and nuclear pre-mRNA splicing have a few things in common: both form an intron lariat, the stereochemistry of the reactions is the same (Maschhoff and Padgett 1993; Moore, Query *et al.* 1993; Moore and Sharp 1993; Padgett, Podar *et al.* 1994) and both reactions require coordinated metal ions (Sontheimer, Sun *et al.* 1997; Sontheimer, Gordon *et al.* 1999). It is therefore presumed that the actual reaction in nuclear pre-mRNA splicing is catalyzed by the snRNAs (Yean, Wuenschell *et al.* 2000). In contrast to self-splicing introns, nuclear pre-mRNA splicing requires an input of energy and protein cofactors that are organized in the spliceosome.



**Figure 2: Catalytic Steps of the Splicing Reaction**

In the first step of the splicing reaction, the 2'-OH group of the branch point adenosine carries out a nucleophilic attack on the phosphodiester bond of the 5'-splice site. In the second catalytic step, the resulting free 3'-OH group of the 5'-exon performs another nucleophilic attack on the phosphodiester bond on the 3'-exon/intron boundary. The nucleophilic attacks are shown as dashed red arrows.

## 2.2 Spliceosomes

Two distinct types of spliceosomes have been identified: the major and the minor spliceosome (reviewed by (Padgett and Shukla 2002); (Burge, Tuschl *et al.* 1999)). The terms “major” and “minor” refer to the abundance of the respective spliceosomes in the cell: the minor spliceosome is far less abundant than the major spliceosome (Burge, Tuschl *et al.* 1999).

U snRNPs are the main building blocks of both spliceosomes. In each type of U snRNP, different splicing factors are arranged around one U snRNA, which is name-giving to the respective U snRNP. There are two groups of U snRNP-associated proteins: the proteins, which are common to all U snRNPs, and the proteins, which are only associated with a certain U snRNP, the so-called particle-specific proteins. In addition to these common and particle-specific splicing factors, non-snRNP factors transiently or permanently associate with the spliceosome.

## 2.3 The Major Spliceosome

### 2.3.1 Components of the Major Spliceosome

#### The U snRNAs

Each U snRNP contains one central RNA-molecule so that there are five uridine-rich small nuclear RNAs (U snRNAs), named U1, U2, U4, U5 and U6 snRNA. In general, U snRNAs are characterized by their small size (between 106 and 187 nucleotides), their metabolic stability and the high degree of conservation of their primary and secondary structure (Guthrie and Patterson 1988). With  $1 \times 10^6$  copies per cell, the major U snRNAs are very abundant.

Except for U6, all U snRNAs are transcribed by RNA-polymerase II and have a 2,2,7-trimethylguanosine cap ( $m^3G$ -cap) on their 5'-ends (Mattaj 1986). Another characteristic feature for all U snRNAs apart from U6 is their Sm-binding site, which is the uracil-rich consensus sequence YAU<sub>3-6</sub>GY (Y is a pyrimidine). This single-stranded region is usually flanked by hairpins. It serves as the binding site for the Sm-core-proteins (see page 8; (Raker, Hartmuth *et al.* 1999)). The U6 snRNA is exceptional as it is transcribed by RNA-polymerase III, it has a  $\gamma$ -monomethyl-phosphate on its 5'-end and it does not have a Sm-binding site (Kunkel, Maser *et al.* 1986; Reddy, Henning *et al.* 1987; Singh and Reddy 1989). All U snRNAs possess several post-transcriptionally modified nucleotides like pseudo-uridine ( $\Psi$ ), 6-methyladenosine and 2'-O-methyl ribose (summarized in (Massenet, Mouglin *et al.*

1998)). These modifications presumably participate in the modulation of RNA-RNA interactions and the spliceosome assembly as shown for the U2 snRNA (Yu, Shu *et al.* 1998).

U snRNAs exhibit complex secondary structures. U1 snRNA for example forms a cloverleaf structure. Unlike all other U snRNAs, U4 and U6 snRNA are not single entities, but they form a complex via base pairing with a characteristic Y-shaped interaction domain (Rinke, Appel *et al.* 1985; Brow and Guthrie 1988).

### The Common Proteins

U1, U2, U4 and U5 snRNPs all contain the seven Sm-proteins B/B', D1, D2, D3, E, F and G. These proteins are therefore called the common proteins (table 1). B and B' are isoforms of the same protein. The Sm-proteins are relatively small and basic. They are characterized by a highly conserved sequence motif, the Sm-motif, which can mediate protein-protein as well as protein-RNA contacts (Hermann, Fabrizio *et al.* 1995; Seraphin 1995; Kambach, Walke *et al.* 1999; Toro, Thore *et al.* 2001; Urlaub, Raker *et al.* 2001). Sm-proteins are known to initiate snRNP-assembly in the cytoplasm by associating with the conserved Sm site in the U1-U5 snRNAs upon their export from the nucleus (Branlant, Krol *et al.* 1982; Mattaj and De Robertis 1985). By this interaction the characteristic heptameric Sm-ring is formed. The association of the Sm-proteins allows hyper-methylation of the 5'-m<sup>7</sup>G cap of the snRNAs to a 2,2,7-trimethylguanosine (m<sup>3</sup>G) cap, which together with the Sm-proteins provides a signal for import into the nucleus (Fischer and Luhrmann 1990; Hamm, Dathan *et al.* 1990).

A second set of proteins, which carry the Sm-motif, have been identified and, because of the similarity to the Sm-proteins, these proteins were called LSm-proteins (Like Sm-proteins; table1). Analogous to the Sm-proteins, LSm2-8 form a ring-shaped structure, but they specifically bind the uridine-rich region in the 3'-end of U6 snRNA (Achsel, Brahms *et al.* 1999; Vidal, Verdone *et al.* 1999; Schneider, Will *et al.* 2002).

### The Particle-Specific Proteins

After the import of the U snRNAs with the Sm-ring to the nucleus, the remaining snRNP-specific proteins associate with the particles to complete maturation (Zieve and Sauterer 1990). These specific proteins mediate the distinct functions of the respective U snRNPs. Depending on the salt concentration, functional complexes of different sedimentation constants can be isolated. The proteome of the functional complexes is shown in Table 1.



name	kDa	U snRNP proteins				
		12S	17S	13S	20S	25S
		U1	U2	U4/U6	U5	U4/U6/U5
B/B' LSM8	18/29	●	●	●●	●	●●
D3 LSM4	18	●	●	●●	●	●●
D2 LSM3	16.5	●	●	●●	●	●●
D1 LSM2	16	●	●	●●	●	●●
E LSM5	12	●	●	●●	●	●●
F LSM6	11	●	●	●●	●	●●
G LSM7	9	●	●	●●	●	●●
70K	70	●				
A	34	●				
C	22	●				
A'	31		●			
B''	28.5		●			
SF3a	120		●			
	66		●			
	60		●			
SF3b	155		●			
	145		●			
	135		●			
	49		●			
	p14		●			
	14		●			
	10		●			
	90			●		●
	61			●		●
	60			●		●
	20			●		●
	15.5			●		●
	220				●	●
	200				●	●
	116				●	●
	102				●	●
	100				●	●
	52				●	
	40				●	●
	15				●	●
	100				●	
	65				●	
	27				●	

**Table 1: Protein Composition of the Human Major Spliceosome**

The common and specific proteins of the major U snRNPs are listed (according to (Kastner 1998)). The Svedberg-constants (S) are given for the individual U snRNPs. Only the proteins, which have been detected in stoichiometric amounts are included. The common Sm-proteins are shown in dark grey, the LSm-proteins in light grey.

The mammalian 12S U1 snRNP comprises three particle-specific proteins: U1 70K, A and C. The two largest proteins, U1 70K and U1A, bind directly to U1 snRNA stem-loop I and II, respectively (Nagai and Mattaj 1994). The structure of the amino-terminal fragment of U1A in complex with an RNA-oligonucleotide of stem II has been solved by X-ray crystallography (Oubridge, Ito et al. 1994). This structure will be discussed later in more detail, as U1A is a close homolog of the U11/U12 65K protein.

The 17S U2 snRNP contains two stably associated proteins: U2A' and U2B'' (Luhmann, Kastner et al. 1990; Tang, Abovich et al. 1996; Caspary and Seraphin 1998), and two more weakly associated multisubunit complexes: SF3a and SF3b (Brosi, Hauri et al. 1993). *In vivo* U2A' and U2B'' form a dimer, which binds to stem-loop IV of the U2 snRNA. U2B'' is another close homolog to the U11/U12 65K protein and will be discussed later in more detail as well. Almost all proteins of SF3b and every protein in SF3a have been shown to establish contacts to nucleotides around the branch point. SF3b 14a/p14 has even been crosslinked directly to the branch point adenosine (Gozani, Feld et al. 1996; Query, Strobel et al. 1996; Will, Schneider et al. 2001). Thus it is assumed that the SF3a and SF3b proteins are in close proximity to the catalytic core of the spliceosome.

Base pairing between U4 and U6 snRNAs leads to the formation of a hetero-dimeric complex. This 13S U4/U6 di-snRNP is composed of the Sm- and LSm-proteins and five particle-specific proteins. Together with the 20S U5 snRNP, which contains seven particle-specific proteins, the di-snRNP associates to form a 25S U4/U6.U5 tri-snRNP. In this tri-snRNP the Y-shaped U4/U6 interaction domain is preserved. During spliceosome assembly the tri-snRNP is integrated as a pre-formed unit.

### The Non-snRNP Proteins

Numerous non-snRNP proteins associate permanently or transiently with the spliceosome. A definite number cannot be given as the identification of spliceosomal components is not yet complete. Some of these proteins contain common conserved domains, which allow a grouping of the non-snRNP proteins. The two most extensively studied of these groups are the SR-proteins and the DEXD/H-box helicases.

The SR-proteins are characterized by a variable number of serine-arginine di-peptides in their carboxy-terminus (summarized in (Fu 1993)). N-terminally these highly phosphorylated proteins possess one or two RNA recognition motifs (RRMs; also called RNA-binding domains, RBDs). This modular structure allows them to interact simultaneously with RNA via their RBDs and with proteins via the SR-domain. In higher eukaryotes SR-proteins are essential and exert diverse functions: they recruit snRNPs in the early assembly-stages of the spliceosome and they regulate alternative splicing (Berget 1995; Shin and Manley 2002; Huang and Steitz 2005).

Another non-snRNP protein family is the one of the DEXD/H-box helicases. The name originates from the common consensus sequence in the Walker B motif of the ATP-binding domain. Most of these proteins are ATP-dependent RNA-helicases, which facilitate the dynamic rearrangements during the splicing cycle by the unwinding or annealing of

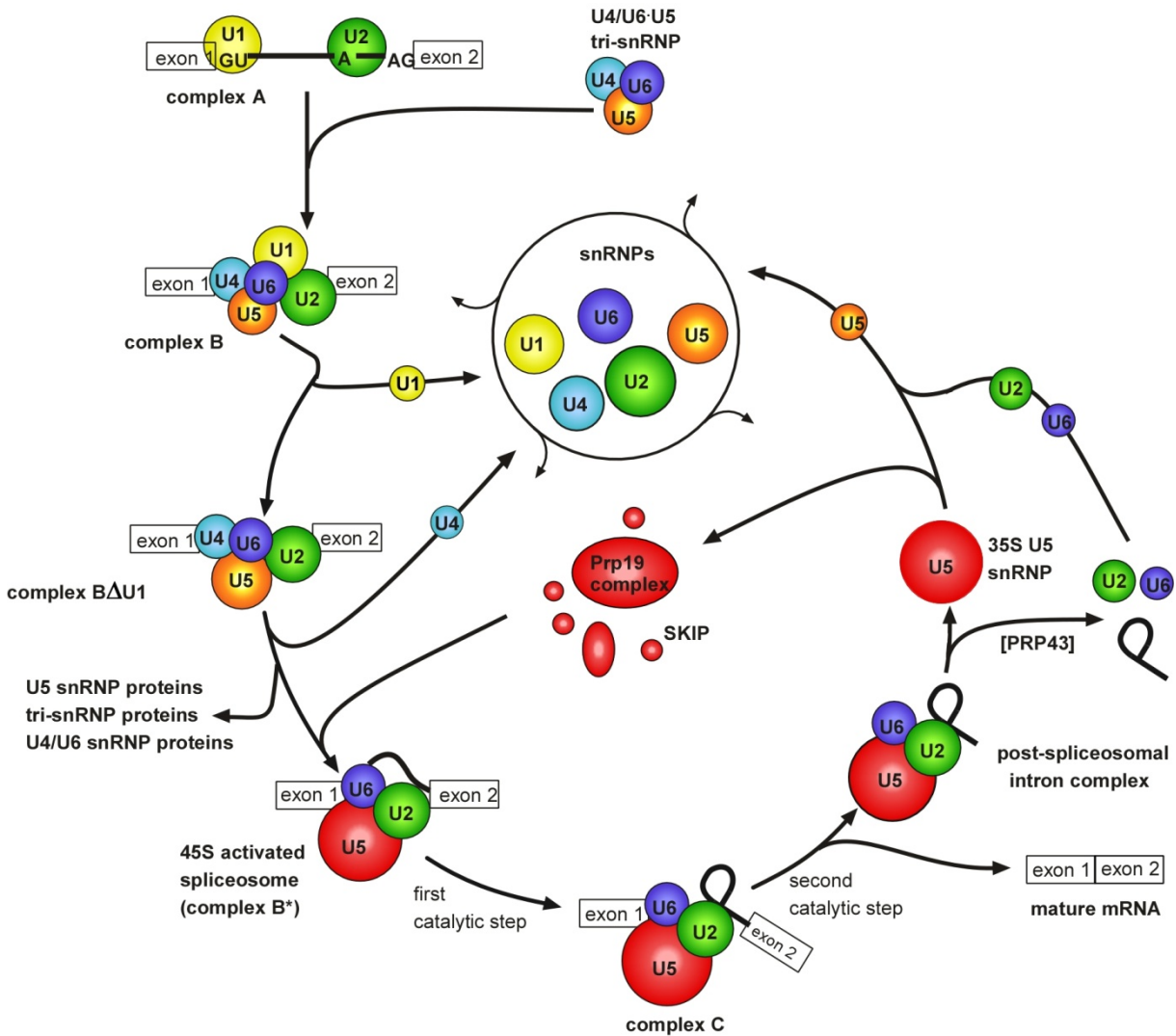
RNA-duplexes (Jankowsky, Gross et al. 2001; Jankowsky, Fairman et al. 2005; Yang and Jankowsky 2005) and the weakening of RNA-protein interactions (Jankowsky, Gross et al. 2001).

Some of the non-snRNP proteins exist as pre-formed complexes. An example for such a pre-organized non-snRNP complex is the Prp19-complex. This complex associates with the spliceosome during the splicing cycle and it is believed to promote spliceosomal activation prior to catalysis. The Prp19-complex in humans and yeast is a main subject of the present study and will therefore be discussed in more detail in chapter 2.3.3 after presenting the assembly pathway of the major spliceosome.

### 2.3.2 Assembly and Splicing Cycle of the Major Spliceosome

Different to many other catalytic particles, the spliceosome does not exist as a pre-formed, active particle in the nucleus. It has to be formed *de novo* for each splicing reaction. Moreover, the spliceosome is subjected to significant rearrangements before, during and after the splicing reaction. Certain spliceosomal intermediates of this rearrangement process with a characteristic snRNP content and protein composition could be isolated and characterized. The isolation of these complexes, led to the postulation of a stepwise assembly pathway of the spliceosome (Reed and Palandjian 1997), which is summarized in Figure 3 and described in detail in the following paragraphs.

Even prior to binding of the first snRNP, the pre-mRNA is bound by heterogeneous nuclear ribonucleoproteins (hnRNPs) in the H-complex. The actual assembly of the spliceosome is initiated by the energy-independent interaction of the U1 snRNP with the 5'-splice site. In this step, base pairing of the 5'-end of the U1 snRNA with the 5'-splice site leads to formation of the early- or E-complex. This interaction is stabilized by the U1-specific proteins U1-C and U1-70K and by proteins from the SR-family (Figure 4 A; (Heinrichs, Bach et al. 1990; Kohtz, Jamison et al. 1994; Reed 1996; Du and Rosbash 2002)). Initial recognition of the 3'-splice site occurs at this stage as well: the branch point is bound by the splice factor 1 (SF1), whereas the polypyrimidine tract and the 3'-AG are contacted by U2AF65 and U2AF35 respectively (Berglund, Abovich et al. 1998; Wu, Romfo et al. 1999). The latter two proteins are subunits of the U2 auxiliary factor U2AF.



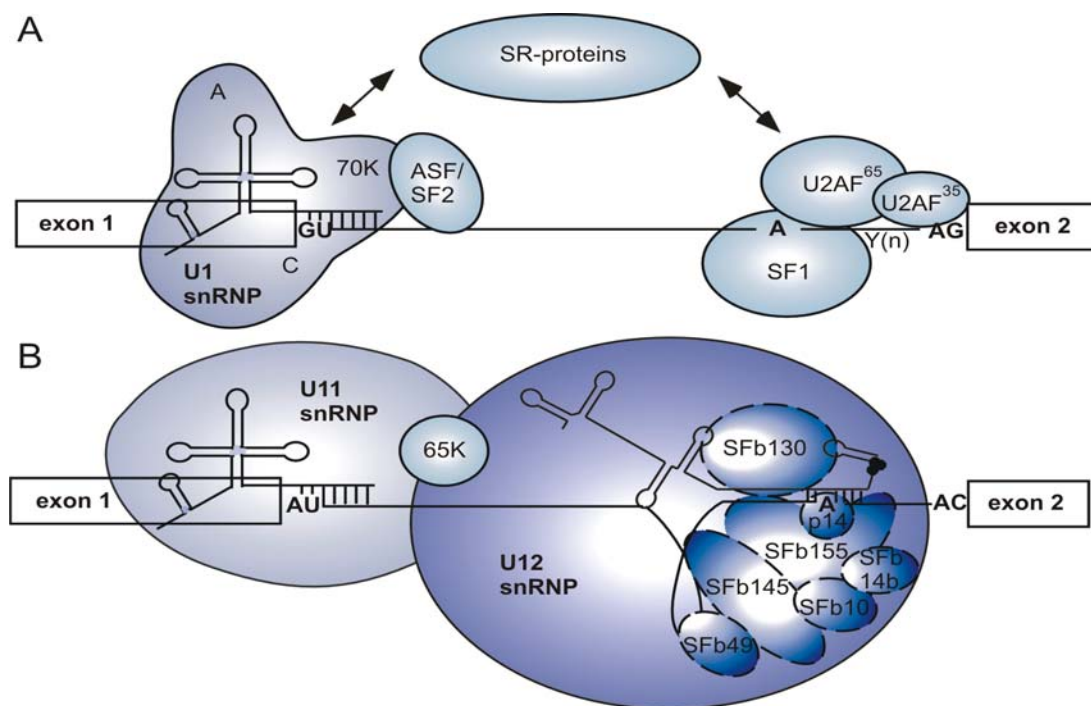
**Figure 3: Splicing Cycle of the Major Spliceosome**

The ordered interaction of snRNPs (colored circles) during the spliceosomal cycle is shown. Complexes, which can be resolved in mammalian splicing extracts by biochemical methods (e.g. native gel electrophoresis or immuno-affinity selection) are represented. The Prp19-complex proteins and Prp19-related proteins are shown as red ellipses. Their stable incorporation into the spliceosome is indicated by the color-switch of U5 from orange to red upon integration of these proteins. Boxes indicate exon sequences; solid lines are intron sequences. The first and last two nucleotides and the branch point adenosine of the intron are also initially shown.

At this early stage of the splicing cycle a molecular bridge is formed between the U1 snRNP, the SF1 and the U2AF65 (Figure 4 A). This molecular bridge is called intron bridge, as it reaches across the entire intron. Intron bridging brings the 5'-ss and the 3'-ss in close proximity and helps to stabilize the U1 snRNP interaction with the 5'-splice site. A known intron bridging factor in the major spliceosome is the SR-protein ASF/SF2, which uses its

SR-domain to simultaneously contact U1 70K, which is bound to the 5'-ss, and U2AF35, which is bound to the 3'-ss (Kramer 1996).

The next step in the splicing cycle is A-complex formation, which is the ATP-dependent association of the 17S U2 snRNP with the branch point. Complementary base pairs are established between the U2 snRNA and the branch point, whereupon the conserved branch point adenosine is bulged out and positioned optimally for the splicing reaction (Query, Moore et al. 1994). Proteins of the SF3a- and SF3b-complexes contact the pre-mRNA in a region 30 nucleotides upstream of the branch point adenosine, called the anchoring site. Thereby the SF3a- and SF3b-complexes facilitate spliceosomal A-complex formation (Gozani, Feld et al. 1996; Query, Strobel et al. 1996).



**Figure 4: Intron Bridging in Early Spliceosomal Complexes**

**A** Schematic depiction of interactions in the major spliceosomal E complex between factors bound at the 5'-ss with those bound at the branch point/3'-ss across an intron. The ability of SR-proteins to bridge the 5'- and the 3'-ss is indicated by arrows. **B** Intron bridging in the minor spliceosomal A complex. Due to concomitant binding of the U11 and U12 snRNP, proteins of this di-snRNP appear to be involved in intron bridging. The U11/U12 65K protein overlaps with both snRNPs to show its involvement in the intron bridge. In the major spliceosome, subunits of SF3b (dashed ellipses) contact the pre-mRNA near the branch site; evidence for analogous interactions in the minor spliceosome is currently lacking. The snRNAs are shown schematically. Boxes indicate exon sequences; solid lines are introns; Y(n) is the polypyrimidine tract. The figure is adapted from (Will and Luhrmann 2005).

The following integration of the 20S U4/U6·U5 tri-snRNP generates the B-complex. Subsequently, a number of dynamic conformational changes occur to form the catalytically active B\* spliceosome. The Prp19-complex plays a pivotal role in this snRNP-remodeling (see chapter 2.3.3). The first rearrangement is the disruption of the U1 snRNA base pairing with the pre-mRNA 5'-splice site. This causes U1 snRNP dissociation and thereby generates the B $\Delta$ U1 spliceosome. The newly released 5'-splice site can afterwards interact with the U6 snRNA. The catalytically active B\* spliceosome is then build by the disruption of the Y-shaped U4/U6 di-snRNA complex, which leads to U4 snRNP dissociation from the spliceosome. In the human spliceosome the Prp19-complex together with Prp19-related proteins become incorporated during this step (Figure 3, Figure 6). After U4 snRNP-release, the U6 snRNA is free to undergo new base pairing interactions (Lamond, Konarska et al. 1988). U6 snRNA forms three short helices with the U2 snRNA and a new intra-molecular helix, which is involved in coordinating a catalytically essential magnesium ion (Sontheimer, Sun et al. 1997). At least two of the U2/U6 snRNA helices together with the U5 snRNA stem-loop I and the U6 intra-molecular helix participate in the spliceosomal active site(s) (Sontheimer and Steitz 1993; Newman 1997).

The presumably RNA-catalyzed first transesterification reaction leads to formation of the C-complex. The second, possibly also RNA-catalyzed reaction brings about the post-spliceosomal complex as well as the mature mRNA. The mRNA is transported to the cytoplasm in form of an mRNP. The post-spliceosomal complex, which consists of U2, U5, U6 and the Prp19-complex proteins, remains in the nucleus.

The nineteen related proteins 1 and 2 (Ntr1, Ntr2) in complex with the DEXD/H-box helicase Prp43 participate in the catalysis of the post-spliceosomal complex disassembly (Tsai, Fu et al. 2005). In this disassembly a novel form of the U5 snRNP, the so-called 35S U5 snRNP, is released. Most likely, 35S U5 snRNP is already generated in the course of the snRNP-remodeling events prior to the first catalytic step of splicing (Figure 3; (Makarov, Makarova et al. 2002; Makarova, Makarov et al. 2004). The Prp19-complex proteins become incorporated into the spliceosome as a sub-complex of the 35S U5 snRNP. Apart from the Prp19-complex proteins the 35S U5 snRNP contains the U5-specific proteins and Prp19-related proteins, which are known to associate with the Prp19-complex. The 35S U5 particle remains nearly unchanged until it is released from the post-spliceosomal complex. After the release 35S U5 snRNP is converted to the 20S U5 snRNP, which is then, like all other U snRNPs, recycled for a new round of splicing.

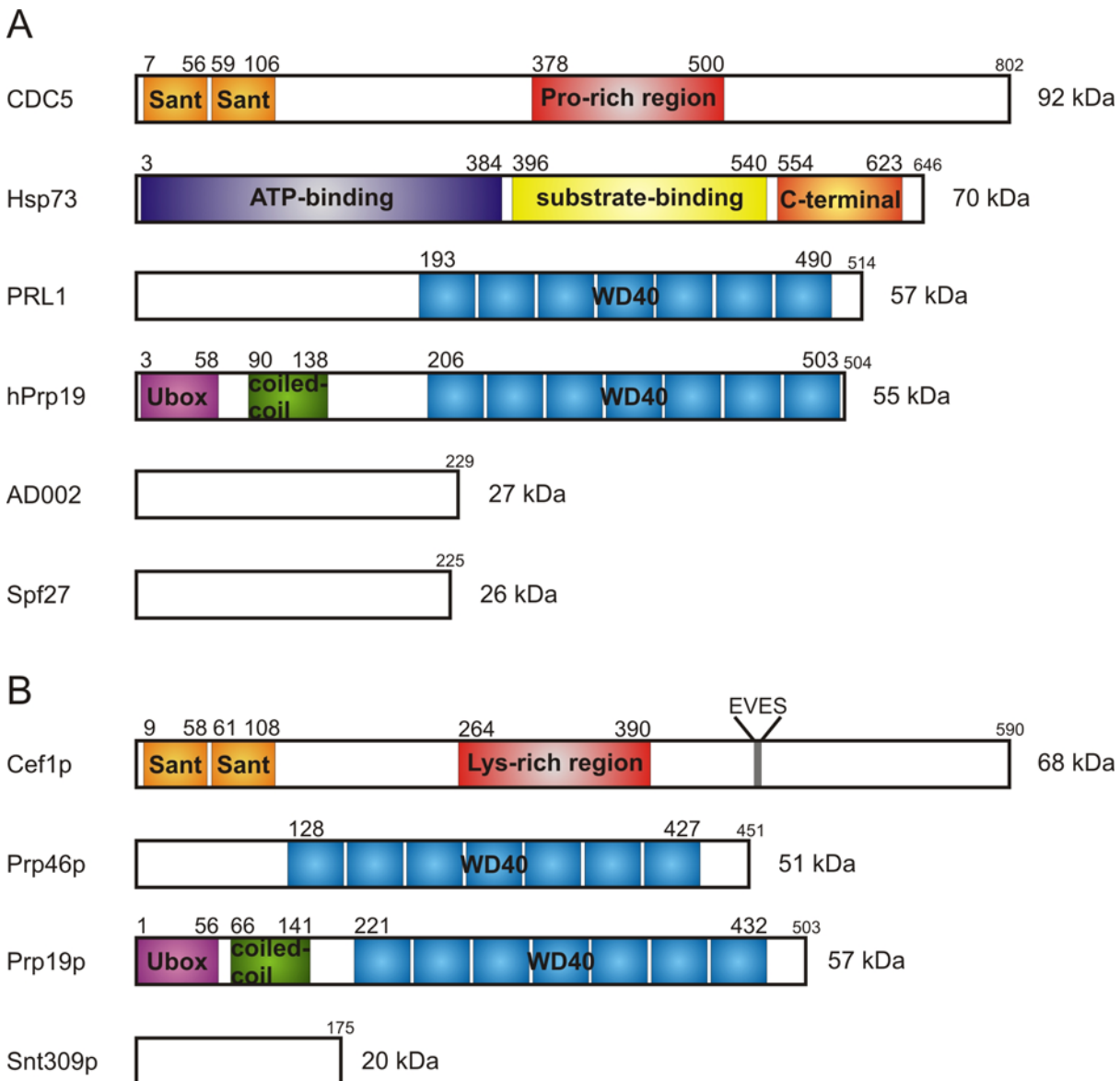
As an alternative to stepwise spliceosome assembly, the so-called holo-spliceosome-model was proposed (summarized in (Brow 2002)). According to that model, a pre-formed penta-snRNP, which comprises all U snRNPs and some splicing factors, exists in the absence of

pre-mRNA. All U snRNPs bind simultaneously to the pre-mRNA. The initially weak interactions with the pre-mRNA are stabilized in a stepwise manner by reorganization of the particle until the catalytically active spliceosome emerges. The complexes of the stepwise assembly model detected *in vitro* are thought to correspond to stages of the spliceosome, in which certain components are bound strongly whereas other more loosely attached components were disassembled during purification. Which of the models is true or whether both positions have to converge is currently unknown.

### 2.3.3 The Prp19-Complex

The Prp19-complex is essential for pre-mRNA splicing, since its depletion, disruption of its structure or deletion of components lead to a loss of splicing activity. The complex becomes stably incorporated into the spliceosome during spliceosomal activation. In *S. cerevisiae* it has been shown to serve as a specificity factor for the base pairing between U5 and U6 snRNA with the pre-mRNA. It may also be required for correct spliceosome assembly. The protein composition of this non-snRNP complex is subject of an ongoing discussion and seems to be highly dependent on the organism and the method of purification. Diverse studies in human, *S. cerevisiae* and *Schizosaccharomyces pombe* (*S. pombe*) shed light on components and possible functions of the Prp19-complex and shall be discussed here. Subsequently the structure and function of individual complex proteins will be introduced.

The human Prp19-complex, which is also called CDC5- or Prp19-CDC5-complex, has first been isolated by immuno-affinity purifications with antibodies against CDC5 (Ajuh, Kuster et al. 2000). Its name is related to one of its central components the human pre-mRNA processing protein 19 (hPrp19; see page 23). Mass spectrometric analysis revealed seven core components and around twenty additional proteins, which co-purify with the complex. Later studies, that used more stringent purification conditions, narrowed the number of components down to seven core proteins and two loosely attached proteins (Tarn, Hsu et al. 1994; Makarova, Makarov et al. 2004). Since association under stringent conditions requires stronger interactions, which in turn may ease structural investigations, the complex defined by Makarova *et al.* was taken as a basis for the present study. It consists of hPrp19, CDC5, PRL1, CCAP1/Hsp73, CCAP2/AD002 and SPF27 (Figure 5). The catenin  $\beta$ -like 1 protein, which has also initially been detected in the complex, was regarded as an impurity, because of its low abundance. The complex sediments at 14S in a glycerol gradient and has an approximate mass of ~500 kDa.



**Figure 5: Domain Structure of Prp19-Complex Components**

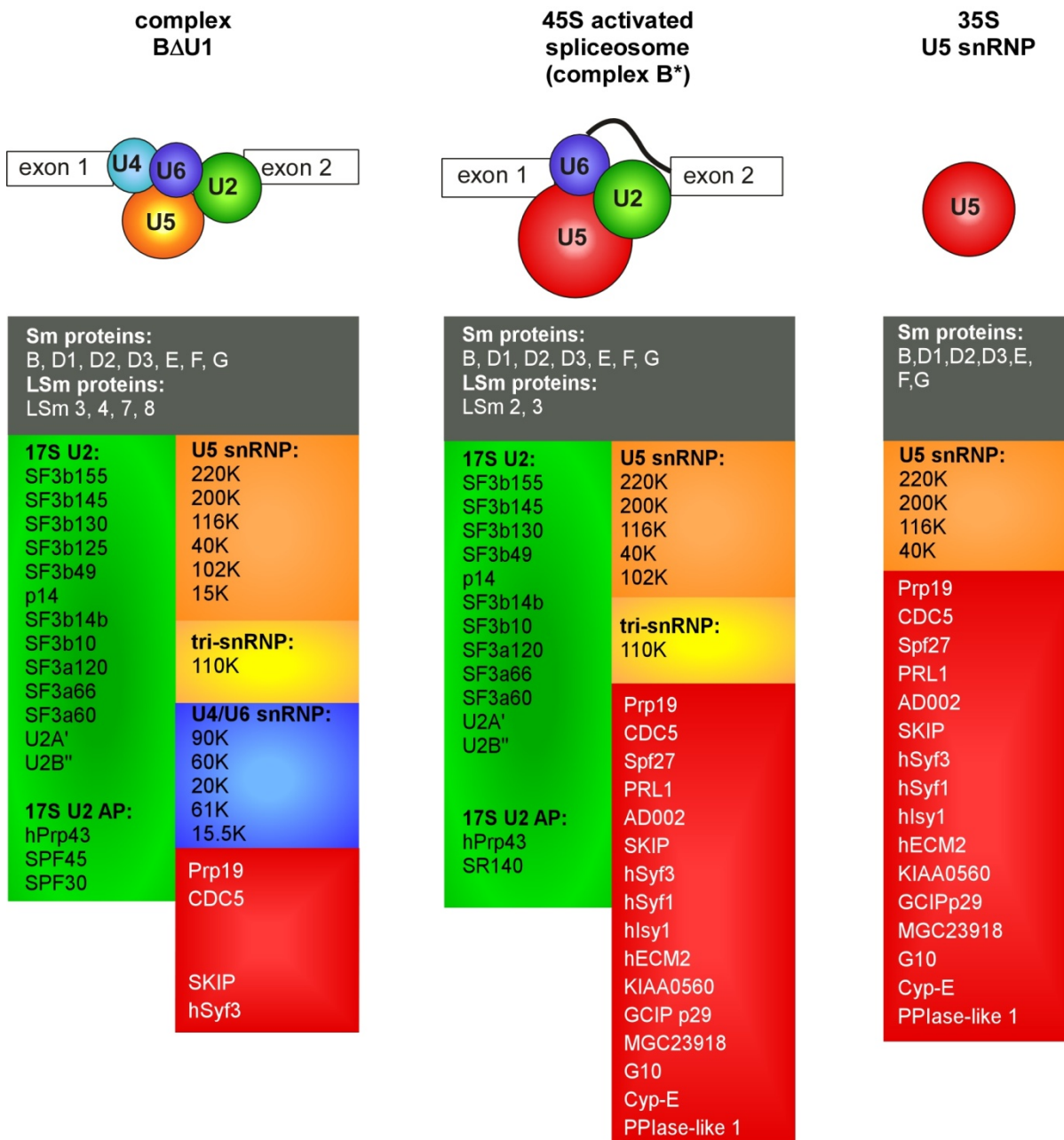
The domain architecture of the human Prp19-complex members and their *S. cerevisiae* homologs are shown. Only those proteins, which were investigated in the course of this study, are shown. The numbers above the domains represent the N- and C-terminal amino acid in the respective domain. The total number of amino acids is shown in smaller letters at the end of each protein. The protein sizes are not scaled. **A** Predicted domains of the human Prp19-complex proteins. The Hsp73 domains are not predicted, but result from an alignment with Hsp70, whose structure is known. Additionally, the domain borders of the hPrp19 U-box were determined by an alignment with the *S. cerevisiae* Prp19p U-box domain. **B** Predicted domains of the *S. cerevisiae* homologs. All Prp19p domain borders apart from the U-box are shown according predictions made by Ohi *et al.* (Ohi, Vander Kooi *et al.* 2005). The visible residues in the Prp19p U-box crystal structure (Vander Kooi, Ohi *et al.* 2006) were chosen as domain borders for the U-box. The Cef1p EVES-sequence motif is shown as well.



In *S. cerevisiae* and *S. pombe* Prp19p is part of the so-called nineteen complex (NTC). This complex contains at least eleven core proteins in *S. cerevisiae*, of which only Prp19p (hPrp19), Cef1p (CDC5) and Snt309p (a functional homolog of Spf27) are homologous to proteins from the human complex (Figure 5; (Tsai, Chow et al. 1999; Chen, Yu et al. 2002; Ohi, Link et al. 2002). Additionally, Prp46p (PRL1) and Cwc15p (AD002) could be detected as more loosely attached components. These five proteins were also present in *S. pombe* (McDonald, Ohi et al. 1999). In *S. pombe* and *S. cerevisiae*, the entire isolated complexes contain around 30 proteins (McDonald, Ohi et al. 1999; Ohi, Link et al. 2002).

Several lines of evidence suggest that the Prp19-complex functions as an integral unit in human and yeast: the core proteins always co-precipitate even under harsh purification conditions (high salt or heparin). Complementation with a single component does not restore splicing in a depleted extract, whereas addition of the purified complex can complement the depletion (Tarn, Hsu et al. 1994; McDonald, Ohi et al. 1999; Ajuh, Kuster et al. 2000; Makarova, Makarov et al. 2004; Grillari, Ajuh et al. 2005). The complex is not assembled or indirectly maintained by a heterologous component as DNA or RNA. Moreover, an extended interaction network among the complex proteins has been detected in *S. cerevisiae* (McDonald, Ohi et al. 1999; Ajuh, Sleeman et al. 2001; Chen, Yu et al. 2002; Ohi and Gould 2002). In some studies, snRNAs co-precipitated with the NTC in *S. cerevisiae*. These cases most likely represent post-spliceosomal complexes (Ohi, Link et al. 2002; Wang, Hobbs et al. 2003; Tardiff and Rosbash 2006; Ohi, Ren et al. 2007). The *S. pombe* complex, conversely, can only be purified in the presence of snRNAs, which indicates a different composition or assembly behavior of the *S. pombe* NTC complex (McDonald, Ohi et al. 1999; Ohi, Link et al. 2002; Ohi, Ren et al. 2007).

Some members of the human Prp19-complex were detected in proteomic analyses of early spliceosomal complexes as the A- and the B-complex (Hartmuth, Urlaub et al. 2002; Deckert, Hartmuth et al. 2006; Behzadnia, Golas et al. 2007). However, this association seems to be loose as it persists only under very mild purification conditions. The stable integration of the human Prp19-complex into the spliceosome occurs during the final spliceosome activation step after U1-release. This was shown by proteome comparisons of the BAU1 spliceosome, which has already released the U1 snRNP, with the B\*-complex spliceosome, which represents the activated spliceosome and has released the U4 snRNP as well (Figure 6; (Makarova, Makarova et al. 2002; Makarova, Makarov et al. 2004). Purification of both spliceosomal complexes was performed in the presence of heparin to promote the dissociation of weakly bound proteins. In the BAU1 spliceosome, the Prp19-proteins were largely absent, whereas in the activated 45S B\* spliceosome all Prp19-complex proteins were present as an integral part of the 35S U5 snRNP.



**Figure 6: Proteome Comparison of Spliceosomal Complexes**

The proteomes of the spliceosomal B $\Delta$ U1, the 45S B\* and the 35S U5 complex are shown. In B $\Delta$ U1, which occurs before spliceosome activation, U4/U6 snRNP-specific proteins (blue box) are found, whereas proteins from the Prp19-complex (red box) are largely absent (only a few detected peptides of some proteins in mass spectrometry). In the active B\* complex, all Prp19-complex members and Prp19-related proteins have been detected. These proteins are also present in the 35S U5 snRNP. AP means associated proteins.

Since most likely the Prp19-complex is integrated into the spliceosome only shortly before it exerts its main function, the time-point of stable integration suggests that the complex plays a pivotal role in spliceosome activation or in splicing catalysis. In accordance to this, the first step of splicing is inhibited in extracts, which are doubly depleted with anti-CDC5- and anti-AD002 antibodies (Makarova, Makarov et al. 2004). The Prp19-complex remains tightly bound to the spliceosome after the first catalytic splicing step, suggesting that it is crucial for maintaining the catalytically active RNA-network (Bessonov, Anokhina et al. 2008).

In *S. cerevisiae* the NTC becomes stably associated with the spliceosome concomitant with or shortly after U4 snRNP is released (Tarn, Lee et al. 1993). Immunodepletion experiments have demonstrated that it is not needed for release of U4 snRNP though (Chan and Cheng 2005). In budding yeast, no B $\Delta$ U1-spliceosome has so far been isolated. Only equivalents of the human B-complex, which was stalled by lowering the ATP-concentration in the splicing reaction, and of the activated B\*-complex have been isolated. Therefore an exact comparison of the steps during spliceosome activation in the human and the yeast system is not possible, because this process is less well characterized in *S. cerevisiae*. But, as the findings are so far consistent with the human system, the time-point of stable integration is likely the same in both organisms.

The reason why the Prp19-complex is loosely attached to the spliceosome in early assembly stages and the question whether this loose attachment has a functional meaning remains unclear. One line of argumentation provides evidence that the presence and loose attachment of the human Prp19-complex is needed for spliceosome assembly: disruption of the complex leads to defects in spliceosome assembly (Ajuh and Lamond 2003; Grillari, Ajuh et al. 2005). The disruption can be achieved by an inhibition of the hPrp19 tetramerization (Grillari, Ajuh et al. 2005) or by an interruption of the interaction between CDC5 and PRL1 (Ajuh and Lamond 2003). Moreover, the presence of a non-functional complex resulted in correctly assembled spliceosomes, which were unable to perform catalysis (Ajuh and Lamond 2003). These findings provide evidence that the presence and loose attachment of the Prp19-complex is needed for spliceosome assembly, whereas its functionality is needed to promote spliceosome activation. As the complex is not stably integrated into the spliceosome at this stage, it has been speculated that it may exert a function (like ubiquitin ligation) on other proteins, which in turn are required for B complex formation.

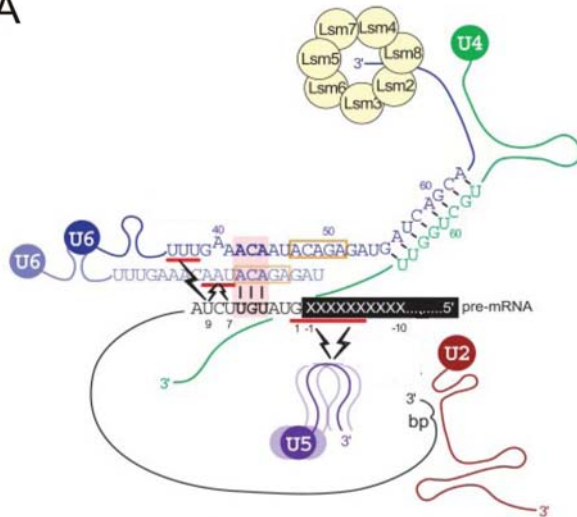
However, another line of evidence argues against this: correct assembly of human spliceosomal B complex was observed in immunodepleted extracts, whereas splicing catalysis is blocked in these extracts due to a defect in spliceosome activation (Ajuh, Sleeman et al. 2001; Ajuh and Lamond 2003; Makarova, Makarov et al. 2004). So these findings suggest that the Prp19-complex is not needed for spliceosome assembly, but only

for spliceosome activation. The detected B complexes were not formed with the help of residual undepleted Prp19-complex activity, as the amounts of B complex in the depleted extracts were almost equal to undepleted extracts.

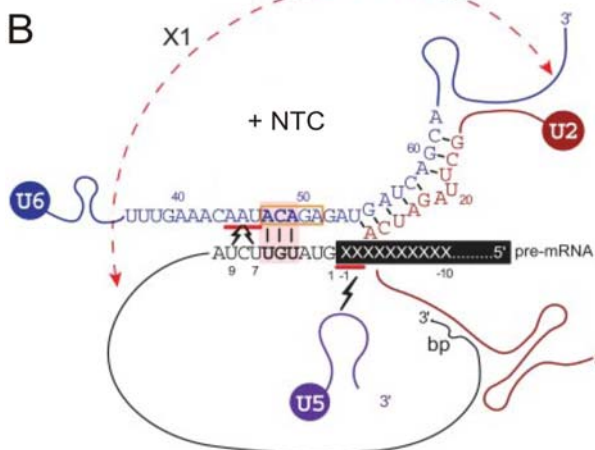
In the end it remains unclear, which of the two argumentations is correct. However, this uncertainty only concerns a putative functional role for the loose attachment of the Prp19-complex in spliceosome assembly. Both argumentations state that the presence of a functional complex is needed in the spliceosome activation step. The stably integrated complex has been proven to be essential for splicing catalysis and not for spliceosome assembly, because it is integrated into the spliceosome after tri-snRNP integration and during activation (Ajuh, Kuster et al. 2000; Ajuh, Sleeman et al. 2001; Chan, Kao et al. 2003; Makarova, Makarov et al. 2004; Chan and Cheng 2005).

The most interesting results concerning the function of the stably integrated Prp19-complex during spliceosome remodeling were obtained in *S. cerevisiae*. These results are based on the investigation of base pairing patterns between U5 and U6 snRNAs and the pre-mRNA (Figure 7). Even though the U1 snRNP is still bound to the 5'-splice site in spliceosomal A and B complexes, several dynamic contacts between the U6 snRNA and a broad region on the pre-mRNA in the vicinity of the 5'-splice site already exist in these complexes (Figure 7 A). Upon U1 (and U4) dissociation, the 5'-splice site is liberated and can undergo new interactions with the U6 snRNA. Compared to the interactions in spliceosomal A and B complexes, these new U6 snRNA/pre-mRNA-interactions are stronger and limited to a single contact (Figure 7 B). This contact involves the invariable sequence (ACAGAG) of the 5'-end of U6 snRNA and the UGU triplet of the pre-mRNA 5'-splice site (Chan, Kao et al. 2003; Chan and Cheng 2005). Thus, U6 undergoes a switch in base pairing from several loose interactions to one predominant stronger interaction with the pre-mRNA 5'-splice site during spliceosome activation. When the NTC-complex is depleted, the interaction pattern after spliceosome activation is different to the one in undepleted extract (Chan, Kao et al. 2003; Chan and Cheng 2005): not the invariable sequence, but a region of U6 snRNA, which lays more upstream, now contacts the pre-mRNA (Chan and Cheng 2005). A similar observation is made for U5 snRNA: normally, the conserved loop 1 of U5 snRNA strongly interacts with the last nucleotides of the 5'-exon prior to the catalytic steps. In extracts, which were depleted of the NTC, multiple and more dynamic interactions to a broader region in the 5'-exon are observed. In conclusion, these results suggest that the NTC functions as a specificity factor that defines base-pairing interactions of U5 and U6 snRNAs with the pre-mRNA in the active spliceosome.

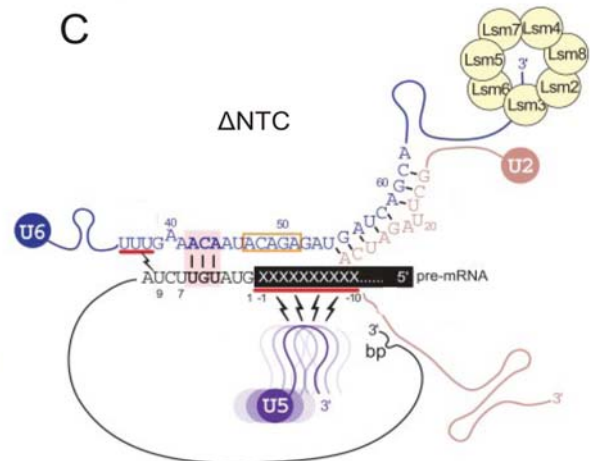
A



B



C



**Figure 7: Dynamic Interactions of U5 and U6 with Pre-mRNA and the Role of NTC**

**A** Interactions in the ATP-stalled spliceosomal complex, which still contains U1 and U4 snRNPs. There are multiple interactions between the U6 snRNA and the intron close to the 5'-splice site. Additionally, the U5 snRNA crosslinks to multiple sites on the exon-intron boundary. **B** Interactions after U1 and U4 snRNP release in the presence of NTC. Only one stronger contact between the U6 snRNA invariable sequence and the intron's UGU triplet is present. U5 snRNA is also detected only with one crosslink to the exon. **C** Interactions after U1 and U4 snRNP release in the absence of NTC. U6 snRNA contacts the UGU triplet with nucleotides upstream of the invariable sequence. U5 snRNA crosslinks to various sites in the exon and the LSm-proteins are still attached to the 3' U-tract of U6 snRNA. The 5'-exon is shown as a black box. X is a non-conserved nucleotide. The intron is a black line with sequences near the 5'-ss and the branch point (bp) shown. Multiple drawings of U5 and U6 indicate dynamic interactions with pre-mRNA. The orange box marks a part of the invariable sequence of U6; the red block marks base pairings between U6 and the 5'-ss; the red bar marks cross-linked residues identified by primer extension analysis; zigzags indicate cross-linking by UV. The dashed line indicates a base pairing. The figure is adapted from (Chan and Cheng 2005).

Moreover, both U5 and U6 snRNPs dissociate more easily from the spliceosome in the absence of NTC. Therefore the nineteen complex not only defines the base pairing interactions between these snRNPs and the pre-mRNA, but also helps to stabilize their interaction with the spliceosome (Chan, Kao et al. 2003).

Another important RNA-RNA contact during spliceosome activation is established between intron-nucleotides, which lay ~30 bases downstream from the 5'-splice site, and the 3'-terminal U-tract of U6 snRNA (Figure 7 B). In order to allow this contact, the LSM-proteins, which bind to the U6 snRNA 3'-U-tract, may have to be destabilized and eventually even be released from the U6 snRNA (Chan, Kao et al. 2003). In NTC-depleted extracts this destabilization is less efficient (Figure 7 B compared to C). Additionally the U6 snRNP and the U5 snRNP dissociate from the spliceosome more easily in depleted extracts. If U6 snRNP is released and the LSM-proteins are still attached to its 3-terminal U-tract may, U6 snRNP may enter a futile pathway: it does not form the U4/U6 di-snRNP anymore and thereby U6 snRNP re-usage in new rounds of splicing is inhibited (Chen, Kao et al. 2006). In conclusion, next to direct participation in spliceosome activation, the Prp19-complex has an indirect role in spliceosome recycling through affecting the biogenesis of U4/U6 di-snRNP via destabilization of the LSM-proteins. However the LSM-destabilization has not been observed in the human system so far.

Additionally it is speculated that the NTC may modulate splicing accuracy. Strains in which *Isy1p*, a nineteen complex protein, is deleted display reduced fidelity in 3'-splice site selection (Villa and Guthrie 2005).

In summary, most information on the function of the Prp19-complex has been obtained in *S. cerevisiae*, but the situation is most likely similar in the human system (some of the U6 snRNA-pre-mRNA crosslinks have been found in humans as well (Makarova, Makarov et al. 2004)). Whether the loose attachment of the Prp19-complex to early spliceosomal complexes has a functional role for spliceosome assembly is still unclear. After the assembly, however, the Prp19-complex is stably integrated into the spliceosome and presumably plays a role in spliceosome remodeling and spliceosome activation. The complex may serve thereby as a specificity factor, which helps to define the correct base pairing interactions between U5 and U6 snRNA and the pre-mRNA. Additionally, the Prp19-complex may be involved in the U6 snRNP-LSM destabilization, which in turn has an influence on the U4/U6 di-snRNP biogenesis. The complex may also be involved in the modulation of splicing accuracy. Whether different components of the complex mediate different functions or whether the entire complex works as a whole in all these processes is not yet known.

Finally, an insight on the potential location of the Prp19-complex in the spliceosome has been gained. A 29 Å resolution structure of the 37S fission yeast post-spliceosomal complex was

obtained by electron microscopy (Ohi, Ren et al. 2007). Different regions of the structure were assigned to the U5 snRNP and the U2/U6 snRNPs. The NTC is thought to correspond to a ridge domain, which protrudes between the upper U5-lobe and the lower U2/U6-lobe of the particle. In another study, which delineated the structure of the human C-complex (Jurica, Sousa et al. 2004), the Prp19-complex was believed to be located at a comparable position. So in the human C-complex as well as in the *S. pombe* post-spliceosomal complex the Prp19-complex is ideally positioned to exert its function on the U5 as well as on the U6 snRNP.

### The Prp19 Protein

The precursor RNA processing protein 19 (Prp19) is an essential splicing factor and is highly conserved throughout evolution. It has first been discovered in a series of screens for temperature sensitive mutants of yeast proteins, which are defective in pre-mRNA splicing (Ruby and Abelson 1991). Prp19 forms homo-tetramers, which may function as an assembly platform for other splicing factors and thereby it constitutes the central stabilizing component of the Prp19-complex.

Three protein motifs can be recognized in the sequence of human and *S. cerevisiae* Prp19 protein: an amino-terminal U-box domain, a central coiled-coil region and a C-terminal WD40 domain, which consists of seven WD-repeats (Figure 5). A crystal structure of the dimeric U-box domain from *S. cerevisiae* Prp19p has been determined at 1.5 Å resolution (Vander Kooi, Ohi et al. 2006). The mixed  $\alpha/\beta$  structure of the U-box domain (Figure 8 C) resembles a RING-finger domain (Ohi, Vander Kooi et al. 2003; Vander Kooi, Ohi et al. 2006). Both U-box and RING-finger domains often are integral components of macromolecular assemblies. In the human and yeast protein a heptad-repeat containing coiled-coil is predicted C-terminally to the U-box domain. Tetramerization via the coiled-coil domain is required for the *in vivo* function of *S. cerevisiae* Prp19p: a mutated protein, which does not form a tetramer, is not able to rescue a temperature sensitive mutant (Ohi, Vander Kooi et al. 2005). In Prp19p a construct of amino acids 66 to 141 corresponding to the predicted coiled-coil domain has been shown to tetramerize (Ohi, Vander Kooi et al. 2005). Tetramerization of human Prp19 is also supposed to occur, because it is present in super-stoichiometric amounts in the Prp19-complex (Ajuh, Kuster et al. 2000; Makarova, Makarov et al. 2004). Additionally the complex can be disrupted by peptides against the potential tetramerization interface in the hPrp19 coiled-coil domain (Grillari, Ajuh et al. 2005). However, tetramerization has never directly been shown for hPrp19. The WD40-repeats (also known as WD or beta-transducin repeats) are short ~40 amino acid motifs terminating in a Trp-Asp (WD) di-peptide (Smith, Gaitatzes et

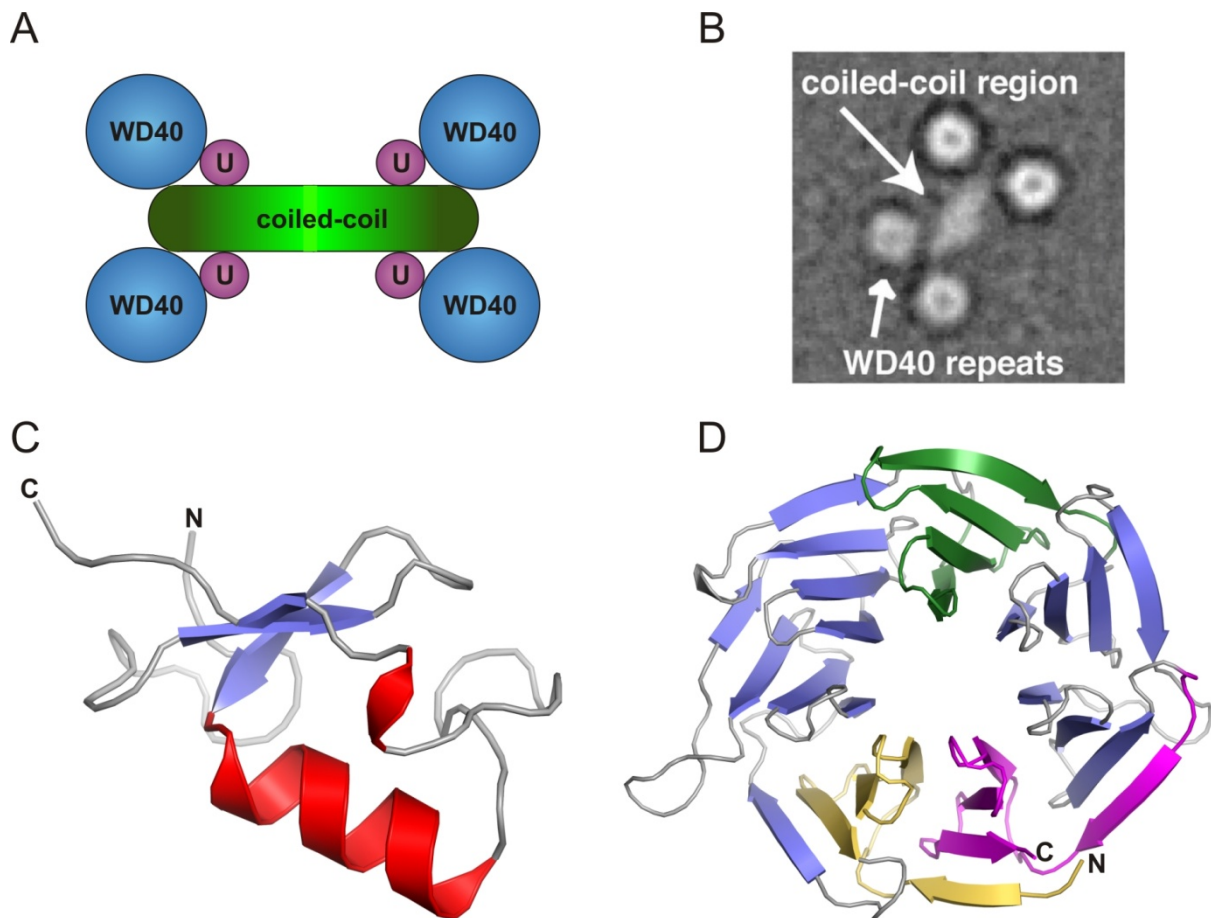
al. 1999). WD-containing proteins have 4 to 16 repeating units, which are thought to form a circularized beta-propeller structure (Figure 8 D). WD-repeat proteins form a large, ancient protein family found in all eukaryotes. The common function of WD-repeat proteins is the coordination of multi-protein assemblies, in which they serve as a rigid scaffold for protein interactions. The binding-specificity of WD40-proteins is determined by sequences outside the repeats.

Based on structural insights into the U-box and WD40-domains and biochemical data, a model of *S. cerevisiae* Prp19p has been proposed: a dimer of U-box domains and two WD40-domains are connected to either side of a central tetrameric coiled-coil domain (Figure 8 A). The WD40-domains are flexibly attached to the coiled-coil. This model was confirmed by negative stain electron microscopy images, which showed that a Prp19p tetramer forms an elongated particle consisting of a central stalk formed by the tetrameric coiled-coil and two WD40- and U-box domains on either side of it (Ohi, Vander Kooi et al. 2005). The WD40-domains are visible as circles in the negative stain image.

Sequence properties of the human and *S. cerevisiae* Prp19 protein, which may influence protein production and crystallization, are listed in Table 20 and explained in chapter 8.5. Both proteins show one extended disordered region and a number of relatively short stretches of local disorder. The human and the *S. cerevisiae* Prp19 share 24 % identical residues in their first 400 amino acids. The similarity in this region is around 43 %.

Several different functions have been assigned to the Prp19 protein: apart from its role in pre-mRNA splicing, human and yeast Prp19 have been implicated in the repair of damaged DNA (Mahajan and Mitchell 2003; Zhang, Kaur et al. 2005). They may also have an E3 ubiquitin-ligase activity (Ohi, Vander Kooi et al. 2003; Vander Kooi, Ohi et al. 2006), which is mediated by the U-box domain. U-box E3 ligases facilitate the transfer of ubiquitin by precise spatial orientation of the E2 and the substrate. This Prp19 function might for example be needed during spliceosome assembly to modulate other proteins by ubiquitination. In addition, hPrp19 is ubiquitinated itself. It has been found that a ubiquitinated hPrp19 species is active in DNA repair (Lu and Legerski 2007). Interestingly, ubiquitinated hPrp19 does not associate with CDC5 and PRL1. This implies that splicing and DNA repair are likely carried out by two different hPrp19 forms and that the function is regulated by ubiquitination. The human Prp19 protein has also been characterized as a nuclear matrix protein (Gotzmann, Gerner et al. 2000). Numerous studies have demonstrated an involvement of the nuclear matrix in the processes of replication, transcription, splicing and DNA repair (Cremer, Kreth et al. 2000). Therefore it is hypothesized that the two species of hPrp19 both execute their jobs in “processing or repair factories”, which are located at the nuclear matrix.





**Figure 8: Model of *S. cerevisiae* Prp19p**

The model of Prp19p, which is based on structural insights in the U-box and WD40-domains and biochemical data, is shown. **A** Schematic representation of Prp19p domain localization. **B** Negative stain image as published in Ohi *et. al.* (Ohi, Vander Kooi *et al.* 2005). **C** Prp19p Ubox-monomer from the dimeric U-box crystal structure as published in van der Kooi *et. al.* (Vander Kooi, Ohi *et al.* 2006). **D** WD40  $\beta$ -propeller structure from the G $\beta$ -subunit of the heterotrimeric G-protein as published in (Smith, Gaitatzes *et al.* 1999). The N-terminal WD40-motif is colored in yellow, the C-terminal motif is colored in magenta. One  $\beta$ -propeller was colored in green.

It has been demonstrated that tetramerization of Prp19 is a pre-requisite for spliceosome assembly and may also be needed for spliceosome stability. Disruption of the Prp19 tetramer by mutations or by peptides directed against the coiled-coil homo-oligomerization domain inhibits splicing in human and yeast (Ohi and Gould 2002; Grillari, Ajuh *et al.* 2005; Ohi, Vander Kooi *et al.* 2005). This may be due to the important contacts between Prp19(p) and other components of the Prp19-complex via the coiled-coil domain (discussed on page 29). All these results indicate that Prp19 constitutes a central interaction platform in human and in yeast, which is needed for the stability of the entire Prp19-complex.

## The Cdc5 Protein

The cell division cycle 5 (CDC5) protein is the second essential protein in the Prp19-complex. Its high degree of conservation suggests an important function. CDC5 was originally discovered in a screen for fission yeast mutants defective in cell cycle progression (Nurse, Thuriaux et al. 1976). Despite the early discovery, CDC5 was only related to splicing after its detection in a spliceosomal complex (Neubauer, King et al. 1998) and a clear splicing function has not yet been allocated to this protein. However, the most obvious function in splicing is to connect hPrp19 and PRL1 in the Prp19-complex. As PRL1 in turn makes contacts to integral spliceosomal components, the CDC5/PRL1-interaction may tether the Prp19-complex to the spliceosome.

In the amino-terminus of CDC5 from diverse species two Myb-repeats and a Myb-like repeat can be found (Figure 5). Myb-repeats are ~50 amino acid helix-turn-helix motifs, which were first characterized in the transcriptional activator c-Myb and are known to specifically bind DNA. Characteristically spaced tryptophan residues build the DNA-binding surface. Even though the CDC5 Myb-domains are important for cell growth (Tsai, Chow et al. 1999), single and even double mutations, which were predicted to disrupt the Myb-structure, did not impair the function of the yeast ortholog (Cef1p) *in vivo* (Ohi, Feoktistova et al. 1998). On the contrary, a 31 amino acid stretch at the C-terminus of Cef1p was essential for cell viability and protein-protein interactions (Tsai, Chow et al. 1999). This region encompasses the so called EVES-sequence (amino acid 480-483), which has been shown to play a role in interactions of Myb to its binding partners. Indeed the essential CDC5-PRL1 contact in the human system is mediated via this region (Ajuh, Sleeman et al. 2001; Ajuh and Lamond 2003). Even though this EVES-sequence cannot be recognized in the human CDC5, other sequence motifs can be detected: four nuclear localization signals, several phosphorylation sites and a proline-rich region, which resembles the activation domain found in Myb-related proteins (Bernstein and Coughlin 1997).

Sequence properties of the human CDC5 and *S. cerevisiae* Cef1p protein, which may influence protein production and crystallization, are listed in Table 20. Notably the yeast protein is far smaller than the human protein (590 amino acids instead of 802), but shares almost the same domain structure (Figure 5). One difference is that the proline-rich region of CDC5 is replaced by a lysine-rich region in Cef1p. Both proteins are predicted to be entirely disordered apart from the Myb-domains. CDC5 proteins from human, *S. cerevisiae*, *S. pombe*, *Xenopus laevis*, *Mus musculus* and *Caenorhabditis elegans* all display a very high sequence similarity. The Myb-domains of human and *S. cerevisiae* CDC5 for example align with 46 % identity and 64 % similarity.

Various functions have been assigned to the CDC5 protein before its implication in pre-mRNA splicing, in particular transcriptional activation by DNA-binding (Ohi, McCollum et al. 1994; Hirayama and Shinozaki 1996; Lei, Shen et al. 2000) and the correlated stimulation of cell cycle progression at the G<sub>2</sub>-phase (Ohi, Feoktistova et al. 1998). A block of cell cycle progression is often observed for spliceosomal protein mutants, also for CDC5 (Ohi, Feoktistova et al. 1998). Likely this block originates from a lack of key mRNAs that are needed for cell cycle progression, but cannot be spliced. The cell cycle arrest is therefore likely only a consequence of a splicing defect.

Immunodepletion of human, *S. pombe* and *S. cerevisiae* CDC5 inhibits splicing (Burns, Ohi et al. 1999; McDonald, Ohi et al. 1999; Ajuh, Kuster et al. 2000), demonstrating that the protein is essential for splicing in several organisms. Moreover, C-terminal truncation mutants of human CDC5 were unable to interact with PRL1 and disrupted the Prp19-complex. Addition of these truncation mutants to the nuclear extract of HeLa cells (immortal human cell line named after Henrietta Lacks) not only inhibited splicing catalysis, but also interfered with spliceosome assembly (Ajuh, Sleeman et al. 2001). For this reason it was speculated that CDC5 either contacts components of the spliceosome itself or at least it connects the core Prp19-complex to proteins, which contact the spliceosome. Thereby CDC5 may tether the Prp19-complex to the spliceosome. The presence of the Prp19-complex in turn may be needed for efficient spliceosome assembly.

In more detail this function of CDC5 could look the following way: CDC5 binds PRL1 via its C-terminus. Additionally, a contact to hPrp19 is established. PRL1 then contacts integral spliceosomal components with its WD40-domains and thereby it tethers the Prp19p-complex to the spliceosome. Alternatively, a direct contact between *S. cerevisiae* Cef1p and Clf1p (human Syf3) was detected. As Syf3 may already be present in the BΔU1 spliceosome, this contact could also tether the Prp19-complex to the spliceosome via CDC5. In any case, the function of CDC5 or Cef1p seems to be the recruitment of the Prp19-complex to the spliceosome.

Because of its ability to bind DNA, another function of CDC5 was proposed, even though it is highly speculative: CDC5 may tether the spliceosome associated Prp19-complex to the DNA during transcription and therefore helps to couple transcription and splicing (Ajuh, Kuster et al. 2000).

The transcriptional activator function of CDC5 and its function in splicing are potentially regulated by phosphorylation and dephosphorylation (Bernstein and Coughlin 1997; Stukenberg, Lustig et al. 1997; Ben-Yehuda, Dix et al. 2000; Boudrez, Beullens et al. 2000).

### Other Proteins from the Prp19-complex

Compared to Prp19 and CDC5 very little is known about the remaining core components of the complex and their implication in splicing. The domain architectures can be found in Figure 5. Sequence properties, which may influence protein production and crystallization, are listed in Table 20.

Although the human pleiotropic regulator 1 (PRL1) protein as well as its *S. cerevisiae* homolog Prp46p are essential for splicing (Ajuh, Sleeman et al. 2001; Albers, Diment et al. 2003), not much is known about their precise function. The only domain predicted in this highly conserved protein is the WD40-domain with seven WD-repeats (for a more detailed description of WD40-domains see page 23). Apart from the C-terminal WD40 domains the human and yeast proteins are predicted to be unfolded.

PRL1 has been described to play a role in the pleiotropic control of glucose and hormone responses in *A. thaliana* (Nemeth, Salchert et al. 1998). Additionally, mutation of its gene in *S. pombe* results in cell cycle arrest and defects in pre-mRNA splicing (Potashkin, Kim et al. 1998; McDonald, Ohi et al. 1999). Multiple interactions to spliceosomal proteins as well as to kinases and glucose signaling regulator proteins have been detected. In the spliceosome interactions to the DEAD-box helicase Prp22p and to Prp45p, the human homolog of SKIP, were found (Albers, Diment et al. 2003). Because of the presence of a WD40-domain, it was speculated that PRL1, similar to Prp19, mediates protein-protein interactions (Albers, Diment et al. 2003). Unlike Prp19, PRL1 does not oligomerize and it does not seem to be a central platform for Prp19-complex members (Chen, Yu et al. 2002; Ohi and Gould 2002). More likely, PRL1 recruits the Prp19-complex to the spliceosome, as it simultaneously contacts spliceosomal proteins and Prp19-complex components (see page 29).

The human heat shock protein 73 (Hsp73/CCAP1 or Hsc70; *S. cerevisiae* SSA2) belongs to the Hsp70 protein family. In most of its amino acids Hsp73 equals the bovine Hsc70 protein, of which a structure was already published (Jiang, Prasad et al. 2005). These residues comprise a 44 kDa substrate-binding domain and an 18 kDa nucleotide-binding domain, which are both well conserved. The remaining residues are 94 % identical to a *Rattus norvegicus* 10 kDa sub-domain of Hsc70 (Chou, Forouhar et al. 2003).

The Hsp70-system was described to be more flexible than other chaperone systems, because of diverse interactions to auxiliary co-chaperone proteins (Young, Barral et al. 2003). In addition to ATP-dependent protein folding, constitutively expressed members have been shown to be involved in vesicle secretion, disassembly of clathrin cages and they may also participate in the post-translational trans-membrane targeting of proteins to cellular organelles. Why Hsp73 is present in the Prp19-complex is currently unknown. The trans-

membrane targeting function together with the notion of Prp19's possible membrane association could point towards a role in membrane-located "processing factories". The *S. cerevisiae* homolog Ssa2 has not yet been shown to be a component of the NTC.

*S. cerevisiae* Snt309p is a functional homolog of the non-essential human spliceosome associated protein 27 (SPF27) and was identified in a screen for proteins, which are synthetic-lethal with Prp19p mutations (Chen, Jan et al. 1998). There are no domains predicted for the entire sequence in human and yeast and both sequences are predicted to be largely unfolded. The Prp19-complex is destabilized in the absence of Snt309p. Addition of the binary Prp19p-Snt309p complex to extracts from a Snt309p-deletion strain could reconstitute the Prp19-complex (Chen, Tsao et al. 1999). Genetic analysis further suggested that the protein plays a role in modulating interactions of Prp19p with other components, which in turn facilitates the formation of the Prp19p-complex.

The least is known about the human adrenal-gland protein 2 (CCAP2/AD002; *S. cerevisiae* Cwc15p protein). Similar to Snt309p, the Cwc15p protein has been identified in a screen for mutants that confer synthetic lethality in the presence of a Prp19p mutation (Chen, Jan et al. 1998). There are no domains predicted for this non-essential protein in human or in yeast. CCAP2 and Cwc15p are identical in 21 % of their amino acids and share 37 % similar residues. Both proteins are predicted to be completely unfolded and nothing is known about their function.

### Interactions among the Complex Proteins

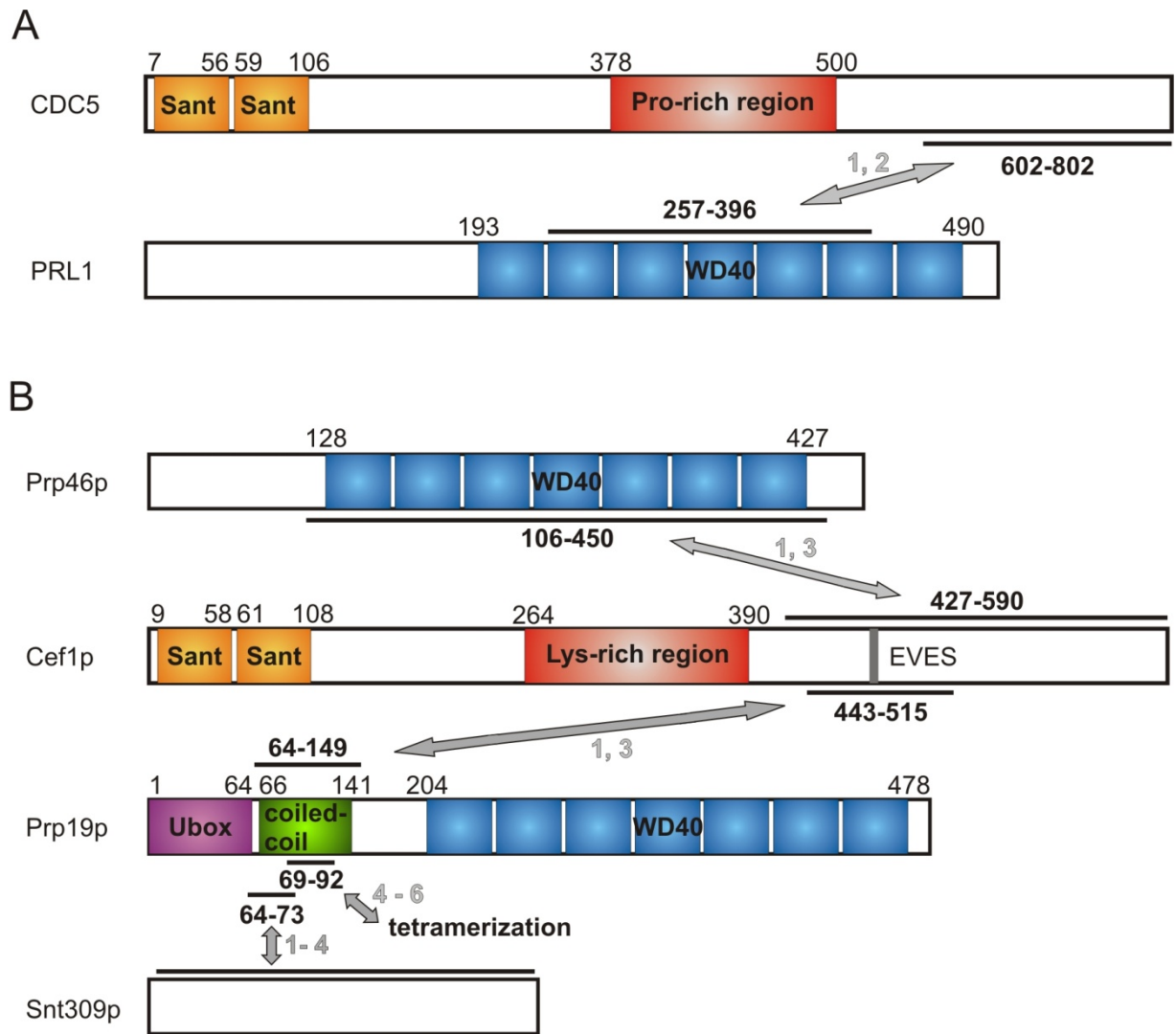
Depending on the purification method, up to thirty proteins are found to constitute the Prp19-complex. The human Prp19 complex is predicted to have fewer components than the yeast complex. Especially in the yeast complexes a number of interactions among the component proteins have been mapped with different methods. In the following paragraphs only the interactions between core components defined for this work (see Figure 5) will be discussed in detail and an interaction scheme will be established.

Yeast two-hybrid, far western, co-immuno-precipitation and *in vitro* binding assays (like pull-down assays, gel filtration and competition binding assays) have revealed the most complete picture of interactions in the *S. cerevisiae* NTC-complex (Tarn, Hsu et al. 1994; Chen, Tsao et al. 1999; Tsai, Chow et al. 1999; Chen, Yu et al. 2002; Ohi and Gould 2002). An overview is given in Figure 9. In a model for complex formation, Prp19p tetramerization is imagined as a first step, because the tetrameric coiled-coil is needed for the contacts to Snt309p and Cef1p. As Prp19p can interact with its associated proteins in a different fashion when Snt309p is bound to it, the interaction between Prp19p and Snt309p is assumed occur next (Chen, Jan

et al. 1998). This interaction between Prp19p and Snt309p is believed to enable stable complex formation. Concomitant to or shortly after Snt309p-binding, a monomer of Cef1p contacts the coiled-coil. Even though Cef1p and Snt309p contact almost the same region on Prp19p (residues 63-73 of the Prp19p coiled-coil), the interactions are not exclusive. The contacts are likely made from opposite sites on the tetrameric coiled-coil. Cef1p itself contacts the Prp46p protein (human PRL1) and as in the human proteins the interaction occurs between the carboxy-terminus of Cef1p and the WD40-domains in Prp46p. This interaction is essential for splicing and has been mapped to the region 427-590 in Cef1p (C-terminus including EVES-sequence) and to residues 106-450 in Prp46p, which comprise all WD-repeats O(Ohi and Gould 2002). Regions N-terminally of the mapped amino acids in Cef1p strengthen the interaction. This suggests that the binding interface between Cef1p and Prp46p is rather extended. In addition Cef1p may mediate numerous interactions to other Prp19p-related proteins. Among these are contacts to the Syf-family proteins (Syf1p/Ntc90p, Syf2p/Ntc31p and Syf3p/Ntc77p/Clf1p) and Isy1p (Ohi, Link et al. 2002).

In the human system only the interaction between the WD40-domains of PRL1 and the C-terminus of CDC5 occurs *in vitro* and *in vivo*. The contact has been shown to be essential for splicing in humans (Ajuh, Sleeman et al. 2001; Ajuh and Lamond 2003). Peptides corresponding to highly conserved sequences in the interaction domain of both proteins inhibit splicing *in vitro*. These peptides cannot disrupt already formed CDC5-PRL1 complexes, which suggests, that the protein-protein interaction interface may be extensive. The fact that the peptides once integrated into the Prp19-complex are not removable by the recombinantly produced interaction partner, demonstrates that CDC5 and PRL1 are deeply buried in the complex. Together with peptides, which mimic the Prp19p tetramerization interface, these small molecules can be used as effective splicing inhibitors (Ajuh and Lamond 2003; Grillari, Ajuh et al. 2005).

Summarized, the Prp19-complex in humans and in yeast seems to function as an integral, pre-formed unit. Among other evidences this notion is supported by a number of interactions, which have mainly been characterized in *S. cerevisiae*. All postulated interaction models of the yeast proteins describe a Prp19p-tetramer as the central platform, on which the other complex components assemble.



**Figure 9: Interactions among Prp19-Complex and NTC Components**

The interactions between the different proteins are shown as grey arrows. The grey numbers next to the arrow stand for the method, with which the interaction has been mapped: 1 = pull-down, 2 = co-immuno-precipitation, 3 = yeast two hybrid, 4 = far western, 5 = gel filtration, 6 = competition binding. The interacting region is shown as a black bar above or below the interacting region. Additionally the amino acids of the interacting regions are written in bold letters on these black bars.

**A** Interactions between proteins from the human complex. **B** Interactions between NTC-proteins.

## 2.4 The Minor Spliceosome

The ATAC-introns are spliced by the minor or U12-dependent spliceosome (Hall and Padgett 1996; Tarn and Steitz 1996; Tarn and Steitz 1996). Similar to the major spliceosome, five U snRNPs, namely U11, U12, U4atac, U5 and U6atac, are the main building blocks of the minor spliceosome. U11 and U12 snRNP are functional analogs of U1 and U2 snRNP. The U4atac/U6atac is analogous to U4/U6 (Hall and Padgett 1996; Tarn and Steitz 1996; Tarn and Steitz 1996; Kolossova and Padgett 1997; Yu and Steitz 1997). The U5 snRNP is common to both types of spliceosome.

The splicing reaction catalyzed by the minor spliceosome is far slower than the one of the major spliceosome, which makes it a rate limiting step in protein biosynthesis (Patel, McCarthy et al. 2002). The reaction also proceeds via the described two-step mechanism (see chapter 2.1.2; summarized in (Patel and Steitz 2003)).

Furthermore, evidence for a comparable assembly pathway of the two spliceosomes was found. U11 is responsible for 5'-splice site recognition, while U12 binds to the branch point. Interaction with the U4atac/U6atac:U5 tri-snRNP and subsequent rearrangements generate the functionally active complex.

### 2.4.1 Components of the Minor Spliceosome

Many components of the minor spliceosome seem to resemble the ones in the major spliceosome. The U snRNAs are functionally and structurally analogous. Even though the protein composition of the minor spliceosome is less well characterized, the common proteins seem to be the same. In addition to this, the tri-snRNP comprises a similar set of proteins.

Differences between the two types of spliceosomes, however, can be found regarding the U11 and U12 snRNP: unlike their major spliceosomal counterparts (U1 and U2 snRNP), these particles exist as a pre-formed U11/U12 di-snRNP. This U11/U12 di-snRNP is the predominant form of U11 and U12 and migrates at 18S in a glycerol-gradient (Montzka and Steitz 1988). A 12S U11 mono-particle exists next to the di-snRNP (Will, Schneider *et al.* 1999), whereas only traces of the U12 mono-particle are present. The U11/U12 di-snRNP also harbors a set of proteins, which are not known from the major spliceosome. Vice versa homologs of all U1-specific proteins as well as U2-A', U2-B'' and subunits of SF3a are absent in the U11/U12 di-snRNP.



### The U snRNAs

The minor spliceosomal U snRNAs are far less abundant than their major counterparts (only  $2 \times 10^3$ - $1 \times 10^4$  compared to  $1 \times 10^6$  copies per cell). They share many structural features with the major U snRNAs. All apart from U6atac possess a Sm-binding site and a tri-methyl cap and all apart from U6atac are RNA-polymerase II transcripts. U6atac is a transcript of RNA-polymerase III, has a simple  $\gamma$ -monomethyl-cap and an LSm-binding site (Tarn and Steitz 1996; Tarn and Steitz 1996).

Remarkable similarities can also be found regarding the secondary structure of the U snRNAs. This is particularly interesting, because their primary sequences differ significantly. U11 resembles the U1 cloverleaf structure and U4atac/U6atac base pair to form the characteristic Y-shaped structure known from U4/U6.

### The Common Proteins

The seven Sm-proteins can be found in each U snRNP apart from U6atac, where the LSm-proteins 2-8 are found (Montzka and Steitz 1988; Tarn and Steitz 1996; Schneider, Will et al. 2002).

### The Particle-Specific Proteins

The protein composition of isolated U11/U12 di-snRNPs and U11 mono-particles has been studied with the help of mass spectrometry (see Table 2; (Will, Schneider et al. 1999)). Around 20 proteins, which were already known from the U2-dependent spliceosome, have been found in the di-snRNP, among them the proteins from the SF3b-complex. This finding suggests a conserved mechanism for branch point recognition by the SF3b subunits. The SF3a complex is not associated with the U11/U12 di-snRNP. Notably, all U1-specific proteins, which are essential for the stabilization of the U1 snRNP/5'-splice site interaction, were absent in the minor spliceosomal particles.

In addition to these findings, seven novel proteins were discovered in the di-snRNP, which could not be detected in the major spliceosome (Will, Schneider *et al.* 2004). These proteins, namely 65K, 59K, 48K, 35K, 31K, 25K and 20K, are therefore called U11/U12-specific proteins. Their high conservation among vertebrates, insects and plants underscores their importance for the U12-dependent splicing.

name	kDa	U snRNP proteins			
		12S U1	17S U2	12S U11	18S U11/U12
B/B' LSM8	18/29	●	●	●	●●
D3 LSM4	18	●	●	●	●●
D2 LSM3	16.5	●	●	●	●●
D1 LSM2	16	●	●	●	●●
E LSM5	12	●	●	●	●●
F LSM6	11	●	●	●	●●
G LSM7	9	●	●	●	●●
70K	70	●			
A	34	●			
C	22	●			
A'	31		●		
B''	28.5		●		
SF3a	120		●		
	66		●		
	60		●		
SF3b	155		●		●
	145		●		●
	135		●		●
	49		●		●
	p14		●		●
	14		●		●
	10		●		●
	65			●	●
	59			●	●
	49			●	●
	35			●	●
	31			●	●
	25			●	●
	20				●

**Table 2: Proteome Comparison of Major and Minor Spliceosomal Particles**

The proteomes of 12S U1 and 17S U2 snRNP are compared to the minor 12S U11-monomer and the 18S U11/U12 di-snRNP.

Interestingly, some U11/U12-specific proteins exhibit similarities to U1 proteins (Benecke 2004). The domain structure of the U1C protein, for example, is mirrored in the U11/U12 20K protein. A similar situation is found for U11/U12 35K, which resembles U1 70K in sequence and structure. But in contrast to their U1 counterparts, both proteins did not bind RNA. The U11/U12 65K protein, however, whose domain architecture resembles that of the U1A protein, binds to the U12 snRNA as well as to the U11-associated 59K protein (Benecke, Luhrmann *et al.* 2005). As the 65K protein was one major focus of the present thesis, further details about it will be given in chapter 2.4.3.

In summary, U11/U12-specific proteins have been detected in the minor spliceosome, whereas the U1-specific proteins are absent. This implies that many of the protein-protein- and protein-RNA-interactions in the early spliceosomal complexes are different between the two types of spliceosomes.

### The Non-snRNP Proteins

The only non-snRNP proteins, which could be detected in the minor spliceosome until now, are the SR-proteins. They seem to function in both types of spliceosomes in a similar manner (Hastings and Krainer 2001). The presence of other factors like for example the Prp19-complex is currently not known.

#### 2.4.2 Assembly of the Minor Spliceosome

The assembly of the minor spliceosome essentially follows similar lines as the one of the major spliceosome. However, exceptions are found in the early stages, because of the existence of the U11/U12 di-snRNP. The first detectable minor complex is the A complex, which is formed by simultaneous binding of the U11/U12 di-snRNP to the 5'-splice site and the branch point (Frilander and Steitz 1999). Like in the major spliceosome, the U snRNAs undergo base pairing with the pre-mRNA (Hall and Padgett 1996; Tarn and Steitz 1996; Tarn and Steitz 1996; Kolossova and Padgett 1997; Yu and Steitz 1997). The intron-bridge, which brings the two ends of the intron into closer proximity, is already pre-formed in the di-snRNP. The incorporation of the U4atac/U6atac/U5 tri-snRNP generates the B complex, which is subsequently rearranged to the catalytically active B\* spliceosome (Frilander and Steitz 2001). After the first transesterification reaction and formation of the C-complex, the second transesterification generates the mature mRNA, which is transported to the cytoplasm for protein biosynthesis. The post-spliceosomal complex is recycled and the intron lariat is degraded.

#### 2.4.3 The U11/U12 65K Protein

The U11/U12-specific 65K protein has only recently been discovered (Will, Schneider et al. 2004). It serves as a bridging factor between the minor spliceosomal U11 and U12 snRNPs (Benecke, Luhrmann et al. 2005). Via its N-terminal half the protein binds to the U11-associated 59K protein and with its C-terminal RRM it binds to stem-loop III of the U12 snRNA (Benecke, Luhrmann et al. 2005). In this way 65K establishes proximity of two pre-mRNA regions: of the 5'-splice site, which is bound by the U11 snRNP, and of the branch point region, which is bound by the U12 snRNP. This proximity is needed for the first transesterification reaction. In other words, the formation of a classical intron bridge is mediated by the U11/U12 65K protein (see chapter 2.3.2 and Figure 4).

U11/U12 65K contains an N- and a C-terminal RNA recognition motif. In the connecting linker, a proline-rich region and a sequence can be found, which is almost completely conserved among 65K orthologs (Figure 10). Apart from this, the region between aa 211 and 380 has been shown to contribute to the binding specificity of the carboxy-terminal RRM to the U12 snRNA (Benecke, Luhrmann et al. 2005). The C-terminal RRM 65K is highly homologous to the N-terminal RNA recognition motifs of the major spliceosomal proteins U1A and U2B<sup>''</sup>. Crystal structures of these two proteins have been solved: U2B<sup>''</sup> in complex with its ancillary protein U2A<sup>'</sup> and 24 nucleotides from the hairpin IV of U2 snRNA (PDB-entry: 1A9N) and U1A in its RNA-free form (PDB-entries: 1OIA, 1NU4) and bound to a 21 nucleotide long RNA, which resembles hairpin II of U1 snRNA (PDB-entry: 1URN). The carboxy-terminal RRM of U11/U12 65K is 28 % identical and 57 % similar to the sequence in the structure of U2B<sup>''</sup> (chain D). To the sequence in the U1A structure (PDB-entry: 1URN, chain A) it shows even 32 % sequence identity and 55 % similarity. This high homology suggests that the three RRM s have evolved from a common ancestor by gene-duplication and subsequent diversification (Bandziulis, Swanson et al. 1989; Benecke, Luhrmann et al. 2005).



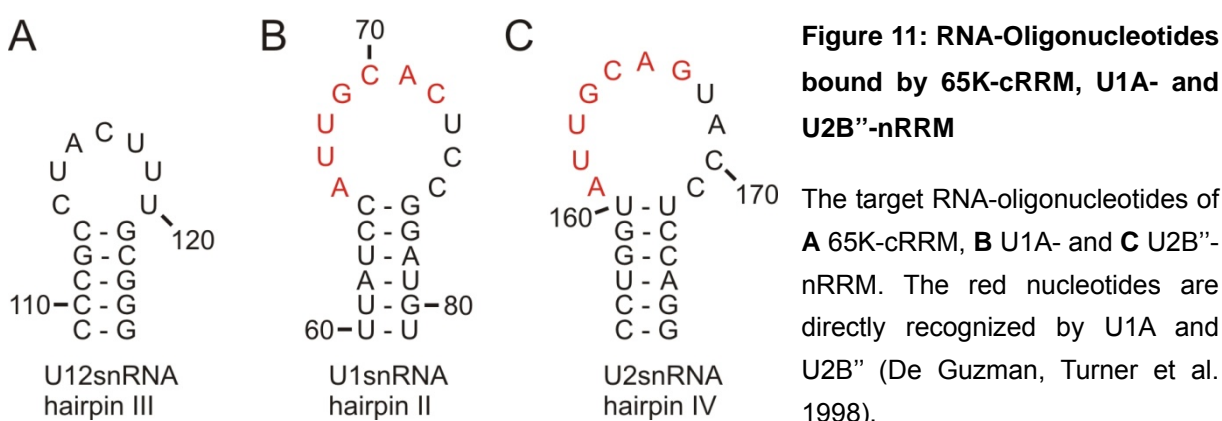
**Figure 10: Domain Structure of U11/U12 65K**

The domains are shown as described in Benecke *et al.* (Benecke, Luhrmann et al. 2005). The grey box represents the conserved sequence QVLHLMN(K/R)MNL, which is characteristic for 65K orthologs.

As the N-terminal RRM s of U1A and U2B<sup>''</sup> (U1A-nRRM, U2B<sup>''</sup>-nRRM) are among the first and best characterized representatives of RNA recognition motifs (Nagai, Oubridge et al. 1990; Oubridge, Ito et al. 1994; Price, Evans et al. 1998), a good basis for functional and structural comparison is given. Structural aspects will be discussed in chapter 2.4.4. Functionally the following facts are known: U1A-nRRM binds with very high affinity and specificity to hairpin II of U1 snRNA ( $K_D \sim 32$  pM; (Katsamba, Myszka et al. 2001)) and to the 3' untranslated region (UTR) of its own mRNA. The latter interaction enables U1A to regulate the polyadenylation of its own pre-mRNA (Boelens, Jansen et al. 1993; Gunderson, Beyer et al. 1994; Allain, Howe et al. 1997). U2B<sup>''</sup>-nRRM, in contrast, binds to hairpin IV of U2 snRNA only in the presence of the ancillary protein U2A<sup>'</sup> (Scherly, Boelens et al. 1990; Bentley and

Keene 1991; Boelens, Scherly et al. 1991). Thus, although the two RRRMs are closely related, their RNA-binding behaviors are dramatically different.

The close homology between the 65K-cRRM and the U1A/U2B"-nRRMs is paralleled by the similarity of their RNA targets (Figure 11). Human 65K-cRRM binds to nucleotides 109-125 at the 3'-end of U12 snRNA. These nucleotides form a hairpin structure with a five base pair stem and a seven residue loop (Benecke, Luhrmann et al. 2005). U12 snRNA hairpin III therefore resembles U1 snRNA hairpin II and U2 snRNA hairpin IV.



#### 2.4.4 RNA Recognition Motifs and the Mechanism of RNA-Binding

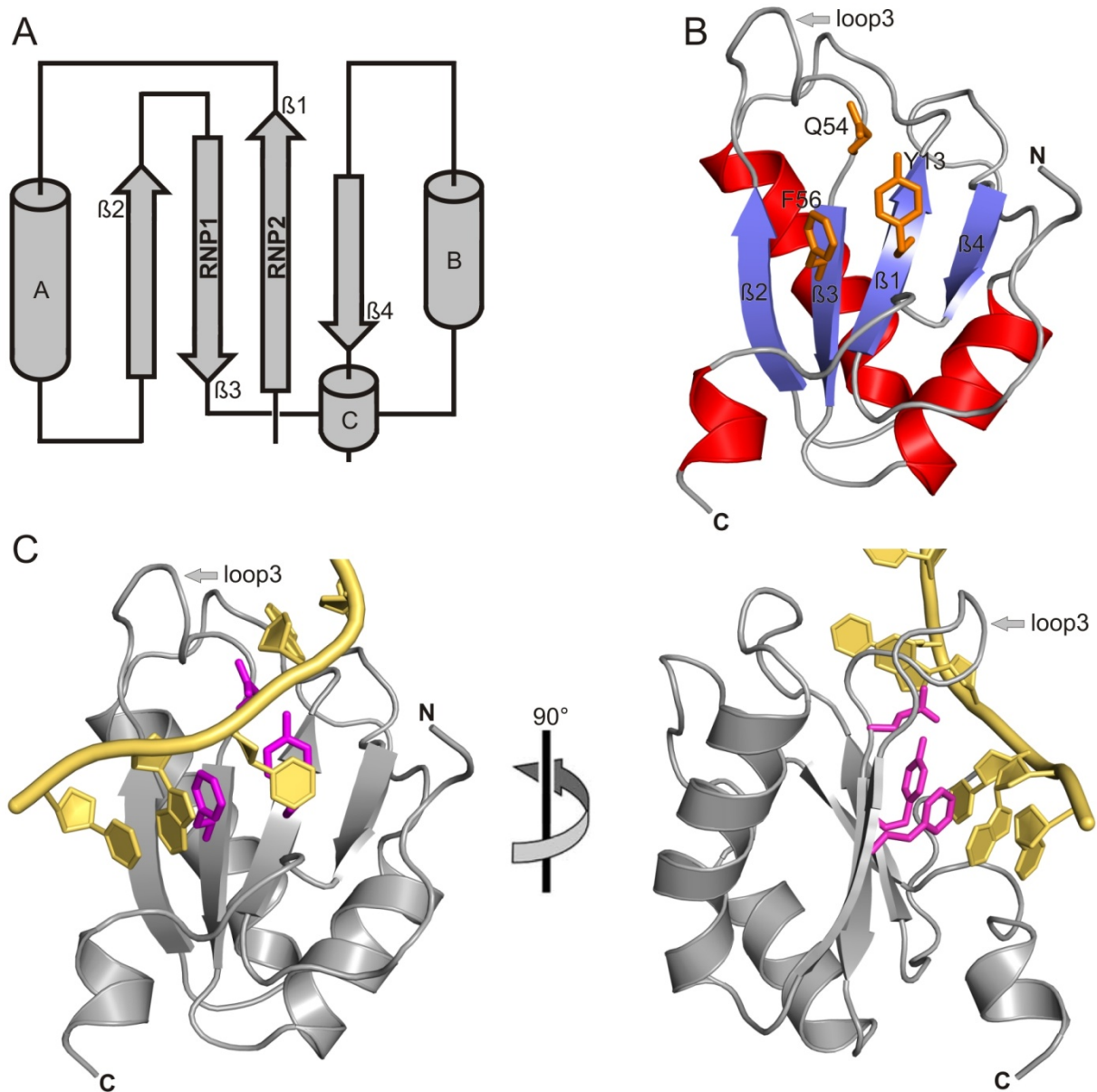
RNA recognition motifs are one of the most abundant domain classes in eukaryotes. They constitute a very versatile macromolecular interaction module, which can support RNA or protein binding by using diverse interaction surfaces (Maris, Dominguez et al. 2005). RRRMs consist of about 90 amino acids that fold into an  $\alpha/\beta$ -sandwich structure with a  $\beta 1-\alpha 1-\beta 2-\beta 3-\alpha 2-\beta 4$  topology (Figure 12 A). The four  $\beta$ -strands form an anti-parallel  $\beta$ -sheet in the spatial order  $\beta 4-\beta 1-\beta 3-\beta 2$  with the two helices packed against one face of the sheet (Figure 12 B). Occasionally a small two-stranded  $\beta$ -sheet can be found in loop 5 ( $\beta 3'$  and  $\beta 3''$ ). An additional C-terminal helix is found in some RRRMs, including the amino-terminal RRM of U1A (U1A-nRRM).

In canonical RNA recognition motifs, two highly conserved sequences, the RNP2 hexamer and RNP1 octamer (Bandziulis, Swanson *et al.* 1989), are located on the two central  $\beta$ -strands (Figure 12 A and Figure 24). Most of the conserved residues in RRRMs build the hydrophobic cores of the domains (Nagai, Oubridge et al. 1990). However, the conserved amino acids on RNP1 positions 1, 3 (and 5) and RNP2 position 2 are found on the surface of the  $\beta$ -sheet (Figure 12 B). In RNA-binding these mostly hydrophobic residues mediate

unspecific stacking interactions to the RNA bases. Thereby they turn the  $\beta$ -sheet surface into a general RNA interaction platform (Maris, Dominguez et al. 2005). Additionally, these critical residues participate in intra-molecular interaction networks, which ensure proper orientation of other parts of the domain (Kranz and Hall 1999). Target-specificity in RNA-binding is contributed by less conserved residues, which are primarily located in loops of variable length that neighbor the  $\beta$ -sheet or by residues on the outer two  $\beta$ -strands (Scherly, Boelens et al. 1990; Allain, Howe et al. 1997; De Guzman, Turner et al. 1998).

A two-step RNA-binding mechanism has been proposed for U1A, in which the RNA is initially recruited to and oriented on the protein via electrostatic interactions of the phosphate backbone with positively charged residues on the protein surface. These positive charges are mainly constituted by lysines and arginines in loop 3, on the C-terminus and between residues 20-25. They form positively charged patches, which follow the backbone of the RNA on the protein surface and thereby these residues mediate the initial RNA-recruitment (Oubridge, Ito et al. 1994; Katsamba, Myszka et al. 2001; DeLucas, Bray et al. 2003; Law, Linde et al. 2006).

In a second step, close-range contacts, including sequence-specific and unspecific contacts, are established (Katsamba, Myszka et al. 2001). In this respect very important regions in U1A and U2B<sup>''</sup> are loop 3 (aa 46-54) and the variable region on the N-terminus of loop 3. As soon as the RNA is bound, U1A residues Ser48 to Met51 protrude through the 10 nucleotide long U1 snRNA hairpin II-loop and ensure a splaying-out of the bases, which is required for subsequent sequence-specific contacts (Figure 12 C). A key contact is established by Arg52 from loop 3: it contacts the loop-closing base pair C65–G76 and thereby it helps discriminating for the right target (Figure 11 and Figure 25 B). Next to this, the first 7 nucleotides of U1 snRNA hairpin III (nucleotides 66-72) are recognized in a sequence-specific manner. The picture is slightly different for the U2 snRNA hairpin IV, in which the loop-closing base pair is altered to UU (instead of CG as in the U1 snRNA; Figure 11). Moreover the last four loop-nucleotides UACC form a stepladder (Price, Evans et al. 1998). Despite the change in sequence, the loop-closing base pair is still recognized by Arg52 in U2B<sup>''</sup>. Different from U1A-RNA interactions, however, the stepladder forming last 4 loop-nucleotides pack against the region between amino acid 44 and 49 (Figure 24). This protein region is called the variable region, because its sequence is significantly different from the sequence in U1A. Taken together, loop 3 is an important determinant for sequence-specific target recognition in U1A and U2B<sup>''</sup>: it helps splaying out the RNA-bases to promote their sequence specific recognition, it comprises a part of the variable region and it comprises Arg52, which discriminates for the loop-closing base pair.



**Figure 12: Topology and Structure of U1A C-Terminal RRM**

U1A is shown as an example for a typical RNA recognition motif. **A** U1A topology. Cylinders represent  $\alpha$ -helices, arrows are  $\beta$ -sheets. The positions of the RNP-motifs are indicated with bold letters on  $\beta$ -sheet 1 and 3. **B** Structure of U1A (PDB-entry: 1URN). To show the conserved RNP residues more clearly, the RNA was removed. Position 2 on RNP2 (Y13) and position 3 and 5 on RNP1 (Q54 and F56 respectively) are shown as orange sticks. **C** U1A with RNA (PDB-entry: 1URN). The  $\beta$ -sheet aromatic residues provide a general binding platform for the RNA (yellow cartoon). Loop 3 protrudes through the RNA-loop to splay out the RNA-bases. The conserved RNP residues are shown as magenta sticks.

The RNA-binding mechanism, which is applied by U1A- and U2B''-nRRM is the most widespread mechanism employed by RRM to bind RNA: a single-stranded RNA-loop is unspecifically bound to the central  $\beta$ -sheet and sequence specific contacts are established by loop regions, which neighbor the  $\beta$ -sheet. However, there are variations of this scheme: especially when longer RNA-stretches have to be accommodated, the RNA can be bound by two RRM simultaneously (Maris, Dominguez et al. 2005). In addition, not only the single-stranded RNA bases can be recognized, but also the double-stranded RNA stem, which is for example shape- and sequence-specifically bound in the major groove by the human RBMY protein (Skrisovska, Bourgeois et al. 2007). Alternatively, the RNA-binding may even be mediated by structural elements distinct from the central  $\beta$ -sheet: so-called quasi RRM use a short  $\beta$ -hairpin and two adjacent loops for RNA-binding and the  $\beta$ -sheet is not involved at all (Dominguez and Allain 2006).

Apart from the variations in the RNA-binding mechanism, the regulation of RNA-binding can be dramatically different even between homologous RRM. Often C-terminal helices are used to regulate RNA-binding, sometimes even to couple RNA-binding with other processes: for example the C-terminal helix of human CstF-64 RRM unfolds upon RNA-binding and thereby it gives a signal for the formation of the polyadenylation complex (Perez Canadillas and Varani 2003). Another way to influence RNA-affinity is the usage of ancillary proteins. The most prominent example is the U2B''-protein, which only binds RNA in the presence of U2A'. Cap-binding protein 20 (CBP20) requires complex formation with CBP80 prior to RNA-binding, because CBP80 induces folding of many CBP20 elements, which in turn enables RNA-binding (Mazza, Segref et al. 2002). These examples emphasize the broad range of variations used to turn a very widespread, general RNA-binding motif to a domain with high target-specificity and distinct functionality.



## 2.5 Aim of the Present Study

The present study was aimed at gaining further insights into the mechanisms underlying spliceosomal activation. In particular, structural information on the Prp19-complex, which plays a crucial role in the rearrangements just prior to spliceosome activation, was sought. In order to build the basis for such investigations, globular domains of the human and yeast Prp19-complexes components were to be defined, cloned, expressed and purified. Subsequently, it was intended to assess the quality of the preparations and to provide global structural information by biochemical and biophysical characterizations. Coexpression and co-purification analyses were to be conducted in order to reveal interaction networks or to improve protein expression and solubility. The structures of suitable target domains, proteins or complexes were supposed to be analyzed by X-ray crystallography or cryo-EM.

Another goal of this study was the structural and functional investigation of the RNA-binding mechanism of the U11/U12 65K C-terminal RNA recognition motif. This work was initiated as part of an effort to better understand the minor spliceosomal U11/U12 di-snRNP. It was hoped that the results could shed light on similarities and differences between components of the major and minor spliceosome. It was planned to design different protein constructs, which comprise the core RRM, on the basis of the known, homologous U1A-structure. Additionally, different RNA-oligonucleotides should be designed. Both components should subsequently be tested for their interaction properties in gel shift assays in order to find optimally interacting targets for crystallization and structure determination.

### 3 Materials and Methods

#### 3.1 Materials

##### 3.1.1 Chemicals and Fine Chemicals

Acrylamide	Roth, Karlsruhe
Rotiphorese Gel 30 (30 % Acrylamide, 0.8 % Bis-Acrylamide)	
Rotiphorese Gel 40 (38 % Acrylamide, 2 % Bis-Acrylamide)	
Agarose	Invitrogen, Karlsruhe
Ammoniumperoxodisulfate (APS)	Merck, Darmstadt
Bovine serum albumin (BSA)	Sigma, Taufkirchen
Bradford-protein staining solution	Bio-Rad, München
Bromphenolblue	Merck, Darmstadt
Coomassie Brilliant Blue R/G250	Serva, Heidelberg
DNA-molecular weight marker (III, VI)	Boehringer, Mannheim
Dithiothreitol (DTT)	Roth, Karlsruhe
Ethylendiamine-N,N,N',N'-tetraacetic acid (EDTA)	Roth, Karlsruhe
Ethidiumbromide solution (10 mg/ml)	Boehringer, Mannheim
Gel filtration standard	BioRad, München
Glutathione, reduced	Sigma-Aldrich, Steinheim
Glycerol	Merck, Darmstadt
Imidazole	Merck, Darmstadt
Isopropyl $\beta$ -D-1-thiogalactopyranoside (IPTG)	Sigma-Aldrich, Steinheim
Magnesium chloride	Merck Eurolab, Hannover
Precision protein standard marker	BioRad, München
Potassium chloride	Roth, Karlsruhe
Potassium phosphate	Sigma, Deisenhofen
Silver nitrate	Merck Eurolab, Hannover
Sodium chloride	Roth, Karlsruhe
Sodium dodecyl sulphate (SDS)	Serva, Heidelberg
Sypro-orange protein gel stain	Invitrogen, Karlsruhe
N,N,N',N'-Tetramethylethylenediamin (TEMED)	Sigma, Taufkirchen
Tris-hydroxymethyl-aminomethan (Tris)	Roth, Karlsruhe
Triton X-100	Sigma, Taufkirchen
tRNA <i>E. coli</i>	Boehringer, Mannheim

**Chemicals and Fine Chemicals (Continued)**

Yeast extract	Roth, Karlsruhe
Xylencyanol FF	Fluka, Switzerland
4-2-hydroxyethyl-1-piperazineethanesulfonic acid (HEPES)	Calbiochem, USA

Any standard chemicals, organic substances and solvents (purification grade p.a.), which are not listed here, were ordered by one of the following companies: Merck (Darmstadt), Roth (Karlsruhe), Sigma (Taufkirchen), Serva (Heidelberg) or Fluka (Switzerland).

**3.1.2 Media and Antibiotics****Antibiotics**

Ampicillin	Sigma, Deisenhofen
Chloramphenicol	Boehringer, Mannheim
Kanamycin sulphate	Sigma, Deisenhofen
Streptomycin	Merck Eurolab, Hannover

**Media**

Luria-Bertani-broth (LB)-Agar	Q-Biogene, USA
LB-Medium	Q-Biogene, USA
auto-inducing medium	own production

**3.1.3 Enzymes and Enzyme Inhibitors**

Complete EDTA-free <sup>TM</sup> tablets	Roche, Mannheim
Lysozyme	Boehringer, Mannheim
Pefa-Bloc	Biomol, Hamburg
Pfu DNA Polymerase (2.5 U/μl)	Stratagene, Heidelberg
Prescission Protease	own production
Restriction endonucleases	New England Biolabs, Frankfurt
RNasin (40 U/μl)	Promega, USA

**Enzymes and Inhibitors (Continued)**

TEV-protease	own production
T4 DNA Ligase (400 U/μl)	New England Biolabs, Frankfurt
T4 Polynukleotid Kinase (20 U/μl)	New England Biolabs, Frankfurt

**Proteases for Limited Proteolysis and Mass Spectrometry**

Carboxypeptidase Y	Sigma-Aldrich, Steinheim
Chymotrypsin	Roche, Mannheim
Endoproteinase Asp-N	Sigma-Aldrich, Steinheim
Endoproteinase Glu-C	Sigma-Aldrich, Steinheim
Leucin Aminopeptidase	Sigma-Aldrich, Steinheim
Papain	Sigma-Aldrich, Steinheim
Protease X (Thermolysin)	Sigma-Aldrich, Steinheim
Trypsin	Sigma-Aldrich, Steinheim

**3.1.4 Antibodies**

anti-CDC5-C1	AG Lührmann
--------------	-------------

**3.1.5 Nucleotides****Nucleotides**

Desoxynukleotid-5'-Triphosphate (dNTPs, 100 mM)	Amersham, Freiburg
---	--------------------

**Radionucleotides**

[ $\gamma$ <sup>32</sup> P]-ATP (6000 Ci/mmol; 10 Ci/l)	Amersham, Freiburg
---	--------------------

### 3.1.6 DNA-Oligonucleotides

The DNA-oligonucleotides used during this study were only employed in the cloning of expression constructs. In order not to go beyond the scope of this methods section, the table with the used DNA-oligonucleotides can be found in table Table 17 in the appendices (page 124).

### 3.1.7 RNA-Oligonucleotides

oligo	position in human U12 snRNA	sequence
U12-wt	109-125	5'-CCCGCCUACUUUGCGGG-3'
U12-short	110-124	5'-CCGCCUACUUUGCGG-3'
U12-long	G-109-125-C	5'-GCCCGCCUACUUUGCGGGC-3'
U12-5'ov	CG-109-125	5'-CGCCCGCCUACUUUGCGGG-3'
U12-27nt	104-130	5'-GGUGACCCGCCUACUUUGCGGGAUGCC-3'

**Table 3: RNA-Oligonucleotides**

### 3.1.8 Plasmids

vector	pro-motor	selector	tags and fusion partners	protease site	replicon	supplier
pGEX6p 1	tac	Amp	N-GST	Presc	ColE1	Pharmacia
pET-28	T7-lac	Kan	N-His, C-His	Thr	ColE1	Novagen
pET-M11	T7-lac	Kan	N-His, C-His	TEV	ColE1	EMBL
pET-M13	T7-lac	Kan	none	none	ColE1	EMBL
pET-M20	T7-lac	Amp	N-TrxA, N-His, C-His	TEV	ColE1	EMBL
pET-M60	T7-lac	Kan	N-NusA, N-His, C-His	TEV	ColE1	EMBL
pET-Duet1	T7-lac	Amp	N-His(MCS1)	none	ColE1	Novagen
pRSF-Duet	T7-lac	Kan	N-His(MCS1)	none	RSF1030	Novagen
pCDF-Duet	T7-lac	Strep	N-His(MCS1)	none	CloDF13	Novagen

**Table 4: Plasmids**

### 3.1.9 Escherichia coli Strains

BL21 (DE3) pLysS	Novagen, Darmstadt
HB101	Sigma, Deisenhofen
Rosetta (DE3) pLysS, ChlorampR	Novagen, Darmstadt
Rosetta2 (DE3) pLysS, ChlorampR	Novagen, Darmstadt
XL-1 Blue, TetR	Stratagene, US

**3.1.10 Common Buffers**

10 fold PBS pH 7.3 or pH 8.0

1.4 M NaCl

27 mM K<sub>2</sub>CO<sub>3</sub>0.1 M Na<sub>2</sub>HPO<sub>4</sub>18 mM KH<sub>2</sub>PO<sub>4</sub>

10 fold TBE

0.89 M Tris

0.89 M Boric acid

25 mM EDTA pH 8.0

**3.1.11 Commercial Kits**

Qiagen Plasmid Midi/Maxi Kit

QIAprep Spin Miniprep Kit

GFX Purification Kit

Human Hela Marathon-Ready™ cDNA Library

Qiagen, Düsseldorf

Qiagen, Düsseldorf

Amersham, Freiburg

Clontech, Heidelberg

**Crystallization screens**

Anions and Cations Suites

Additive Screens I-III

Additive Screens

Classics and Classics lite

Crystal Screen I and II

Index I and II

Mb class I and II

MPD suite

JCSG screen

Natrix Screen

(NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub> screen

Nucleix Suite

PACT screen

PEG I and II

pH clear I and II

Pre-Crystallization test

Qiagen, Hilden

Hampton Research, USA

own production

Qiagen, Hilden

Hampton Research, USA

Hampton Research, USA

Qiagen, Hilden

Qiagen, Hilden

Qiagen, Hilden

Hampton Research, USA

Qiagen, Hilden

Qiagen, Hilden

Qiagen, Hilden

Qiagen, Hilden

Qiagen, Hilden

Hampton Research, USA

**Crystallization Screens (Continued)**

Protein Complex screen	Qiagen, Hilden
SM I, II and III	Qiagen, Hilden
Salt Rx Screen	Hampton Research, USA

Buffers and reagents which were used for refinement (purification grade p.a.) were purchased at Hampton Research, Sigma-Aldrich or Roth.

**3.1.12 Working Materials**

CD-cuvette, 1 mm	Hellma
Dialysis mebranes MWCO 3500-8000 Da	SpektraPor, USA
Electroporation cuvettes	Bio-Rad, München
Glutathione Sepharose 4B	Amersham, Freiburg
Ni-NTA agarose	Quiagen, Hilden
Pipettes (adjustable)	Eppendorf, Hamburg
Pipetting aid „pipettus-akku“	Hirschmann, Eberstadt
Probe Quant™ G-25 Micro Columns	Amersham, Freiburg
Sterile filter 0.2 µm, 0.45 µm	Millipore, Frankreich
X-ray films Biomax MR	Kodak, USA

**3.1.13 Technical Devices**

Åkta explorer/prime/purifier and columns	Amersham, Freiburg
Autoklave, Varioklav 300-EPZ	H+P Labortechnik, Oberschleißheim
Biofuge (pico/fresco)	Heraeus, Hanau
Chirascan, CD-spectrometer	Applied Photophysics, UK
Crychem plates (sitting drop, 24-well)	Hampton Research, USA
Electroporation device, Gene Pulser	Bio-Rad, München
Eppendorf tubes (0.5, 1.5, 2 ml)	Eppendorf, Hamburg
Falkon tubes (5, 15, 50 ml)	Greiner, Kremsmünster
Image Plate, Mar345	Mar, Norderstedt

**Technical Devices (Continued)**

Geldocumentation device	Bio-Rad, München
Gelelectrophoresis device	Bio-Rad, München; Institutes workshop
Head over tail Rotor (HOT)	Cole-Parmer, USA
Heating block Dri-Block DB-3	Techne, UK
NanoDrop ND-1000	Thermo Fischer, USA
NanoDrop Robot, Cartesian	Zinsser Analytic, USA
pH-meter	Mettler, Toledo
Phosphorimager Typhoon 8600	Molecular Dynamics
Rotating anode, Rigaku RU300	MSC, USA
Scintillation counter LS 1701/TRI-CARB 2100TR	Beckman/Packard, USA
Shaking incubator	New Brunswick, USA
SMART-system	Amersham, Freiburg
Sonifier (Branson 250D)	Heinemann Labortechnik
Sorvall RC 5B Centrifuge	Kendro, USA
Sorvall SA-300/600, SLC-.6000 rotor	Kendro, USA
Spectrophotometer Ultrospec 3000 pro	Biorad, München
VDX plates (hanging drop, 24-well)	Hampton Research, USA
Vivaspin, concentrator MWCO 5-100 kDa, 500 µl – 15 ml	Vivascience, Hannover
Vortex	Janke & Kunkel, Staufen
X-ray film developer X-Omat 2000	Kodak, USA
96-well crystallization plates	Greiner, Kremsmünster
DNA Engine OPTICONTM, real-time PCR	MJ Research, USA



## 3.2 Methods

The methods section first explains applied laboratory methods and in a second part it gives an introduction to crystallization techniques, to data collection and processing strategies, to structure solution techniques and to model building and refinement methods (3.2.6 and 3.2.9). The applied methods are always described at the end of the respective introductory chapter. These chapters shall provide background information to readers, who are not familiar with technical details in protein crystallography.

### 3.2.1 Molecular Biology Standard Methods

#### Concentration Determination of Nucleic Acids

For nucleic acid concentration determination, 1 µl of an aqueous solution was measured against water on the NanoDrop ND-1000, which calculated concentrations from the absorbance at 260 nm according to Lambert-Beer's Law.

#### Agarose Gel Electrophoresis of Nucleic Acids

To separate nucleic acids, agarose gel electrophoresis was performed according to standard protocols (Sambrook, Fritsch et al. 1989). Depending on the size of the analyzed fragments, the agarose concentration was varied between 0.5 and 3 %. A 1-kb or a 3-kb DNA ladder at a concentration of 0.05 mg/ml was loaded in one lane as a marker. The nucleic-acid samples were mixed with DNA-loading buffer. Electrophoretic separation was performed in 1 fold TBE buffer at 50-100 V and the gel was stained in 0.5 µg/ml ethidium-bromide (EtBr). DNA was visualized under UV light.

5 fold DNA loading buffer

30 % glycerol (v/v)

0.25 % bromophenol blue (w/v)

0.25 % xylene cyanol FF (w/v)

### Polymerase Chain Reaction (PCR)

The polymerase chain reaction (PCR) was used for target amplification from cDNA libraries, from yeast genomic DNA and from plasmid sources. Forward and reverse primers were designed to introduce compatible restriction enzyme sites. 3-6 additional bases were added at the 5' ends to allow efficient digestion by restriction enzymes (see <http://www.pepcore.embl.de/cloning/index.html>) and NEB-catalogue for primer design protocols). The annealing temperature was chosen on the basis of the calculated melting temperatures of the primers. A typical PCR-protocol is shown below:

PCR reaction mixture (50 µl)	PCR cycling program
1 µl DNA sample (~100 ng)	95 °C 2'
1 µl 5' oligo (20 pmol/µl)	95 °C 30" } 30 repetitions
1 µl 3' oligo (20 pmol/µl)	55 °C 1' }
5 µl 10 fold cloned Pfu-buffer	72 °C 1'/kb }
5 µl DMSO	72 °C 5'
2 µl dNTP (25 mM each)	hold temperature at 4 °C
2 µl Pfu polymerase	
33 µl H <sub>2</sub> O	

### Site-Directed Mutagenesis

Site-directed mutagenesis was performed according to the protocol provided with the stratagene Quickchange™ kit (<http://www.stratagene.com/manuals/200518.pdf>).

### Enzymatic Digestion and Ligation of PCR-Products or Plasmids

For digestion with the respective restriction enzymes, buffers and temperatures were chosen according to the manufactures instruction (see NEB catalogue). PCR-products were purified before digestion using the GFX purification kit and the DNA was eluted in 30 µl H<sub>2</sub>O. Usually, 5 µg DNA (PCR-product or plasmid DNA) were digested. In cases where the digestion was performed to verify a cloning result, the DNA was separated on an agarose gel and checked for a digestion product of appropriate size. If the digestion was performed to produce DNA suitable for ligation, the DNA was purified via a preparative gel. For this purpose the digested DNA was purified by agarose gel electrophoresis (see page 49), the band was excised from the gel and the DNA was extracted using the GFX purification kit.

Ligation was performed at 16 °C overnight with T4 DNA ligase in a 20 µl reaction volume and a molar ratio of 1:10 (plasmid:insert).

### Transformation of Plasmids in Bacterial Cells and Their Isolation

Prior to transformation, the ligation reaction was precipitated using butanol according to standard protocols (Sambrook, Fritsch et al. 1989). 1-2 µl of the butanol-precipitated DNA were mixed with either 60 µl electro-competent HB101/XL1blue cells for plasmid isolation or with 60 µl electro-competent BL21/Rosetta(2) cells for protein expression. The mixture was transferred to a electroporation cuvette and subjected to a 4.8 msec short pulse of 2.3 kV. After resuspension in 1 ml LB-medium, the cells were grown at 37 °C for 1 h without selective antibiotics. Subsequently, the cells were pelleted, resuspended in a small volume of fresh LB and streaked out on an agar-plate containing the selective antibiotics. For protein expression as well as plasmid purification, single colonies were picked after overnight growth to inoculate a pre-culture for protein expression or a culture for plasmid isolation. For plasmid isolation with the Qiagen kits, a single colony was used to inoculate the respective volumes of LB-medium, which are specified by the manufacturer. The cultures were grown overnight at 37 °C and alkaline lysis was performed according to the manufacturers instruction.

### Cultivation of *E. coli* Cells

*Escherichia coli* cells were cultivated using standard methods (Sambrook, Fritsch et al. 1989). The cells were either grown in liquid media like LB or auto-inducing medium in a shaking incubator at temperatures depending on the experiment or they were grown overnight at 37 °C on LB-Agar plates. The auto-inducing medium was prepared as described in Studier *et al.* (Studier 2005).

Depending on the selection marker of the plasmid, antibiotics in the following concentrations were added to the medium:

Ampicillin	100 µg/ml
Chloramphenicol	34 µg/ml
Kanamycin	50 µg/ml
Streptomycin	50 µg/ml.

### Storage of *E. coli* Cells

For long-time storage at -80 °C, cells with an optical density at 600nm (OD<sub>600</sub>) ~0.6 were mixed with glycerol in a 5:1 ratio (cells to glycerol) and shock frozen with liquid nitrogen.

### Production of Electro-Competent *E. coli* Cells

Bacterial cells were grown in 0.5 l LB-medium to OD<sub>600</sub>=0.4-0.6 and sedimented at 4000 times g for 15 min. The pellet was washed twice with pre-chilled ddH<sub>2</sub>O and twice with ice-cold 10 % glycerol. Afterwards the pellet was resuspended in 4 ml 10 % glycerol, divided into aliquots of 60µl, shock frozen with liquid nitrogen and stored at -80 °C.

### DNA Sequencing

DNA sequencing was either performed at an in-house facility or the samples were sent to Seqlab, Göttingen. In both cases only the appropriate amounts of DNA and sequencing oligonucleotide needed to be supplied.

### 5'-End Labeling of RNA-Oligonucleotides

25 pmol RNA-oligonucleotide were incubated in a volume of 25 µl with 1.5 µl T4 polynucleotidyltransferase (PNK) and 2.5 µl 10 fold PNK-buffer in the presence of 3 µl [ $\gamma$ -<sup>32</sup>P]-ATP (6000 Ci/mmol; 10 µCi/ml). The mixture was left at 37 °C for 1 h. The 5'-end labeled product was diluted to 20 µl and purified using a G25 gel filtration column according to the manufacturer's instruction.

## 3.2.2 Protein Biochemical Standard Methods

### Denaturing SDS-Polyacrylamide Gel Electrophoresis

The denaturing SDS polyacrylamide gel electrophoresis (SDS-PAGE) was performed according to Laemmli *et al.* (Laemmli 1970). Depending on the protein-mixture that had to be separated, the crosslinking-degree and the size of the gel were varied. In this study acrylamide gels of 10 %, 13 %, 17 % and 20 % (37.5:1 acrylamide: bis-acrylamide, 1mm thickness) were used with an increased concentration of TEMED (0.33 % v/v).

Before application on the gel, the protein samples were mixed with protein-loading buffer and incubated 5 min at 95 °C to ensure complete denaturation. After loading the samples on the gel in a gel chamber filled with protein running buffer, the proteins were focused in the stacking gel at 15-25 mA and subsequently separated in the resolving gel at 30-45 mA.

Protein loading buffer

75 mM Tris/HCl pH 6.8  
1.25 mM EDTA  
2.5 % (w/v) SDS  
20 % (w/v) glycerol  
0.1 % (w/v) bromphenolbue  
50 mM DTT

Protein running buffer

25 mM Tris/HCl pH 8.8  
192 mM glycine  
0.1 % (w/v) SDS

Stacking gel buffer (4 fold)

500 mM Tris/HCl pH 6.8  
0.4 % (w/v) SDS

Resolving gel buffer (4 fold)

1.5 M Tris/HCl pH 8.8  
0.4 % (w/v) SDS

Rotiphorese gel 30 with 30 % acrylamide and 0.8 % bis-acrylamide was used for the gel mixtures. To achieve the desired percentages of the gels, the respective amount of Rotiphorese gel 30 was calculated, diluted with water and mixed with 1/4<sup>th</sup> of the corresponding gel buffer. TEMED was added at 0.33 % v/v and the polymerization was started by addition of 10 % APS at 10 times the volume of TEMED in stacking gels and equal volume as TEMED in resolving gels.

### **Gel Staining Methods**

Proteins on SDS-polyacrylamide gels were visualized either by staining and destaining them with Coomassie brilliant blue R250 (Sambrook, Fritsch et al. 1989) or by silver-staining the gel (Blum, Beier et al. 1987). Whereas Coomassie-staining reveals bands of up to 1 µg protein per band, silver-staining is far more sensitive and can detect up to 5 ng in a single protein band.

### **Analytical and Large Scale Expression of Fusion Proteins in *E. coli***

High amounts of purified protein are needed for X-ray crystallographic studies. To optimize the expression levels of the proteins, different conditions for expression were usually tested.

The most commonly tested conditions were:

<i>E. coli</i> strain	medium-type	IPTG [mM]	OD <sub>600</sub> at induction	temperature [°C]	expression time [h]
Rosetta(2)/BL21(DE3)	LB	1 or 0.5	0.7, 1.7	30	3
Rosetta(2)/BL21(DE3)	LB	1 or 0.5	0.7, 1.7	20	3-5
Rosetta(2)/BL21(DE3)	LB	1 or 0.5	0.7, 1.7	16/18	5-12

**Table 5: Commonly Tested Expression Conditions with LB-Medium**

Rosetta(2) stands for Rosetta(2)(DE3)pLysS as listed above (see 3.1.9).

<i>E. coli</i> strain	medium-type	temperature [°C]	expression time [h]
Rosetta(2)/BL21(DE3)	auto-inducing	30	~18-24
Rosetta(2)/BL21(DE3)	auto-inducing	20	~24-36
Rosetta(2)/BL21(DE3)	auto-inducing	16/18	~24-36

**Table 6: Commonly Tested Expression Conditions with Auto-Inducing Medium**

Rosetta(2) stands for Rosetta(2)(DE3)pLysS as listed above (see 3.1.9).

The culture volume ranged from 50 to 100 ml. Shaking flasks had at least 3 times the volume of the culture. The IPTG-concentration was lowered to 0.1 mM in some cases.

Prior to induction, 500 µl culture at OD<sub>600</sub>=0.6 served as a non-induced control. The sample was centrifuged 1 min at 13000 rpm, the cell pellet was resuspended in 50 µl protein loading buffer and boiled for 5 min at 95 °C. After over-expression a sample of equal OD<sub>600</sub> was subjected to the same procedure. The non-induced and the induced sample were separated via SDS-PAGE and visualized with Coomassie staining. Overexpression of target proteins is detected by comparing the thickness of bands of a target size. Following expression, the cells were sedimented at 4000 rpm for 15 min and resuspended either in 1 fold PBS-buffer for glutathione-sepharose affinity purification or in W0-buffer for Ni-NTA purification (see page 56).

In both cases the buffers contained Pefa-Bloc and 2 mM DTT. Finally the cells were shock frozen in liquid nitrogen and stored at -80 °C.

The optimal expression condition was not only determined by the overexpressed amount of protein, but rather chosen after initial solubility tests according to the amount of soluble protein (see page 56).

Having determined the optimal expression condition, the culture volume was scaled up (usually to 6 l). The cells were pelleted after expression and resuspended in the buffer required for the respective affinity purification. EDTA-free complete<sup>TM</sup> was added as a protease inhibitor. Afterwards the cells were shock frozen in liquid nitrogen and stored at -80 °C.

### Coexpression of Proteins

For most proteins of the human Prp19-complex and some proteins from the yeast Prp19p-complex, coexpression was tested, because the interaction of coexpressed proteins may have a positive effect on protein solubility. The two plasmids of choice were co-transformed into an *E. coli* strain as described in the previous section (see page 51).

Standard conditions for the coexpression were the following:

<i>E. coli</i> strain	medium-type	IPTG [mM]	OD <sub>600</sub> at induction	temperature [°C]	expression time [h]
Rosetta	LB	1	0.6, 1.7	30	3
Rosetta	LB	1	0.6, 1.7	18	min. 13

**Table 7: Conditions for Coexpression**

Rosetta(2) stands for Rosetta(2)(DE3)pLysS as listed above (see 3.1.9).

The combinations of vectors will be described in the respective results section.

After expression, the culture was divided in two parts, which were then subjected to affinity purification with either glutathione sepharose or with Ni-NTA (see page 56).

### Cell Lysis

The *E. coli* cells were disrupted by sonication. Proper cooling was accomplished with a NaCl-ice-bath and a number of short pulses (duration 5-10 s) with pauses (duration 10-30 s) to re-establish a low temperature. Subsequently, the lysate was centrifuged at 15000 rpm for 45 min to pellet the cell debris. The supernatant was supplied to affinity purification as described below (page 56).

### Purification and Cleavage of GST-Tagged Fusion Proteins

Glutathione sepharose 4B was pre-swollen twice with PBS-buffer pH 7.3 containing 2 mM DTT. 8 ml sepharose-slurry was taken for 100 ml lysate volume. The beads were mixed with the lysate and incubated  $\frac{1}{2}$ -1 h at 4 °C on a head-over-tail rotor. After sedimenting the beads at 2000 rpm for 3 min, the supernatant was discarded and the beads were washed 3 times with a 10 fold volume of PBS-buffer pH 7.3. Before elution 3 washing steps with either 0.5 M Tris pH 8.0 (very stringent condition) or with 50 mM Tris pH 8.0, 150 mM NaCl were performed. In the latter case, another washing step with 50 mM Tris pH 8.0, 700 mM NaCl was sometimes included to remove residual impurities. The protein was eluted from the beads by incubating either 0.5 M Tris pH 8.0, 100 mM reduced glutathione (stringent conditions) or 50 mM Tris pH 8.0, 150 mM NaCl, 100 mM reduced glutathione for 5 min at 4 °C on the head-over-tail rotor. Elution steps were repeated until all protein was eluted from the beads. All buffers contained 2 mM DTT.

The GST-fusion tag was cleaved from the protein in three possible ways: directly in the elution buffer, in the elution buffer during dialysis or on the beads. The latter two methods were normally applied in case of large scale protein preparations. Under all circumstances Prescission protease was added in a 1:20-1:40 ratio (Prescission/protein; mg/mg). The ratio was dependent on the completeness of the cleavage, which was monitored using SDS-PAGE and subsequent Coomassie staining.

### Purification and Cleavage of His-Tagged Fusion Proteins

Ni-NTA agarose was pre-swollen twice in  $W_0$ -buffer containing 2 mM DTT and mixed with the lysate of the overexpressed cells. After  $\frac{1}{2}$ -1 h incubation on a head-over-tail rotor at 4 °C, the beads were sedimented at 2000 rpm for 3 min. The supernatant was removed and the beads were washed twice with  $W_0$  using a buffer volume, which was 10 times the volume of beads. The washing step was repeated with  $W_1$ ,  $W_2$  and finally  $W_3$  (all containing 2 mM DTT). For elution, only half of the volume of beads was taken from the elution-buffer and incubated with the Ni-NTA for 10 min on the head-over-tail rotor at 4 °C. The elution step was repeated until all protein is eluted from the beads.

Buffer  $W_2$  was only applied in some cases, because the high salt concentration leads to very stringent conditions. Especially for the purification of coexpressed proteins it was not used.

From each purification step starting at the removal of cell debris after lysis a sample is taken, applied to SDS-PAGE and visualized with Coomassie staining.



W<sub>0</sub> or lysis buffer

20 mM Tris HCl pH 8.0  
10 mM imidazole pH 8.0  
150 mM NaCl  
0.2 % (v/v) NP-40  
2 mM DTT

W<sub>1</sub> buffer

20 mM Tris HCl pH 8.0  
10 mM imidazole  
150 mM NaCl  
2 mM DTT

W<sub>2</sub> buffer

20 mM Tris HCl pH 8.0  
10 mM imidazole  
1 M NaCl  
2 mM DTT

W<sub>3</sub> buffer

20 mM Tris HCl pH 8.0  
50 mM imidazole  
150 mM NaCl  
2 mM DTT

Elution buffer

10 mM Tris HCl pH 7.0  
300 mM imidazole  
150 mM NaCl  
2 mM DTT

Cleavage of the 6-His-fusion tag was achieved using a ratio of 1:10-1:20 TEV-protease:fusion-protein (mg/mg). In most cases the reaction was carried out directly in the elution buffer. Completeness of the cleavage was controlled using SDS-PAGE and Coomassie staining.

**Determination of a Purification Protocol**

To determine which column material is best suited for the intermediate purification, the following column resins were tested in a small scale for binding and elution properties of the target protein:

1. Sepharose CL4B (buffer A)
2. DEAE-sepharose
3. SP-sepharose
4. CM-sepharose
5. Heparin
6. Hydroxylapatite
7. Methyl-sepharose
8. Butyl-sepharose
9. Octyl-sepharose
10. Phenyl-sepharose
11. Sepharose CL4B (buffer B)

**Buffer A**

50 mM Tris pH 7.0

50-150 mM NaCl

2 mM DTT

**Buffer B**

50 mM Tris pH 7.0

1 M NaCl

2 mM DTT

100 µl of the matrix slurry of materials 1-6 were equilibrated in buffer A, the same amount of materials 7-11 were equilibrated with buffer B. The protein solution was rebuffed into the respective buffers and 200 µl of ~0.5 mg/ml were mixed with the beads and incubated for a few minutes at 4 °C on a head-over-tail rotor. The supernatant was removed after a short centrifugation at 3000 rpm and 100 µl of it were mixed with protein loading buffer. The beads were washed 2 times with 1 ml of the respective equilibration buffer. To elute the protein, 200 µl elution buffer was added to the beads and incubated at 4 °C for 1 min on a head-over-tail rotor. Subsequently the sample was spun down and 100 µl of the supernatant were mixed with protein loading dye. The elution step was repeated to ensure proper elution especially in the case of reverse phase materials. The samples were analyzed with SDS-PAGE and visualized with Coomassie staining. The column resin, which bound and eluted the target protein, but preferably no other proteins, was chosen for large scale purification.

Only for very few proteins this procedure was not necessary, because little contamination was detected after the capture step (glutathione sepharose or Ni-NTA purification) and the proteins could be separated from their cleaved fusion partner in size exclusion chromatography.

To remove residual contaminants and transfer the protein in a final working buffer, size exclusion chromatography was applied. Before this polishing step, the protein was concentrated to at least 3 mg/ml. As a general running buffer 20 mM Tris pH 7.5, 150 mM NaCl, 2 mM DTT was chosen. The pH and the salt concentration of the buffer were adjusted after testing the behavior of concentrated protein under different other buffer conditions. Depending on the size of the target protein, Superdex 75 or Superdex 200 were chosen for size exclusion chromatography. For large scale purifications only HiLoad 26/60-columns were utilized.

**Purification Protocol for Truncated 65K Proteins**

The cells were sonicated and the soluble protein was captured using glutathione sepharose 4 FF beads pre-swollen in PBS-buffer with 5 ml beads/l culture volume and 30 min incubation. Beads were washed with PBS-buffer and subsequently 50 mM Tris, 150 mM NaCl, 2 mM DTT. Concentration of protein on the beads was estimated, eluting

protein from 100 µl beads with 50 mM Tris pH 8.0, 150 mM NaCl, 2 mM DTT, 100 mM reduced glutathione and measuring the protein concentration in a Bradford-assay. Cleavage of the GST-tag was performed by adding 1/40 (mg/mg) Prescission-protease to the beads and incubating overnight. The supernatant was applied on a Heparin column and the protein was eluted using a salt-gradient from 150 mM-1 M NaCl. Peak fractions were pooled and subjected to gel-filtration on a Superdex 75 26/60 column with 20 mM Tris 7.5, 150 mM NaCl, 2 mM DTT as a running buffer.

### Concentration of Proteins and Determination of the Concentration

Protein solutions were concentrated using a membrane with a molecular weight cutoff, which was around 3 times less than the molecular weight of the respective protein. The concentration was determined with a Bradford assay: up to 20 µl protein was diluted to 800 µl with ddH<sub>2</sub>O and mixed with 200 µl Bradford solution. After 5 min incubation, the extinction at 595 nm was measured in a spectrophotometer. By comparison with a BSA standard curve, the concentration of the protein solution was determined.

To obtain the optimal concentration for crystallization, a pre-crystallization test was carried out. Alternatively the protein was crystallized in the Hampton classics screen and the number of drops, in which the protein precipitated, was counted. The target concentration was reached when around 1/3<sup>rd</sup> of all conditions showed precipitation 1 h after crystallization.

### 3.2.3 Bioinformatic Methods

Only the most frequently used online-servers and programs are listed and grouped according to application. Crystallography software is specified in the methods sections 3.2.7 to 3.2.9.

#### Alignments

Clustal W–multiple sequence alignment (<http://align.genome.jp/>)

BLAST-Basic Local Alignment and Search Tool (<http://www.ncbi.nlm.nih.gov/blast/Blast.cgi>)

Clustal X-program: version 1.81

Vector NTI-program: version 9-10 (invitrogen)

### Domain Prediction

SMART-Simple Modular Architecture Research Tool (<http://smart.embl-heidelberg.de/>)

PROSITE-Database (<http://www.expasy.ch/prosite/>)

### Fold Prediction

Foldindex (<http://bip.weizmann.ac.il/fldbin/findex>)

IUPred-Prediction of Intrinsically Unstructured Proteins (<http://iupred.enzim.hu/>)

DisEMBL-intrinsic protein disorder prediction (<http://dis.embl.de/>)

Globplot-intrinsic protein disorder, domain and globularity prediction (<http://globplot.embl.de/>)

### Other Programs

Common databases for literature and structures are not listed.

BIND-Biomolecular interaction database (<http://bond.unleashedinformatics.com/Action>)

COILS-prediction of coiled coil regions in proteins

([http://www.ch.embnet.org/software/COILS\\_form.html](http://www.ch.embnet.org/software/COILS_form.html))

SGD-*Saccharomyces* Genome Database (<http://www.yeastgenome.org/>)

## 3.2.4 Special Methods

### Electron Microscopy

For electron microscopy, protein samples with a concentration of ~0.25 mg/ml were supplied. The following steps were performed by Dr. M. Golas and Dr. B. Sander. The sample was rebuffed in 20 mM HEPES pH 7.9, 100 mM NaCl and either directly adsorbed onto a carbon film for electron microscopy (EM) (as described in (Golas, Sander et al. 2003)) or subjected to a glycerol-gradient centrifugation prior to EM specimen preparation. In either case, both unfixed and glutaraldehyde-fixed samples, were analyzed. To provide unfixed and fixed gradient fractions, two gradients – one with and one without glutaraldehyde – were prepared. The isokinetic ultracentrifugation in the presence of glutaraldehyde was performed as described by Kastner *et al.* (Kastner, Fischer et al. 2008). The gradient ranged from 5-20 % glycerol, its total volume was 4 ml. After removal of the upper 200 µl of the gradient, a 200 µl 5 % glycerol-cushion was applied and covered with 200 µl of the sample. Sedimentation was achieved by centrifugation at 35000 rpm for 18 h at 4 °C. The gradient

was harvested in 300  $\mu$ l fractions. The peak fractions were adsorbed on carbon-films and stained in a 2% uranyl formate staining solution for 2 min (Golas, Sander et al. 2003). The incubation times of the sample to the carbon were varied from 5 min to 45 min. Negative stain images of the particles were visualized at room temperature using a Philips CM200FEG transmission electron microscope operated at 200kV.

### Domain Mapping with Limited Proteolysis

Limited proteolysis and subsequent mass spectrometric analysis was applied to detect globular folded domains in the proteins of interest. In order to find stable fragments, which persist during proteolysis, either the incubation times are varied at a constant protease concentration or the incubation time is kept constant and the protease concentration is varied. In this study the latter method was chosen. 8  $\mu$ l of 1 mg/ml protein was incubated with a protease:protein-ratio (mg/mg) of 1:10, 1:100 and 1:1000 for  $\frac{1}{2}$  h at room temperature. The reaction is stopped by addition of 1  $\mu$ l PMSF and 5  $\mu$ l protein loading buffer. Subsequently, the samples were boiled for 5 min. Separation of the sample was achieved by SDS-PAGE and the bands were visualized with Coomassie staining.

In case of the U11/U12 65K-cRRM protein, the limited proteolysis was performed on the protein alone and on a protein:RNA-mixture of 1:1.5 molar ratio.

Stable fragments were analyzed by mass spectrometry. The analysis was performed by U. Plessmann and M. Raabe. Analysis of the data was carried out as described in chapter 3.2.4.

### Gel Shift Assays and $K_D$ -Determination

To detect and quantify protein/RNA-interactions, the ability of the protein to shift a radioactively labeled RNA in a native gel was tested. For this purpose, 0.25 pmol of the 5'-end labeled RNA-oligonucleotide was incubated with the denoted amounts of protein and 1  $\mu$ l *E. coli* tRNA (10 mg/ml) in a reaction volume of 10  $\mu$ l for 1 h at 4°C. Subsequently, 5  $\mu$ l native loading buffer were added to the reaction and 5  $\mu$ l of the mixture were loaded on a native gel. The best type of native gel was determined to consist of a 6 % mixture of 1:29 bis-acrylamide:acrylamide in 1 fold TBE. The gel was polymerized with 0.05 % TEMED (v/v) and 0.5 % APS (v/v) and pre-chilled at 4°C prior to usage. Protein separation was performed at 8 W and 4 °C in 0.5 fold TBE.

Thereafter, the gel was dried on Watman paper at 80 °C for 1 h. The radioactively labeled RNA was visualized with autoradiography on X-ray films or on a phosphoimager plate.

Quantification was performed with the program ImageQuant on scans from the phosphoimager. The apparent  $K_D$  is the protein concentration, at which half of the RNA is shifted.

### 3.2.5 Methods for Biophysical Characterization

#### CD

In order to record a circular dichroism spectrum and to monitor thermal denaturation, 250-350  $\mu$ l of a ~0.5 or 1 mg/ml concentrated protein solution (concentration is specified in the respective result section) were filled into a 1 mm thick CD-cuvette. The optimal buffer substance was 20 mM Na/K- $PO_4$ . As a salt NaF was usually chosen instead of NaCl, because chloride-ions interfere with the signal in the far UV. Salt concentration and pH were chosen depending on the experiment.

Spectra were recorded in the range of 190-260 nm with a Chirascan CD-spectrometer. The pathlength was 10 nm, the bandwidth 1 nm and the accuracy was set to 0.2 °C. A minimum of 3 spectra were averaged and a blank spectrum (only buffer) was subtracted. To monitor thermal unfolding, spectra were recorded before thermal melting at 4 °C, during thermal melting at the peak temperature (82 °C or 95 °C) and after melting at 4 °C.

Thermal denaturation was monitored at a constant wavelength: 222 nm for  $\alpha$ -helical proteins and 220 nm for mixed  $\alpha$ - $\beta$ -proteins. The temperature was increased from 4-82 °C or to 95 °C with a slope of 1 °C/min.

#### Thermal Denaturation using Real-Time PCR

The thermal denaturation with real-time PCR monitoring was only applied in the case of U11/U12 65K-cRRM and its deletion mutants. A titration curve with different protein concentrations ranging from 0.2 nmol/ $\mu$ l to 2.3 nmol/ $\mu$ l was prepared with 20 mM HEPES pH 7.5, 150 mM NaF, 2 mM DTT. NaF was chosen as a salt to ensure comparability to CD-assays. The protein was mixed with 2  $\mu$ l 0.5 fold Sypro-orange protein gel stain and buffer was added to reach a reaction volume of 20  $\mu$ l. In a PCR-plastic stripe with small volume tubings the samples were set in the heating block of the real-time PCR machine. The samples were subjected to a temperature gradient from 4-95°C with a slope of 1 °C/30 s and one read-out per 1 °C.

### Analytical Gel Filtration

Analytical gel filtration runs were performed on a SMART-system at 4 °C. Superdex 75 or 200 PC 3.2/30 columns were utilized, depending on the protein size. A typical buffer was: 20 mM Tris pH 7.5, 150 mM NaCl, 2 mM DTT, whereas the type of salt and its concentration were often varied as stated in the respective results section. Gel filtration standard was applied to the column in the mentioned buffer.

Normally the flow rate was set to 40 µl/min and the fraction size was 40 µl. In the majority of the cases, 50 µl of 1 mg/ml protein were subjected to gel filtration.

### 3.2.6 Crystallization

In this chapter the general theory behind protein crystallization, the methods, which are nowadays employed in protein crystallization and crystallization screens, will be discussed. Additionally the applied methods will briefly be explained. For an overview on possible purification strategies, domain definition techniques and biophysical methods for protein characterization and quality assessment the reader is referred to chapter 8.2. Parts of this chapter are based on (McPherson 1999).

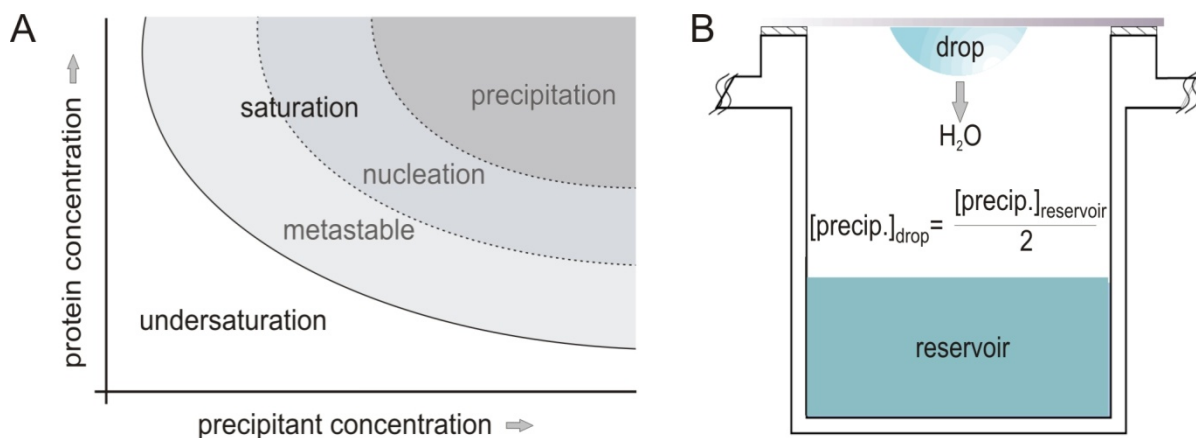
#### General Overview on Protein Crystallization

The aim of protein crystallization is to create a solution supersaturated with protein, which produces single, well ordered crystals of an appropriate size. A protein or any other molecule in solution can exist in three different saturation states: it can be undersaturated, saturated and supersaturated. The best way to visualize this is a phase diagram like the one in Figure 13. The solubility of most compounds depends on properties such as pH and temperature, but in any case solubility is limited to a defined amount of material per volume. Below this solubility limit the solution is said to be undersaturated, above the limit it is supersaturated and material becomes excluded from solution with a net accumulation of solid state.

The supersaturated zone in a phase diagram can be divided into three regions, which merge in a contiguous manner (dotted lines in Figure 13): the precipitation zone, the labile zone and the metastable zone. In the precipitation zone, a protein will come out of solution rapidly as an amorphous precipitate, because of extreme supersaturation. In the labile or nucleation zone crystal nuclei can form spontaneously and grow. The higher the degree of

supersaturation in the labile zone is the higher is the amount of nuclei formed. Therefore it is the clear aim of a crystallization approach to bring the system to a state of low supersaturation in the labile zone. Under these conditions a few nuclei would be formed and thereby the system would be driven out of the labile into the metastable zone, where no further formation of nuclei, but only the growth of existent nuclei occurs.

A vast number of crystallization techniques have been developed, but by far the most popular is the vapor diffusion method using sitting or hanging dropa (Figure 13). For this setup, a protein is mixed with precipitant solution and equilibrated over a reservoir in a sealed environment. Due to the mixing of protein and precipitant, the precipitant solution in the drop is more diluted than the one in the reservoir. A driving force acts upon any sealed system, which is out of its thermodynamic equilibrium, that drives it back to equilibrium. The two solutions finally reach equilibrium when the osmolarity of the drop, by losing water due to evaporation, becomes equal to the osmolarity of the reservoir. By means of this vapor diffusion, the drop decreases in volume and as a consequence the concentrations of all components, including that of the macromolecule, in the drop rise significantly. In this way the drop slowly becomes supersaturated and under ideal circumstances the protein will undergo a phase transition and form a highly ordered crystal.



**Figure 13: The Process of Protein Crystallization**

**A** Phase diagram for protein crystallization. **B** Setup for hanging drop crystallization. The precipitant concentration is abbreviated as [precip.].

Even though the pace of discovery in recent years in fields like molecular biology, crystallization techniques, cryo-crystallography, synchrotron-radiation, detector-techniques and computing has lead to tremendous improvements, crystallization itself remains a trial and



error process. As an example: improvements in synchrotron radiation nowadays allow a crystal with dimension of only tens of a micrometer on one edge to produce the desired data. But under which conditions even crystals of this small size can be produced is still not predictable. First attempts to rationally predict crystallization conditions with a neural network algorithm in a structural genomics project have been reported though (DeLucas, Bray et al. 2003) and will surely be a focus of future developments.

Screening chemical composition space for crystallization is an enormous task: 383 different compounds have been used in crystallizations reported in the Biological Macromolecular Crystallization Database (Kundrot 2004). There are four general categories of precipitants: salts (*e. g.* ammonium sulfate), volatile organic solvents (*e. g.* ethanol), polymers (*e. g.* polyethylene glycol (PEG)) and non-volatile organic alcohols (*e. g.* 2-Methyl-2,4-pentanediol (MPD)). Ideally many combinations of these precipitants with other additives in a broad pH-range should be tested in initial screening. Two major types of screens have been developed: sparse matrix and grid screens. A sparse matrix screen consists of trial conditions selected from known crystallization conditions for macromolecules in the Macromolecular Crystallization database (Gilliland, Tung et al. 1994; Gilliland, Tung et al. 2002). Contrary to this random sampling of different conditions are the grid screens. In grid screens a certain precipitant concentration is varied systematically against the pH. Because they consume a high amount of material, grid screens are nowadays employed primarily in refining steps. Intermediate between sparse matrix and grid screens are screens, which vary one variable more systematically (*e.g.* pH) and others somewhat randomly (*e.g.* salt). Most screens, which are solely based on one type of precipitant (*e.g.* the PEGs-, MPD-,  $(\text{NH}_4)_2\text{SO}_4$ -screen) are designed following this principle.

Besides the screening technique and the type of screen, physical and chemical parameters like temperature, protein concentration and drop size can be varied in initial screening. A good recommendation is to screen at least 250 sparse matrix conditions at 4 °C and 20 °C with a common drop size (Kundrot 2004).

Successful hits in initial screening often result only in small crystals that have to be optimized in size and shape. This is done by refining the condition usually in a grid-like fashion. Often the addition of additives or the application of a seeding technique (macro-, micro- or streak seeding) facilitates crystal growth and quality.

Once diffraction-quality crystals are obtained, a suitable condition for cryo-protecting the precipitant solution in the drop has to be found. Cryo-protection is necessary, because datasets are usually recorded under cryogenic conditions in order to minimize the radiation damage during exposure. A cryo-protectant preserves the crystal from breaking due to the volume gain of water during freezing. Additionally it prevents scattering of ordered water

molecules in an ice-crystal by disrupting the local order of water molecules, so that it freezes glass like. After freezing, data can be collected on the crystal.

### Applied Methods

In this project initial screening was performed in a 96-well format, sitting drop setup. The protein was spun down prior to crystallization at 13 krpm for 5 min. Drop volumes of 200  $\mu$ l with a 1:1 (protein:precipitant) ratio were pipetted by the Cartesian NanoDrop robot at 4 °C and 20 °C. The screens listed above (see page 46) were usually tested in initial screening. The exact number of tested conditions and the protein concentration for crystallization will be denoted in the respective chapters. Crystal growth was monitored 1, 3 and 7 days and 2 and 4 weeks after setup. In most cases, crystal trays were also checked again at later time points.

Initial hits were refined in a grid-like fashion and by the use of additive screens. In many cases, streak and micro-seeding were performed.

In case of the 65K<sup>380-517</sup> protein, annealed RNA was mixed with the protein 1 h before crystallization in an RNA-binding buffer, which contained 1.5 mM MgCl<sub>2</sub>, 150 mM NaCl, 20 mM Tris pH 7.5 and 2 mM DTT. Annealing of the RNA was achieved by heating the RNA dissolved in water for 2 min in a 80 °C water bath and rapidly cooling it afterwards. Directly before crystallization the sample was spun down at 13 krpm for 5 min.

Cryo-protectants were tested by checking a titration curve of cryo-protectant mixed with reservoir for the scattering behavior upon exposure to an X-ray beam. A crystal was transferred to a droplet with cryo-protected reservoir (if necessary in steps of increasing concentrations) and shock frozen in liquid nitrogen for data collection.

### 3.2.7 Data Collection and Data Processing

#### General Overview on Data Collection and Data Processing

In order for an object to diffract light, the wavelength  $\lambda$  of the light must approximately be not larger than the object itself. In a protein molecule bonded atoms are only about 0.15 nm or 1.5 Å apart. Electromagnetic radiation of this wavelength falls into the X-ray range (0.1-100 Å) and wavelengths of particular interest for crystallography are 0.4-2.5 Å. Conventional X-ray sources are: the sealed tube, the rotating anode and synchrotron radiation. Synchrotron radiation is characterized by a high intensity, low divergence and the tunability of the wavelength in a continuous spectrum.

When a crystal is placed in the path of an X-ray beam, the electrons in the molecules which make up the crystal diffract the source beam into many discrete beams. W.L. Bragg showed that the angles at which diffracted beams emerge from a crystal can be computed by treating diffraction as reflection from sets of distant parallel planes in a crystal. A crystal is built up from small building blocks called unit cells. A unit cell is the smallest unit from which an entire crystal can be built up only by translational operations. The most obvious sets of planes that can therefore be assigned in a crystalline lattice are those determined by the faces of the unit cells. If a beam is scattered by a certain set of lattice planes (identified by the coordinates  $h$ ,  $k$  and  $l$  also called Miller indices) it will interfere with beams diffracted by other planes in that set. These beams will only lead to constructive interference if they have an equal phase. This in turn can only be the case when they can travel a multiple integer of their wavelength before they interfere with each other. The mathematical expression of this phenomenon is the Bragg's law:

$$2d_{hkl} \sin \theta = n\lambda$$

with  $d_{hkl}$ : spacing between planes,  $\lambda$ : wavelength,  $\theta$ : incoming = reflecting angle,  $n$ : order of diffraction (integer).

When Bragg's law is fulfilled, each of the resulting beams will produce a distinct spot (reflection) on a detector. In a diffraction experiment each atom contributes to the intensity and the phase of a reflection. These many contributors make the mathematical description of the waves very complicated. The French mathematician Fourier showed that even the most intricate periodic functions can be described as the sum of simple sine and cosine functions whose wavelengths are integral fractions of the wavelength of the complicated function. Such a sum is called a Fourier series. The Fourier series that describes a diffracted ray is called a structure-factor equation, which is called  $F_{hkl}$  for the reflection  $hkl$  and looks as follows:

$$F_{hkl} = \sum_{i=1}^N f_i [e^{2\pi i(hx_i + ky_i + lz_i)}] = |F_{hkl}| e^{i\Phi_{hkl}}$$

with  $f_i$ : atomic scattering factor of atom  $i$ ,  $N$ : number of atoms,  $hkl$ : Miller indices,

$x_i, y_i, z_i$ : coordinates of the  $i^{\text{th}}$  atom,  $hkl$ :  $\Phi_{hkl}$  phase of  $F_{hkl}$ ,  $i = \sqrt{-1}$ .

To collect a full set of reflections, it is necessary to rotate the crystal, in small steps, through an angle determined by its symmetry and orientation. Recent developments even enable a data collection with continuous rotation of the crystal using a special detector type, which can record data "on the fly". This method is faster and it eliminates partial reflections.

After collecting the reflections, the data have to be processed and scaled. Processing data is a broad expression for a procedure that includes many steps like refining setup and crystal

parameters, indexing reflections, obtaining unit cell dimensions and spot intensities and calculating standard deviations. After the processing, all images are merged in a procedure called scaling. As a first step in scaling the space group is defined. Then all symmetry operations for this particular space group are carried out, symmetry related reflections are averaged and so only unique reflections are filtered out. The result is a list containing all unique reflections with their respective indices (h, k and l) and their intensities. Scaling yields the data statistics such as completeness of the dataset, and is therefore important for judging the quality of the data.

### Applied Methods

Data were collected at the Swiss Light Source (SLS) in Villigen, Switzerland.

For the U11/U12 65K-cRRM structure the crystal was measured at 0.984 nm wavelength, rotated by 0.5 ° and exposed for 0.5 s with Al1 as attenuating filter. 360 frames were recorded at a detector distance of 240 mm. The data were processed using the HKL2000 suite and scaled with the Scalepack program. Statistics on the data can be found in the respective chapter (chapter 4.4.3).

### 3.2.8 Structure Solution

#### General Overview on Structure Solution and Model Building

Diffracted X-rays contain the complete information to obtain the proteins three-dimensional structure. This information is encoded in the three parameters that define an electromagnetic wave: the wavelength  $\lambda$ , the amplitude  $A$  and the phase  $\Phi$ . The problem about X-rays is that they cannot be focused as other light or electrons can, but they simply penetrate most materials. Therefore the diffracted beams are visualized as exposed spots on a detector plate. These spots still contain information about amplitude as it is the square root of the measured intensity of a spot, but all information about the phase is lost. This phase information has to be obtained computationally.

Several methods have been developed in order to deduce phases, including direct methods, molecular replacement (MR), single and multiple isomorphous replacement (SIR and MIR) and single- and multi-wavelength anomalous dispersion (SAD and MAD). Combinations of several of the abovementioned approaches, e.g. MIRAS (multiple isomorphous replacement using anomalous scattering) and SIRAS (single isomorphous replacement using anomalous scattering) proved to be useful for experimental phasing of a number of structures. Three of

the most common methods to obtain phases are MIR, MAD and MR. As MR was applied in this work, the method shall briefly be explained.

Molecular replacement (MR) can be used when a suitable model structure for (a part of) the protein in the crystal is already known. As a rule of thumb, MR is straightforward if the known model and the unknown protein share approximately 40 % sequence identity. It is also applicable if, for another reason than sequence identity, the two structures are expected to have a very similar fold of their polypeptide chain. So called Patterson maps of the model structure and the data from the data collection are calculated. A Patterson map is a vector map, which is constructed from the processed raw data. In this vector map, vectors between all atoms in the unit cell are plotted. The Patterson map does not need phase information. To position the model, the model is first rotated around its three rotation axes and then translated. Matching of the model position to the recorded data is checked with correlation-coefficients between the Patterson maps and analysis of the unit cell packing. When a correct position is found, the phases of the model in that position are used for the phase calculations for the measured data. A disadvantage of this method is the possibility to obtain a structure, which is biased by the model structure.

After phase determination, the electron density function  $\rho$  as the final result of an X-ray experiment can be calculated for every point in the crystal ( $x$ ,  $y$  and  $z$ ) from the structure factor amplitude  $|F_{hkl}|$ , the corresponding phase  $\Phi_{hkl}$  and the volume of the unit cell  $V$ :

$$\rho_{xyz} = 1/V \left[ \sum_{hkl} F_{hkl} e^{i\Phi_{hkl}} e^{-2\pi i(hx_1+ky_1+lz_1)} \right]$$

This last equation points out the importance of the phases for the calculation of the electron density.

### Applied Methods

To calculate the initial phases of U11/U12 65K-cRRM, molecular replacement with the U1A structure as a search model (PDB-entry: 1URN\_A) was performed. Molrep from the CCP4-suite was the employed program for this purpose (CCP4-program-suite 1994).

### 3.2.9 Model Building and Refinement

#### General Overview on Model Building and Refinement

In the calculated electron density, a model for the target protein is built by one of two ways: with the help of molecular graphic programs the protein is either built from scratch or an

existing model structure is adjusted in the density. The built model then has to be refined to agree with the data, which have been recorded during data collection. A number of iterative steps are performed during refinement: first the x, y and z coordinates of the atoms and their respective occupancies are adjusted by a manual inspection of the protein structure. In a second step an energy minimization as well as a refinement of the B-factors (factor for mobility or flexibility of an atom in the structure) is carried out by computer programs.

The refined parameters are used to calculate new structure factors  $F_{\text{calc}}$ , which are then compared to the structure factors observed in the experiment  $F_{\text{obs}}$ . Their correlation is quantified in the so called R-factor:

$$R = \left[ \frac{\sum_{hkl} (|F_{\text{obs}}(hkl)| - |F_{\text{calc}}(hkl)|)}{\sum_{hkl} F_{\text{obs}}(hkl)} \right]$$

The lower the R-factor, the better the built model resembles the measured data. Refinement and model building are carried out iteratively to decrease the R-factor. For intermediate resolutions (2-3 Å) the R-factor usually is around 20 %. A small set of reflections (usually 5–10 %) are excluded from the actual refinement process and kept in a test set, of which an independent R-factor,  $R_{\text{free}}$ , is calculated (Brunger, 1992). Comparison of this unbiased value, to the refined R-factor can avoid 'over-refinement'. These two factors should be within 5% of each other for a well-refined structure.

A quality assessment of the final structure can be performed with various programs. Most important is an analysis of the Ramachandran plot, the van der Waals distances and the hydrogen bonding network in the protein.

### Applied Methods

Model building in the COOT program (Emsley and Cowtan 2004) was iterated with refinement using the CNS program-suite (Brunger, Adams et al. 1998) and Refmac (Murshudov, Vagin et al. 1997). Statistics are given in chapter 8.2. Quality of the structure was assessed with programs from the CCP4-suite and programs, which are implemented in COOT.

## 4 Results

### 4.1 Native Purification and *In Vitro*-Translation of the Human Prp19-Complex

Presently not much is known about the overall structure of the human Prp19-complex or the structure of the individual complex components. Only three proteins from the human core complex have also been detected in the *S. cerevisiae* nineteen complex. The interaction between CDC5 and PRL1 is the only described interaction among the human core complex proteins. Moreover, the human Prp19 protein is far less well characterized than its yeast counterpart. Therefore it is unclear, whether a similar interaction network as in *S. cerevisiae* is established among the human proteins. To examine the human Prp19-complex more closely, I followed a dual strategy: the entire complex should be purified from native sources and additionally individual complex proteins should be produced in a recombinant approach. I aimed to use the natively purified Prp19-complex for two purposes: (i) to define a minimal complex for structural analysis using limited proteolysis and (ii) to determine its overall shape using electron microscopy. Results from the native purification approach will be presented in the following paragraphs.

To obtain the human Prp19-complex from native sources, purification from HeLa nuclear extract was performed with a polyclonal antibody against a peptide of CDC5 (amino acids 106–124; the antibody is called anti-CDC5-C1) as described in Makarova *et al.* (Makarova, Makarov *et al.* 2004). In order to scale up the purified amounts, dimethyl pimelidate hydrochloride (DMP)-coupling (Schneider, Newman *et al.* 1982) of the antibody to protein A-sepharose beads and antibody-recovery methods were tested. Moreover, different techniques to concentrate the purified complex were compared. The natively purified human Prp19-complex was subsequently subjected to electron microscopy.

The purified amounts, however, were fairly low: only 14.4 µg of material could be obtained from 6 ml HeLa nuclear extract. This yield is comparable to the experiments performed by Makarova *et al.*. The DMP-coupled antibody purified the complex with a significantly decreased efficiency (~50 %) compared to the uncoupled antibody. In principle the anti-CDC5-C1 antibody could be recovered using a high salt (3.5 M) wash with MgCl<sub>2</sub>, but it lost ~50 % of its affinity in this step. Repeated use of the antibody was therefore not possible. This raised the costs for purification. The best technique to concentrate the complex was pelleting the eluate at 35 krpm for 3.5 h. After re-suspension, the Prp19-complex had a concentration of 1.7 mg/ml. The concentrated complex was analyzed using electron microscopy with and without prior centrifugation on a glycerol gradient and with and without fixation by glutaraldehyde. Generally the electron microscopy negative stain images were

very noisy, because the peptides used for elution of the Prp19-complex from the antibodies tended to stick to the carbon film. In addition to this, many aggregated particles were detected. Whether the aggregation was promoted by the pelleting-procedure is unclear. Taken together, the amounts of natively purified human Prp19-complex were too low to perform conclusive limited proteolysis experiments (~250 µg protein required). As neither the yield nor the quality of the samples for EM seemed improvable in the framework of this project, the native purification approach was abandoned.

In addition to native purification, wheat germ extract was chosen for *in vitro*-translation of the complex proteins. I cloned human Prp19, CDC5 and PRL1 into a suitable vector (pEU3NII) and added an N-terminal strep-tag. The expressions and all described experiments, which involved *in vitro*-translated proteins, were carried out by Dr. D. Agafonov from my department. All three proteins could be expressed individually. The expression levels of CDC5 were 3-4 times lower than for hPrp19 and PRL1. None of the proteins could be purified in significant amounts from the extract. Coexpression of all three proteins as well as pair-wise coexpression was performed, but did not increase protein yield. A sample of hPrp19 bound to CDC5, which has been purified with the anti-CDC5-C1 antibody (see 3.1.4), was analyzed by electron microscopy. The resulting pictures revealed many aggregates and were not suited to obtain a clear image of the two proteins.

## 4.2 Recombinant Production of the Human Prp19-Complex Proteins

As an alternative to native purification and *in vitro*-translation, I aimed to assemble the human Prp19-complex from recombinant sources. This approach permits analyses of the individual Prp19-complex proteins as well as of smaller assembly-intermediates. I primarily focused on the expression in *Escherichia coli*. Various constructs were cloned and tested for expression and protein solubility under a number of different conditions. The commonly tested expression conditions are listed in Table 5. A list of all analyzed human proteins and protein domains is given in Table 8.

An inventory of all proteins and domains, which could be obtained in soluble form, can be found in Table 16. This table also specifies the best expression condition for the respective protein. Details about the conditions, which yielded insoluble proteins, are compiled in Table 15. All results of the recombinant approach will be discussed more detailed in the following chapters.



protein name	length [aa]	amino acid range	MW [kDa]
CDC5	802	1-802	92.3
CDC5_C-term	97	706-802	11.7
Hsp73	646	1-646	70.9
PRL1	514	1-514	57.2
PRL1_WD40	319	173-490	35.5
PRL1_WD40	326	173-498	36.3
PRL1_WD40	305	187-490	34.0
PRL1_WD40	312	187-498	34.8
PRL1_WD40	322	193-514	36.0
PRL1_WD40	296	195-490	32.9
PRL1_WD40	304	195-498	33.7
hPrp19	504	1-504	55.2
hPrp19_cc	81	58-138	9.0
AD002	229	1-229	26.6
Spf27	225	1-225	26.1

**Table 8: Analyzed proteins from the human Prp19-complex**

All analyzed full-length proteins and protein domains from the human Prp19-complex are listed. A more detailed list with several protein parameters, which are relevant for crystallization and a prediction of the crystallization likelihood, can be found in the supplementary section (Table 20). cc means coiled-coil.

#### 4.2.1 The Human Prp19 Protein

A tetramer of the Prp19p protein constitutes the central assembly platform of the nineteen complex in *S. cerevisiae*. A model for the overall architecture of this central protein has been proposed based on the crystal structure of its U-box domain (Vander Kooi, Ohi et al. 2006), electron microscopic negative stain images and various interaction studies (Ohi, Vander Kooi et al. 2005). Less is known about the human Prp19 protein: it has also been detected in super-stoichiometric amounts in the Prp19-complex (Makarova, Makarov et al. 2004), but tetramerization – even though it is likely to occur in the same manner as in yeast – has not been demonstrated yet. Moreover, its interaction with other complex members still awaits investigation. I was interested to see, to which extent the two systems are comparable with respect to Prp19. Therefore I intended to purify the protein and gain insights into its structural properties and its interactions to other proteins from the Prp19-complex.

Under most of the tested expression conditions, however, the protein turned out to be insoluble. Small quantities of soluble GST-tagged hPrp19 (~0.4 mg/l culture volume) could be produced under the conditions listed in Table 16. In all preparations bacterial GroEL

co-purified with the protein. After cleavage of the fusion-tag, hPrp19 and GroEL have a similar molecular weight, which impedes further purification using size exclusion chromatography. Moreover, the two proteins could not be separated by any of the tested column resins (see page 57). Therefore, GST-hPrp19 bound to glutathione-sepharose beads was subjected to repeated washing steps with a buffer containing ATP, MgCl<sub>2</sub> and KCl to remove the chaperone contaminant. Subsequently, the fusion tag was cleaved with PreScission protease and the sample was separated from GST on a Heparin column. Human Prp19 eluted at ~250 mM NaCl in the gradient. The resulting amounts were too low to subject them to a final polishing step (only a few µg protein from 6 l culture volume) or use them for any further analysis.

In conclusion the human GST-hPrp19 fusion-protein could not be produced in *E. coli* successfully in sufficient amounts. In general, hPrp19 may co-purify with the *E. coli* chaperone GroEL due to improper folding. Coexpression with other complex proteins could help to stabilize the hPrp19 protein and therefore it may increase the yields. To produce the protein individually, other expression hosts might be better suited.

#### 4.2.2 The Human Prp19 Coiled-Coil Domain

Tetramerization via the coiled-coil domain is required for the *in vivo* function of *S. cerevisiae* Prp19p. The region between amino acids (aa) 66-141, which comprises the coiled-coil, has been shown to form tetramers (Ohi, Vander Kooi et al. 2005). This region aligns with human Prp19 from residue 60 to residue 135. Therefore a construct was designed, which ranges from amino acid 58 to 138 of human Prp19. Assuming that CDC5 and Spf27 bind to hPrp19 in the same way as their yeast counterparts, the designed construct should also comprise the interaction interface for CDC5 and Spf27 (aa 58-67).

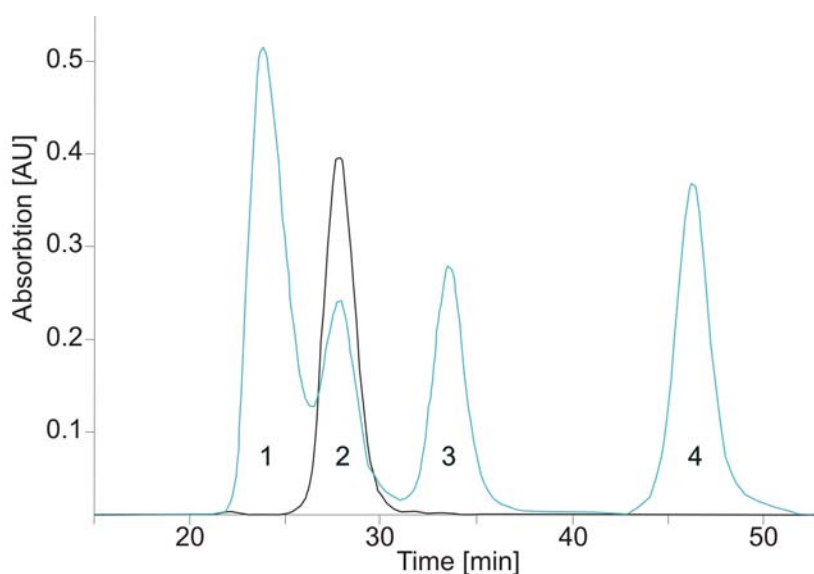
The human Prp19 coiled-coil was soluble with an N-terminal hexa-His-, a TrxA- and a GST-fusion tag. Because the amounts from the hexa-His-tag (His-tag) construct were low and the TrxA-fusion tag was not removable (~10 % cleavage), the purification protocol was determined for GST-hPrp19 coiled-coil. After cleavage of the GST-tag, the protein was subjected to a salt gradient from 50 mM to 1 M NaCl on a CM-sepharose column. Finally, the pure protein was desalted. The average yield of the purifications amounted to ~0.5-1 mg/l culture volume. Remarkably, the protein was most stable in a buffer without salt (10 mM MES pH 6.0, 2 mM DTT). At NaCl concentrations, which exceeded 50 mM, and at protein concentrations >1 mg/ml, precipitation was observed over time. Samples, which contained 50 mM NaCl, could only be concentrated up to 2 mg/ml. Whereas concentrations as high as 24 mg/ml were reached when no salt was present in the buffer.

Analytical gel filtration of hPrp19 coiled-coil was performed on a Superdex 75 column with buffers, which contained different salts (NaCl,  $(\text{NH}_4)_2\text{SO}_4$ ,  $\text{Li}_2\text{SO}_4$ ) or different concentrations of NaCl (50, 100 and 500 mM). Judged by a molecular weight standard, hPrp19 coiled-coil peaked at 44 kDa (Figure 14) and neither the type of salt nor the NaCl concentration made a significant difference. According to the observed precipitation during the purification procedure, a shift of the samples with higher salt towards the column's void-volume or a smeared peak as signs for aggregation would be anticipated. The absence of such effects may be due to low sample concentrations (<1 mg/ml) and relatively short experiment times. In general, a molecular weight of 36 kDa is expected for a hPrp19 coiled-coil tetramer. As the protein is likely very elongated, it might elute at slightly higher molecular weight than anticipated for a globular protein. Therefore it is assumed that hPrp19 coiled-coil forms a homo-tetramer under the tested conditions. These findings indicate that human Prp19 forms a homo-tetramer like its yeast counterpart.

Circular dichroism was carried out with hPrp19 coiled-coil samples, which contained either no salt or 50 mM NaCl, at protein concentrations of 1 mg/ml (for a detailed description of the method see 3.2.5). Both CD-spectra are typical for predominantly  $\alpha$ -helical secondary structure (data not shown). A spectrum of *S. cerevisiae* Prp19p coiled-coil (aa 66-141) looks almost identical and corresponds to 70 % helical content and 30 % coil-structure (Ohi, Vander Kooi et al. 2005). Protein stability was examined in thermal unfolding. The resulting curves did not show a perfect sigmoidal shape as expected for a cooperative unfolding event. The rather continuous slope at the beginning of the temperature gradient may suggest a slight *a priori* instability, but still an extensive signal change (~70 mdeg) with a moderate slope was recorded around the melting point (data not shown). Both samples showed comparable stability, as their melting points ( $T_m$ ) were at ~63 °C.

The protein seems to be functional, since GST-hPrp19 coiled-coil could effectively pull down *in vitro*-translated CDC5. The experiment will be described in more detail below (see chapter 4.2.5).

Initial crystallization trials with 200 conditions were set up with hPrp19 coiled-coil in 10 mM MES pH 6.0, 2 mM DTT at a protein concentration of 10 mg/ml. Another 300 conditions were screened with 6 mg/ml protein concentration. Some quasi crystalline blocks and tiny needles appeared after one day. The needles were not singular, but rather clusters of needles, which were grown together (Figure 15). In a first refinement step, different additives were tested on these conditions. The quasi crystalline blocks could not be improved, whereas the needle bundles could be improved towards more single needles. However, their size did not increase significantly. Lower protein concentrations (even as low as 1 mg/ml) or macro-seeding did not lead to any reduction of nucleation or size gain either.

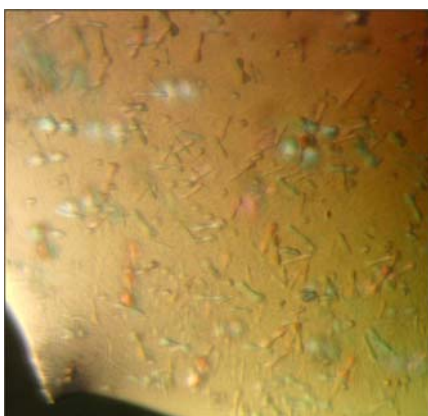


**Figure 14: Human Prp19 Coiled-Coil Analytical Gel Filtration**

The elution profile of hPrp19 coiled-coil (aa 58-138) is shown in black. A column MW-standard is depicted in blue. Peak 1: > 158 kDa proteins in void-volume, Peak 2: 44 kDa, Peak 3: 17 kDa, Peak 4: 1.4 kDa. The running buffer contained 50 mM NaCl.

Staining of the improved needle-shaped crystals with a protein staining dye revealed that they are protein crystals. However, when the crystals were cryo-protected with 30 % glycerol and tested in an X-ray beam, no diffraction was observable. Likely, this was due to the very small size of the crystals.

In summary, the hPrp19 coiled-coil protein seems to be tetrameric. This demonstrated directly that the human Prp19 protein forms homo-oligomers in the same way as its yeast counterpart. Very small needles could be obtained in initial crystallization screening, but they could not be improved to diffraction quality crystals. These needles diffracted poorly in the X-ray beam. A reason for the cessation of crystal growth and the slight protein instability, which was observed in CD, could be the construct length: the N-terminus may be too long compared to the predicted coiled-coil and therefore flexible amino acids could potentially interfere with crystal growth. Future NMR experiments may clarify this point.



**Figure 15: Human Prp19 Coiled-Coil Initial Crystals**

Initial crystals of hPrp19 coiled-coil (aa 58-138), which were grown in Hampton Index screen, condition 13 (0.1 M BIS-TRIS pH 5.5, 0.3 M Magnesium formate) are shown. The crystals form clusters of needles, which are grown together in the middle.

### 4.2.3 The CDC5 C-Terminal Domain and the PRL1 WD40 Domain

The interaction of CDC5 and PRL1 is essential for pre-mRNA splicing (Ajuh, Sleeman et al. 2001). The minimal interaction interface has been mapped to the carboxy-terminus of CDC5 (aa 706-800) and parts of the WD40-domain in PRL1 (aa 257-396). The entire interface may be more extended and involve multiple contacts. Presently, the exact nature of the interaction is unknown. The carboxy-terminus of CDC5, however, was predicted to be unfolded. Therefore it was speculated that folding could be induced upon interaction with the PRL1 WD40-domain. I was interested to reconstitute this interaction and investigate its structural basis. For this purpose the interacting domains from both proteins were thought to be either co-purified after individual expression or the proteins should be coexpressed and purified.

I cloned the carboxy-terminal residues 706-802 of CDC5 as N-terminal His-, ZZ- and GST-fusion proteins. Soluble protein with a cleavable tag could be obtained from all constructs. Mixing this tagged protein with purified PRL1 WD40-domain can potentially induce structural order and enable the co-purification and analysis of the two proteins in complex.

The minimal region of PRL1, which was found to interact with CDC5 (aa 257-396), starts in the middle of the second predicted WD-motif and ends close to the C-terminus of the fifth WD-motif. Since the seven WD-motifs together form the  $\beta$ -propeller-shaped WD40-domain in PRL1, I thought that disrupting this entity may also disturb its function. Therefore I designed a construct (aa 193-514), which covers the entire WD40-domain (predicted aa 193-490) and the remaining C-terminal residues. An N-terminal His-tagged fusion of this construct was expressed in good amounts, but only very little of the expressed protein turned out to be soluble, which prevented its use in co-purifications. To increase the amounts of soluble protein a set of six different PRL1 WD40-constructs was designed. The construct-design was based on alignments of known WD40-domain crystal structures (PDB-entries: 1ERJ, 1NEX, 1VYH, 1P22) and secondary structure predictions. All of these C-terminal His-tag fusion proteins were expressed, but none of them could be obtained in a soluble form. In summary, too little soluble PRL1 WD40-domain was obtained to mix and co-purify it with the CDC5 C-terminal domain. Coexpression trials might improve the solubility of the PRL1 WD40-domain. These coexpressions are discussed in chapter 4.2.5.

### 4.2.4 Other Proteins from the Human Prp19-Complex

The remaining proteins from the Prp19-complex were tested for their expression and solubility. A summary of the best expression conditions for each protein is shown in Table 16.

CDC5 could be expressed as an N-terminal His-tagged fusion protein (in addition to a GST-fusion and a NusA-fusion), which indicates that the protein is not artificially solubilized by the fusion tag. However, the protein was massively degraded under all conditions. Most probably, the degradation occurred from the C-terminus, which is why C-terminal His-tag fusion constructs were designed. These constructs did not express the protein, though.

Similar to hPrp19, PRL1 co-purified with chaperones. Additionally, the expression level and the amount of soluble protein were low. In accordance with these findings, myosin-tagged PRL1, which was expressed in the amoeba *Dictyostelium discoideum* (Howard, Ahern et al. 1988), was not produced in detectable amounts. The expression trials in *D. discoideum*, for which I provided the constructs, were carried out by M. Kollmar from the NMR-department at our institute.

Hsp73 was soluble and the GST-tag could be removed, but as the structure of the highly homologous bovine Hsp70 protein was already known, Hsp73 was not analyzed any further. AD002 was only soluble to a very little extent (~10 %). Despite fairly high expression levels, soluble Spf27 could not be obtained at all.

#### 4.2.5 Interactions among the Prp19-Complex Proteins

One way to obtain interacting proteins is to coexpress them. Compared to expression of the individual proteins, coexpression of interacting proteins can increase the expression level and enhance the solubility. Additionally, coexpressing proteins may protect each other against degradation. Different combinations of the human Prp19-complex proteins, which were known or suspected to interact, were tested in coexpression trials. The protein combinations are listed in Table 9.

Remarkably, in none of the cases expression or solubility was improved. The results rather indicated the opposite tendency: some proteins, which were soluble at least to a little extent in individual expressions (GST-PRL1, GST-hPrp19), could in coexpressions only be purified in smaller amounts (see GST-hPrp19/His-CDC5-expression) or not at all (see GST-PRL1/His-CDC5-C-term-expression). In almost all cases only one protein was soluble. For pGEX6p\_1-CDC5/pET-M11-PRL1 the degradation pattern of soluble GST-CDC5 looked identical to the one found in individual expressions. This means that CDC5 was not protected against degradation by concomitant expression of PRL1. Positive coexpression controls were successful, though. First coexpression trials using a vector series, which is designed specifically for coexpressions (the duet-vector series; see Table 4), did not succeed either.

Besides the proteins, which are listed in Table 9, AD002 was subjected to initial coexpression trials with all other Prp19-complex members. But as AD002 coexpression was not

exhaustively tested, results from these tests can only be taken as an indication. However, the tendency remains the same as in the other coexpressions: in each case some soluble AD002 was obtained, but none of the co-transformed proteins was expressed or purified.

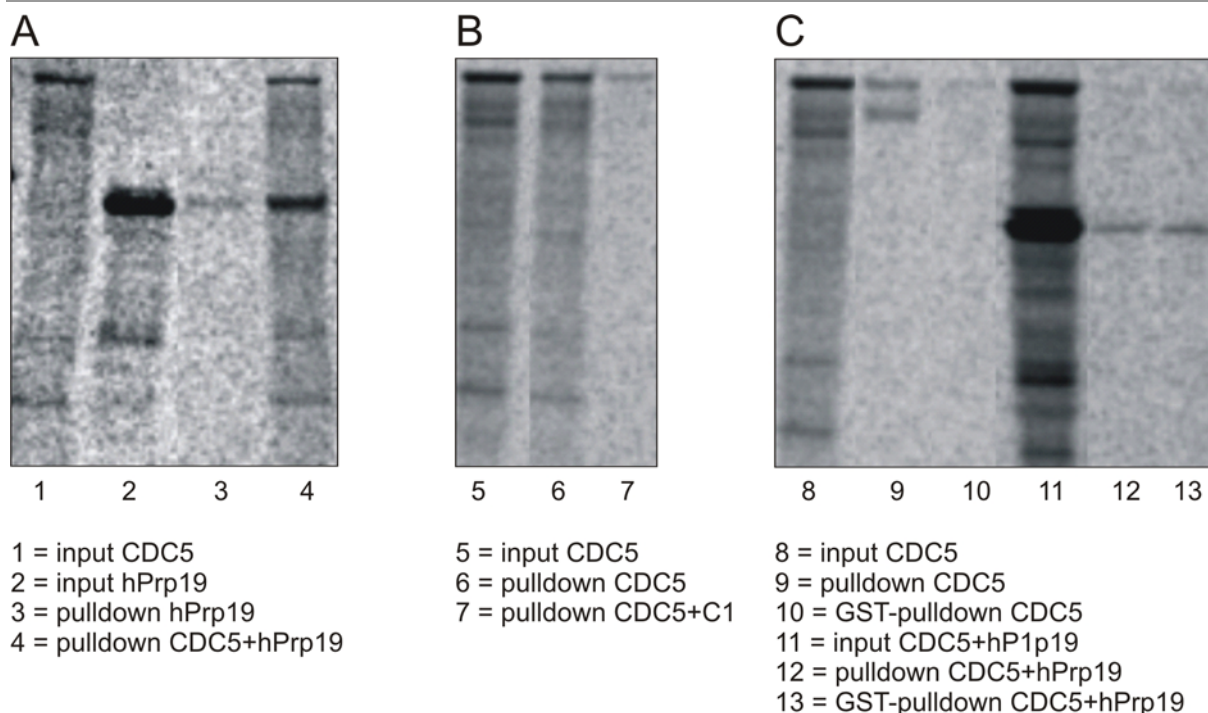
protein 1	vector	fusion tag	protein 2	vector	fusion tag	result
PRL1	pET-M11	N-His	CDC5	pGEX6p_1	N-GST	some soluble protein 2
PRL1	pGEX6p_1	N-GST	CDC5	pET-M11	N-His	some soluble protein 1
PRL1	pGEX6p_1	N-GST	CDC5_C-term	pET-M11	N-His	soluble protein 2
PRL1	pET-M11	N-His	CDC5_C-term	pGEX6p_1	N-GST	soluble protein 2
PRL1	pGEX6p_1	N-GST	CDC5_C-term	pET-MZZ	N-ZZ	soluble protein 2
PRL1 WD40 <sup>93-514</sup>	pGEX6p_1	N-GST	CDC5_C-term	pET-M11	N-His	soluble protein 2
hPrp19	pGEX6p_1	N-GST	CDC5	pET-M11	N-His	soluble protein 2
hPrp19cc <sup>58-138</sup>	pCDF-duet	N-His	CDC5	pET-duet	none	soluble protein 1

**Table 9: Coexpression of Human Prp19-Complex Proteins**

The table lists the human Prp19-complex proteins, their vectors and the fusion tags, which were coexpressed. Additionally the result of the respective coexpression is commented.

Some proteins that are difficult to produce in *E. coli*, may more readily be expressed using a cell free system from a higher organism. These *in vitro*-translation systems also offer the opportunity to easily coexpress proteins. Subsequently, co-purification or co-immunoprecipitation may be employed to study the interactions among the proteins.

In *S. cerevisiae* Cef1p interacts with the Prp19p coiled-coil domain. We obtained information about the corresponding interaction between the human proteins by *in vitro*-translation in wheat germ extract and recombinant protein production in *E. coli*. In cooperation with Dr. D. Agafonov from my department, the following findings were made: *in vitro*-translated CDC5 and hPrp19 could effectively be pulled-down by anti-CDC5-C1 antibody, which indicated that they interact (Figure 16 A). This interaction was mediated by the hPrp19 coiled-coil domain, because GST-hPrp19 coiled-coil, which was produced recombinantly in *E. coli*, could pull-down the *in vitro*-translated CDC5 (Figure 16 C). The pull-down was suppressed by the addition of *in vitro*-translated hPrp19. Therefore the interaction seems to be specific and it is mediated by residues between amino acid 58-138 of the hPrp19 protein. These experiments showed for the first time that the interaction of hPrp19 and CDC5 is also present in the human Prp19-complex. For these experiments I provided the GST-hPrp19 coiled-coil (aa 58-138) and the clones for *in vitro*-translations, whereas the *in vitro*-translation experiments and the pull-down assays were performed by Dr. D. Agafonov.



**Figure 16: *In Vitro*-Translations and Pulldown Assays of Human Prp19 and CDC5**

**A** In vitro-translated hPrp19 and CDC5 pulled down by anti-CDC5-C1 antibody. Both CDC5 and hPrp19 are pulled down by the anti-CDC5-C1 antibody (lane 4). Therefore hPrp19 interacts with CDC5. The efficiency of the hPrp19-pulldown was estimated to be ~4 fold of the efficiency of the CDC5 pulldown, which supports the idea of hPrp19 tetramerization. **B** In vitro-translated CDC5 pulled down by the anti-CDC5-C1 antibody. The pulldown efficiency was approximated to be ~50 % of the total in vitro-translated (lane 7). **C** In vitro-translated hPrp19 and CDC5 pulled down by GST-hPrp19 coiled-coil (aa 58-138). Around 20 % of the in vitro-translated CDC5 (and a truncated fragment) are pulled down by GST-hPrp19 coiled-coil (lane 9). This pulldown is blocked by the presence of in vitro-translated hPrp19 (lane 13). Pulldowns with GST only are shown in lane 10 and 13. The experiments were carried out by Dr. D. Agafonov. The clones and the GST-hPrp19 coiled-coil protein were provided by me.

### 4.3 Proteins from the *Saccharomyces cerevisiae* Prp19p-Complex

The nineteen complex from *S. cerevisiae* consists of at least 30 proteins, of which around 11 have been identified as core proteins (Tsai, Chow et al. 1999; Chen, Yu et al. 2002; Ohi, Link et al. 2002). All interaction models of the NTC place a Prp19p tetramer as a central assembly platform at the core of the complex (Tarn, Hsu et al. 1994; Chen, Tsao et al. 1999; Chen, Yu et al. 2002; Ohi and Gould 2002). Based on this knowledge, I aimed to study the



structure and the interactions of the core nineteen complex proteins: Prp19p and proteins, which directly interact with it. The major focus was laid on Prp19p itself.

A list of all analyzed *Saccharomyces cerevisiae* full-length proteins and protein domains is given in Table 10. Since the human proteins were difficult to produce in *E. coli*, I hoped that their yeast counterparts would be better suited for recombinant production in this expression host. As for the human proteins, expression was tested under a range of conditions (see Table 5). The soluble proteins and the best conditions to express them are specified in Table 16, whereas the conditions leading to insoluble proteins can be found in Table 15.

Protein name	length [aa]	amino acid range	MW [kDa]
Cef1p	590	1-590	67.7
Prp46p	451	1-451	50.7
Prp19p	503	1-503	56.6
Prp19p_Ubox-cc	132	1-132	14.7
Prp19p_Ubox-cc	141	1-141	15.7
Prp19p_Ubox-cc	149	1-149	16.6
Snt309p	175	1-175	20.7

**Table 10: Analyzed Proteins from the *S. cerevisiae* Prp19p-complex**

Analyzed proteins and protein domains from the Prp19p-complex are listed. A more detailed list with several protein parameters relevant for crystallization can be found in the supplementary section (see Table 20). cc means coiled-coil.

#### 4.3.1 The Prp19p Protein

The *S. cerevisiae* Prp19p protein is the most intensively studied component of the nineteen complex. It comprises an N-terminal U-box domain, a central coiled-coil and a C-terminal WD40 domain (Figure 5). The protein has been shown to tetramerize via its coiled-coil domain (Ohi and Gould 2002; Ohi, Vander Kooi et al. 2005). A crystal structure of the dimeric U-box domain has been determined at 1.5 Å resolution (Vander Kooi, Ohi et al. 2006) and electron microscopy negative stain images provided first insights into its domain architecture (Ohi, Vander Kooi et al. 2005). Nevertheless, many questions remain unanswered. It is unknown how the U-box domains are attached to the coiled-coil and whether there is any interplay between the U-box and the WD40-domains in splicing or E3-ligase function. Additionally, it remains to be investigated whether there are different Prp19p conformations, probably even induced by binding partners like Snt309p or by modifications like ubiquitination, which may commit Prp19p to exert different functions.

To answer these questions I expressed Prp19p as an N-terminal His- and GST-fusion protein. His-Prp19p was only soluble to a little extent and it was co-purifying with a number of other proteins. The GST-fusion protein however could be obtained in large amounts and it could be purified close to homogeneity. After removal of the GST-tag, the protein was purified on a MonoQ-column (gradient from 100 mM to 1 M NaCl) and polished in size exclusion chromatography on a Superdex 200-column (running buffer: 20mM Tris pH 8.0, 100mM NaCl, 2mM DTT). Comparison of Prp19p to size exclusion standards suggested that the protein is tetrameric (data not shown). The total yield of the Prp19p purifications amounted to 4-5 mg protein/l culture volume.

Analytical gel filtration was performed at different salt concentrations (between 100-800 mM NaCl). In all cases the main protein peak could be located between the 670 kDa- and the 158 kDa-peak of a molecular weight standard, which was used for column calibration (BioRad-standard; see chapter 3.1.12). Prp19p was eluting more closely to the 158 kDa-peak. A tetrameric oligomerization-state of Prp19p (226.4 kDa) is therefore supported by analytical gel filtration. The salt concentration did not have any effect on the oligomerization-state. Thus, the Prp19p tetramer is not disrupted even by very high salt concentrations.

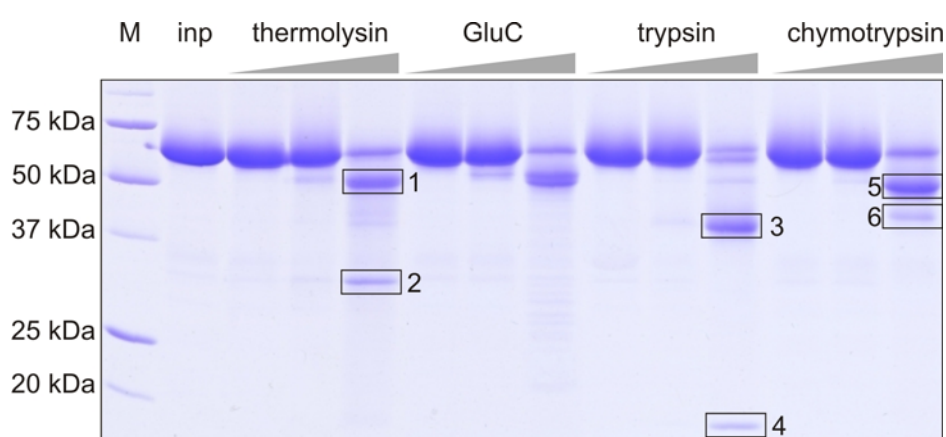
A mixed  $\alpha/\beta$  secondary structure was detected in circular dichroism. The thermal unfolding, however, was ambiguous (data not shown).

Limited proteolysis with different proteases more or less confirmed the domain borders, which have been defined by Ohi et al. using truncated Prp19p-constructs (Ohi, Vander Kooi et al. 2005). Stable fragments were obtained with chymotrypsin, trypsin, endo-protease Glu-C and thermolysin (Figure 17). The results of the mass spectrometric analysis of the fragments are given in Table 11. The fragments from the digestion with endo-protease Glu-C were not analyzed, as the double bands were too close together. Band 2 was not clearly attributable to a certain domain. In summary, the C-terminus of the coiled-coil domain and the WD40-domain could clearly be mapped. They are in well agreement with the previous results by Ohi et al.. The coiled-coil domain is only a bit shorter according to the proteolysis assay (up to aa ~130) compared to residues 1-141 as predicted by Ohi *et al.*. Other domain borders could not be defined in the assay.

band number	protease	sequence coverage [aa]	fragment MW [kDa]	mapped protein region
1	thermolysin	1-423	47.7	up to end of WD40
3	trypsin	131-448	36.6	end of coiled-coil to end of WD40
4	trypsin	1-139	15.5	up to end of coiled-coil
5	chymotrypsin	1-426	48.2	up to end of WD40
6	chymotrypsin	128-504	42.6	end of coiled-coil to C-terminus

**Table 11: Mass Spectrometric Results of Prp19p Limited Proteolysis**

Residues ranges (sequence coverage) that were identified in individual bands from the gel in Figure 17 are listed. The used protease, the calculated mass of the fragment and a description, which region of the protein was mapped, is listed as well.



**Figure 17: Limited Proteolysis of Prp19p**

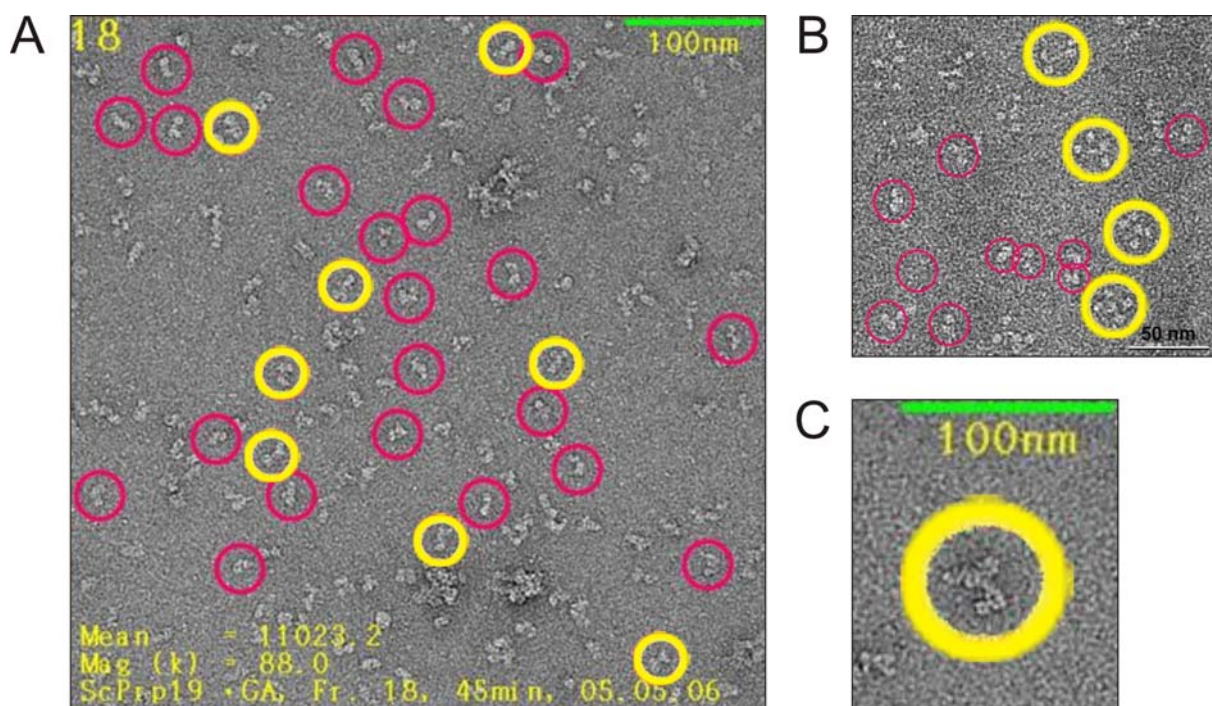
Polyacrylamide gel of the limited proteolysis of Prp19p. The used proteases are mentioned above the grey triangles, which indicate decreasing amounts of the applied protease: protease/protein-ratios (mg/mg) drop from 1/10 in the first lane, to 1/100 and 1/1000 in the second and third lane. Numbered rectangles highlight the bands, which were analyzed by mass spectrometry.

Electron microscopic analysis of Prp19p was performed by Dr. M. Golas and Dr. B. Sander. The best results were achieved after separation of the sample on a 5-20 % glycerol gradient with glutaraldehyde fixation (Kastner, Fischer et al. 2008). The protein peaked in the fractions around number 18, which was calculated to roughly correspond to a sedimentation coefficient of around 6S. Negative stain images clearly revealed the circular WD40-domains (Figure 18 C). The major portion of the particles appeared to be dimers and not tetramers as expected (Figure 18 A). Only some particles showed a clear tetrameric shape (Figure 18 C). Interestingly, a closer analysis of the negative stain images (performed by Dr. M. Golas), which have been published by Ohi *et al.*, resulted in the same discovery: more dimeric than tetrameric particles are visible in their images as well (Figure 18 B), but only the tetrameric

particles were integrated to yield the final image. Since other lines of evidence suggest that Prp19p forms tetramers, the observed Prp19p dimers are most likely only artifacts, which occur during the electron microscopic sample preparation.

The Prp19p protein was subjected to initial crystallization trials at a concentration of 12 mg/ml. Around 1500 conditions were screened at room temperature and at 4 °C. Different kinds of precipitate and some quasi-crystalline aggregates, but no crystals appeared over time.

In conclusion, Prp19p was found to tetramerize according to its behavior in size exclusion chromatography. The predominant dimeric form, which was found in our and in Ohi's electron microscopic studies, can be regarded as an artifact. Limited proteolysis revealed similar domain boundaries as already proposed by Ohi *et al.*. Only the Prp19p coiled-coil domain was mapped slightly shorter on the C-terminus. Finally, crystallization trials were not successful.



**Figure 18: Electron Microscopy Negative Stain Images of Prp19p**

**A** Prp19p EM negative stain images. Fraction 18 of a 5-20 % glycerol gradient with glutaraldehyde fixation was absorbed 45 min on the carbon film before taking the image. The protein comes from my preparations. **B** Negative stain images of Prp19p as published in Ohi *et al.* **C** Enlarged tetrameric Prp19p particle from my preparations. The circular WD40-domains are clearly visible. Yellow circles mark the particles, which were identified as tetramers of Prp19p. Pink circles mark the homo-dimers.

### 4.3.2 The Prp19p U-box and Coiled-Coil Domain

Exclusion of the flexible linker between the coiled-coil and the WD40 domain may improve the chances for crystallization. Therefore, constructs were designed, which only include the U-box and the coiled-coil domain. Taking the proteolysis results as a basis, the first construct terminates around amino acid 130 (aa 1-132). The second construct spans the domains as defined by Ohi *et al.* (aa 1-141) and the last construct extends to aa 1-149 (see Table 10). All three constructs were obtained soluble in high amounts under the expression conditions listed in Table 16.

After cleavage, the GST-tags were removed in glutathione-sepharose affinity purification. Final size exclusion chromatography steps on a Superdex 200-column yielded the pure proteins (running buffer: 20 mM Tris pH 8.0, 150mM NaCl, 2mM DTT). The main protein peaks corresponded to molecular weights between 50 kDa and 90 kDa (according to EMBL column calibration standards; for a reference see <http://www.pepcore.embl.de>). Approximately 1-2 mg protein/l culture volume could be obtained for purifications of the long (aa 1-149) and the short (aa 1-132) protein. The protein of intermediate length (aa 1-141) yielded over 10 mg protein/l culture volume.

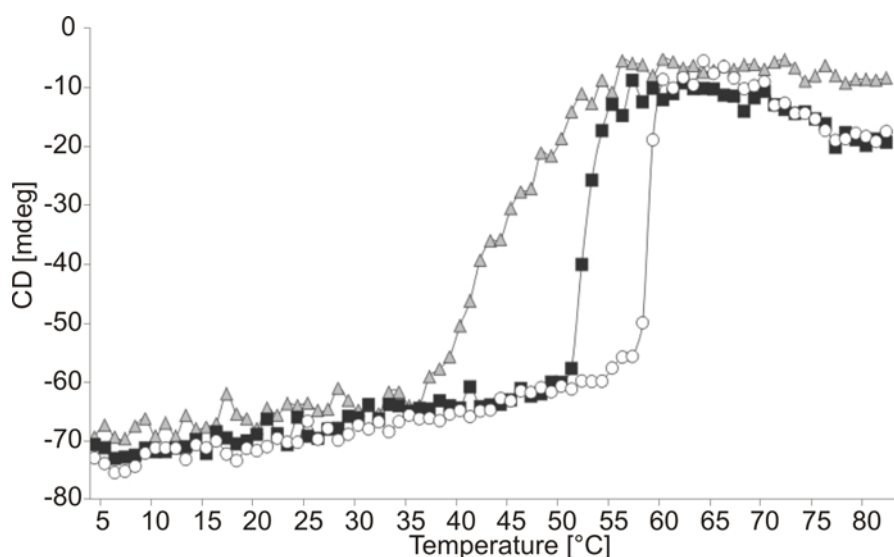
Analytical gel filtration runs with the individual proteins on a Superdex 200-column revealed that all three proteins are running between 158 kDa and 44 kDa – more closely to the smaller molecular weight peak. The elution behavior corresponds well to a tetramer (~60 kDa) and certainly excludes a dimer. The three protein peaks appear in the chromatogram in the order of their different sizes (data not shown). Notably, even though the minimal region, which is required for tetramerization, has been mapped to residues 66-141 by Ohi *et al.*, the shortest Prp19p protein (aa 1-132) still seems to be tetrameric. In addition, this protein was subjected to different NaCl concentrations between 100 mM and 1.5 M to test the stability of the oligomerization. Interestingly, the salt concentration did not change the protein migration in analytical gel filtration runs, which indicates high stability of the tetramer.

All three proteins appeared to consist mainly of  $\alpha$ -helices judged by their circular dichroism spectra. The spectra were recorded at protein concentrations of 0.5 mg/ml. Thermal melting showed cooperative unfolding with a perfect sigmoidal shape for the two larger proteins (Figure 19). The shorter Prp19p protein (aa 1-132) protein unfolded not as cooperatively as the other ones, but still yielded a sigmoidal curve. Thermal stability was highest for the protein of intermediate length (aa 1-141) with a  $T_m$  of ~58 °C.

Initial screening of around 1500 crystallization conditions was performed with all three proteins at 4 °C. Thousand conditions were tested in parallel at 21 °C with the shortest protein. A difference between the temperatures was not observable. Even though protein

concentrations between 7 mg/ml and 21 mg/ml were used, most of the drops remained clear. In the other drops, mainly phase separation occurred – no crystals were obtained. Increasing the salt concentration up to 300 mM NaCl in the protein sample had no effect on crystallization.

Taken together, the designed proteins, which comprise the U-box and the coiled-coil domain, form homo-tetramers. Also the shortest of these proteins (aa 1-132) formed a tetramer. Ohi *et al.* showed that neither the U-box nor the WD40 domain, but only the coiled-coil region between residues 66-141, contributes to tetramerization of Prp19p. Combining these results with my results, the residues, which are necessary for tetramerization, can be narrowed down to amino acids 66-132. Additionally, it was found that this oligomerization is salt-resistant.



**Figure 19: Prp19p U-box and Coiled-Coil Melting Curves**

CD-signal during thermal unfolding plotted against temperature. Rectangles are data points for the short protein (aa 1-132), squares for the intermediate (aa 1-141) and circles for the long protein (1-149).

#### 4.3.3 Other Proteins from the *Saccharomyces cerevisiae* Prp19p-Complex

In *S. cerevisiae*, Cef1p and Snt309p are interaction partners of Prp19p. I was interested to examine these proteins for their ability to interact with the shortened Prp19p coiled-coil constructs. Moreover, the Cef1p-Prp46p contact should be investigated.

Expression conditions for Cef1p, Prp46p and Snt309p were tested as described in chapter 3.2.2.. pET-duet and pCDF-duet served as expression vectors and different *E. coli* strains were employed as expression hosts. Nevertheless, neither of the proteins showed a reasonable expression level and no soluble protein could be obtained. Coexpression of Cef1p and Prp46p did not improve their expression level. Possible reasons and ways to circumvent these problems will be discussed in chapter 5.1.

#### 4.4 Studies on the U11/U12 65K Protein

The U11/U12 65K protein serves as a bridging factor in the minor spliceosome: it binds the U11-specific 59K protein via its N-terminal half and a stem loop structure in the U12 snRNA with its carboxy-terminal RNA recognition motif (65K-cRRM). This C-terminal RRM is highly homologous to the N-terminal RRM of U1A and U2B<sup>''</sup> (U1A-nRRM and U2B<sup>''</sup>-nRRM). Additionally, the recognized RNA-targets, namely the U12 snRNA hairpin III, the U1 snRNA hairpin II and the U2 snRNA hairpin IV, resemble each other (Figure 11). U1A- and U2B<sup>''</sup>-nRRM both bind their RNA target on the surface of the central  $\beta$ -sheet and mediate the sequence-specific interactions mainly with residues from loop 3 and the variable region (Figure 12). Recently, a number of alternative modes to bind RNA and to regulate the binding have been demonstrated in various RRM (see chapter 2.4.4). I was interested to know, whether the U11/U12 65K C-terminal RRM applies the same mode of RNA-binding as its homologs U1A and U2B<sup>''</sup> and whether the binding is regulated in any way. To answer this question, I studied the RNA-binding of 65K-constructs, which comprised the core RRM-domain, in gel shift assays.

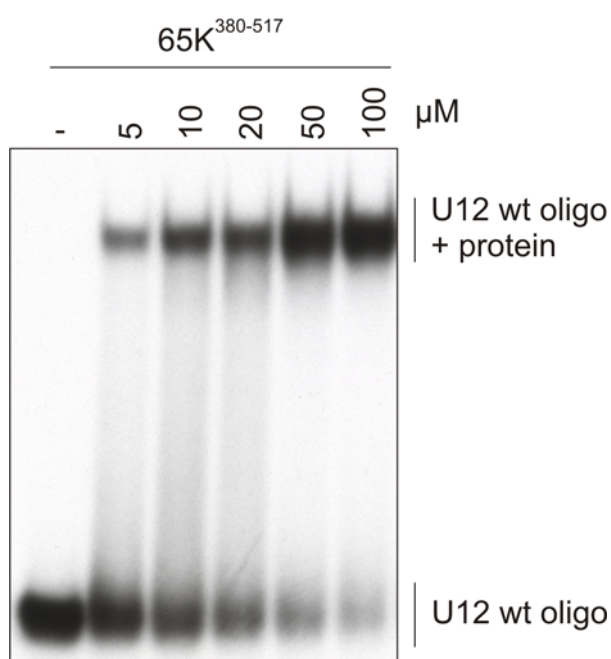
##### 4.4.1 RNA-Binding of Truncated 65K Proteins

In order to compare the RNA-binding of U11/U12 65K-cRRM to U1A and U2B<sup>''</sup>, I intended to design a construct, which comprised the core RNA recognition motif as it is found in U1A and U2B<sup>''</sup>. In the crystal structure of U1A in complex with its cognate RNA, residues 5-85 formed a compact core domain (Oubridge, Ito et al. 1994). This residue range corresponds to amino acids 415-500 in 65K. The visible residues in the U1A and U2B<sup>''</sup> structure, however, comprise a few additional residues N- and C-terminally. Five constructs of U11/U12 65K were designed around this core RRM domain. Since the C-terminal part of U11/U12 65K (aa 380-517) is known to bind to nucleotides 109-125 of U12 snRNA hairpin III (Benecke, Luhrmann et al. 2005), this part of the protein was cloned as well in order to have a positive control for RNA-binding. Whenever experiments with this construct were performed, the construct will be mentioned as 65K<sup>380-517</sup>. All six investigated domains and details about their sequences are summarized in Table 12. Details on the expression conditions and the purification procedure can be found elsewhere (see page 58 and Table 16).

protein name	length [aa]	amino acid range	MW [kDa]	comment
65K <sup>380-517</sup>	138	380-517	16.0	positive control for RNA-binding
65K <sup>404-505</sup>	102	404-505	11.8	long construct around core RRM
65K <sup>411-505</sup>	95	411-505	10.9	middle construct around core RRM
65K <sup>417-505</sup>	89	417-505	10.3	short construct around core RRM
65K <sup>380-506</sup>	127	380-506	14.7	long C-terminal truncation of positive control
65K <sup>380-501</sup>	122	380-501	14.1	short C-terminal truncation of positive control

**Table 12: U11/U12 65K Domains**

Analyzed U11/U12 65K domains are listed. A more detailed list, which provides information on some sequence features, can be found in the supplementary section (see Table 20).



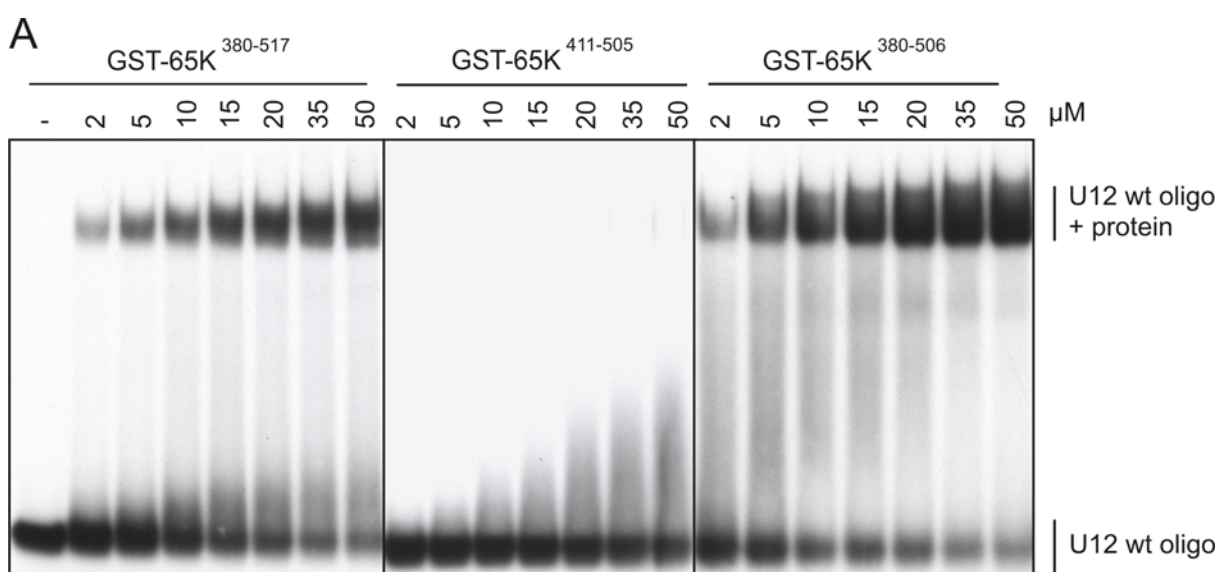
**Figure 20: Gel Shift Assay with 65K<sup>380-517</sup> and U12-wt RNA**

The gel shift assay of the untagged 65K<sup>380-517</sup> protein and the U12-wt RNA is shown. The assay was performed as described in chapter.3.2.4. The apparent  $K_D$  was  $\sim 11 \mu\text{M}$ .

To measure the RNA-affinity of the constructs, gel shift assays with an oligonucleotide comprising the bases 109-125 of U12 snRNA hairpin III (called U12-wt oligo or U12-wt RNA) were performed (Figure 20). The apparent  $K_{Dapp}$  of  $11.4 \mu\text{M}$  of the 65K<sup>380-517</sup> construct was comparable to the  $K_{Dapp}$  of  $\sim 10 \mu\text{M}$ , which has been reported previously (Benecke, Luhrmann et al. 2005). For the 65K<sup>380-517</sup> N-terminal GST-fusion, binding was slightly worse ( $K_{Dapp} = 16.2 \mu\text{M}$ ). Even though both binding constants are in the usual range for RRM containing proteins (*i. e.*,  $10^{-11}$  to  $10^{-6}$  M; (Varani and Nagai 1998)), they are significantly higher than that of U1A binding to the U1 snRNA hairpin II ( $K_D \sim 32 \text{ pM}$ ; (Katsamba, Myszka et al. 2001)).



Of the three constructs, which range up to aa 505, only 65K<sup>411-505</sup> could be purified. Gel shift assays were performed with the untagged and the GST-tagged fusion protein. Surprisingly, for this 65K construct, which comprises the entire C-terminal RRM core domain, almost no binding to the U12-wt oligo was detectable (Figure 21). This finding indicates that 65K<sup>411-505</sup> lacks a part of the protein, which is necessary for RNA-binding. The 65K<sup>380-506</sup> protein, in contrast, bound the RNA with ~2 times higher affinity ( $K_{Dapp} = 8.6 \mu\text{M}$ ) than the positive control protein 65K<sup>380-517</sup>. In conclusion, the N-terminus between residues 380-410 is required for RNA-binding. The question, which role these N-terminal residues may play in 65K RNA-binding, will be discussed later in more detail (see chapter 4.4.5).



**Figure 21: Gel Shift Assays of Truncated 65K Proteins**

The gel shift assays of the truncated 65K proteins with U12-wt RNA are shown. **A** The shifts were performed with different concentrations of RNA oligo, which are listed below the black bar. Above the black bar the protein name is given. The  $K_{Dapp}$  of GST-65K<sup>380-517</sup> was calculated to be 16.2  $\mu\text{M}$ , the one of GST-65K<sup>380-506</sup> was 8.6  $\mu\text{M}$ . **B** Structure of 65K<sup>380-517</sup> highlighting the N-terminal expansion that was deleted in GST-65K<sup>411-505</sup> in green.

A further truncation of the C-terminus, represented by 65K<sup>380-501</sup>, leads to a complete loss of RNA-affinity. This finding is in accordance with U1A, where truncation of the C-terminus directly after the last  $\beta$ -sheet, drastically diminishes RNA binding (Boelens, Scherly et al. 1991; Zeng and Hall 1997).

GST-65K<sup>380-517</sup> has formerly been shown to exhibit a 20 fold higher affinity to full-length U12 snRNA than to the hairpin III of U12 snRNA only (Benecke, Luhrmann et al. 2005), suggesting that residues apart from hairpin III contribute to affinity. Therefore I asked, whether RNA-oligonucleotides, which are altered in their length and their terminal sequence, may show a better affinity to 65K<sup>380-517</sup> than the U12-wt RNA. Sequences of the designed RNA-oligonucleotides are listed in Table 3. All RNA-oligonucleotides showed a similar affinity to 65K<sup>380-517</sup> in gel shift assays as U12-wt. An improved binding of a hairpin structure with an elongated stem containing 10 instead of only 5 stem base pairs could not be detected. This implies that additional stem nucleotides, which directly neighbor the U12-wt oligo form hairpin III, do not contribute significantly to the RNA-affinity. The higher affinity of GST-65K<sup>380-517</sup> observed for full-length U12 snRNA must therefore be mediated by nucleotides more distant from the hairpin III. Moreover, this finding suggests that RNA-binding via the stem nucleotides as observed for the RBMY (Skrisovska, Bourgeois et al. 2007) protein is rather unlikely to occur in 65K.

#### 4.4.2 Crystallization of U11/U12 65K<sup>380-517</sup>

The 65K<sup>380-517</sup> protein was crystallized under a number of conditions and with all designed RNA-oligonucleotides. The details of sample preparation are described elsewhere (page 58). In most cases, crystallization of the 65K-RNA-complexes was performed with and without a preceding gel filtration step. The gel filtration was intended to remove surplus RNA and potentially aggregated particles, which may have formed upon mixing of protein and RNA. All gel filtration runs showed a clear shift of the protein peak in the presence of RNA (data not shown), indicating that the RNA was bound to the protein. A summary of the initial crystallization trials is given in Table 13.

Crystals appeared after one to two days in several conditions, which all contained polyethylene glycol (PEG) of low to intermediate molecular weight (PEG 400-4000). In most cases 0.2 M salt (MgCl<sub>2</sub>, KCl or LiCl or 0.2 M Na-acetate) were present in these conditions and the pH of most conditions was close to neutral. The crystals, which formed under these conditions, were bundles of very fine needles (Figure 22 A). Refinement of initial crystallization hits was performed, additives were used and different temperatures were tested, but the thickness of the needles could not be improved significantly (Figure 22 B). Single needles, which were detached from these bundles, were cryo-protected in 15% propylene glycol and a dataset was recorded to 2.5 Å resolution at the Swiss Light Source in Villingen, Switzerland (for details see 3.2.7). The data statistics are given in Table 14.

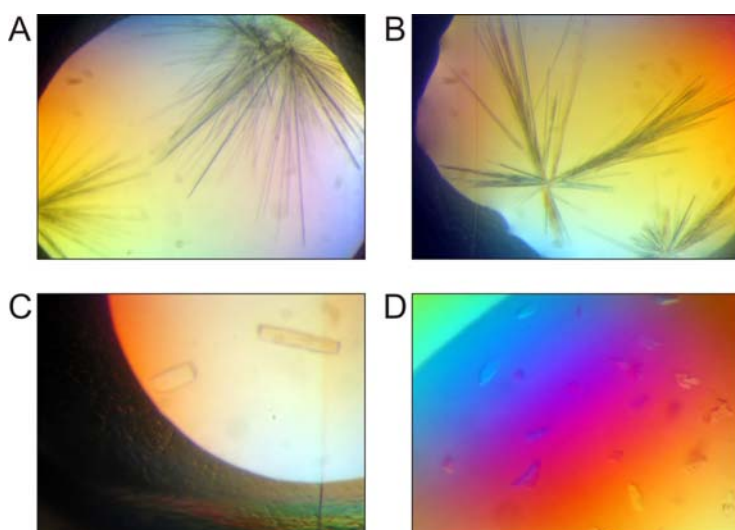
In addition to this crystal form, small rectangular crystals appeared with the 65K<sup>380-517</sup>, which was subjected to reductive lysine methylation prior to crystallization (see chapter 8.5; (Walter, Meier et al. 2006). The more hydrophobic nature of a methylated lysine may favor protein-protein interactions and thereby increase the chances for crystallization. The crystals appeared after 2 days in a condition, which contained 25 % PEG 4000, 0.2 M (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub> and 0.1 M Na-acetate (Figure 22 D). These crystals were cryo-protected in 15 % butandiol and they diffracted only very poorly, so that the unit cell constants and the space group could not be determined.

A third crystal form appeared after 3 days with the short RNA-oligo (Figure 22 C). These rod-shaped crystals diffracted to ~5 Å. Also in this case, identification of the crystal parameters was not possible due to the low quality of the data.

protein and additives	RNA oligo	protein concentration [mg/ml]	gel filtration	number of screened conditions
65K <sup>380-517</sup>	wt	10,2	yes	192
65K <sup>380-517</sup>	wt	8	no	1248
65K <sup>380-517</sup> with 1:2000 chymotrypsin	wt	8	no	480
65K <sup>380-517</sup> with 1:2000 thermolysin	wt	8	no	672
65K <sup>380-517</sup> methylated	wt	8,8	yes	672
65K <sup>380-517</sup> methylated	wt	10,2	no	768
65K <sup>380-517</sup>	5ov	11.7	yes	384
65K <sup>380-517</sup>	short	8	no	864
65K <sup>380-517</sup>	short	10.9	yes	480
65K <sup>380-517</sup>	long	6	yes	576
65K <sup>380-517</sup>	27nt	8	no	768

**Table 13: Initial Crystallization Trials of 65K<sup>380-517</sup>**

The table lists the conditions for initial crystallization trials. The RNA-oligonucleotides, which were co-crystallized, are abbreviated according to Table 3.



**Figure 22: Crystals of 65K<sup>380-517</sup>**

**A** Initial crystals of 65K<sup>380-517</sup> with U12-wt oligo. **B** Refined crystals of 65K<sup>380-517</sup> with U12-wt oligo. **C** Crystals of 65K<sup>380-517</sup> with U12 short oligo. **D** Crystals of methylated 65K<sup>380-517</sup> with U12-wt oligo.

Data Collection	
Wavelength [Å]	0.984
Temperature [K]	100
Space Group	C2
Unit Cell Parameters [Å, °]	
a	91.0
b	33.5
c	49.3
$\beta$	96.6
Resolution [Å]	50-2.5 (2.59-2.50) <sup>a</sup>
Reflections	
Unique	5299 (527)
Completeness [%]	99.9 (100)
Redundancy	3.5 (3.3)
$I/\sigma(I)$	9.6 (2.2)
$R_{\text{sym}}(I)$ <sup>b</sup>	10.8 (56.3)

**Table 14: Statistics from the Data Collection**

The statistics from the data collection on the 65K<sup>380-517</sup> crystals are listed. A table with the refinement statistics can be found in chapter 8.2. a = data from the highest resolution shell;  $b = R_{\text{sym}}(I) = \sum_{\text{hkl}} \sum_i |I_i(\text{hkl}) - \langle I(\text{hkl}) \rangle| / \sum_{\text{hkl}} \sum_i |I_i(\text{hkl})|$ ; for  $n$  independent reflections and  $i$  observations of a given reflection;  $\langle I(\text{hkl}) \rangle$  – average intensity of the  $i$  observations.

#### 4.4.3 Crystal Structure of 65K<sup>380-517</sup>

From the dataset, which was obtained from the needle-shaped 65K<sup>380-517</sup>/U12-wt crystals, the structure was solved by molecular replacement and refined as described in 3.2.8. The refinement statistics are given in Table 19. The resulting structure is the first structure of a minor spliceosome-specific protein. Residues 387-506 could be unequivocally traced in the electron density of the protein, comprising the core of the C-terminal RRM of 65K (aa 415-500) as well as 30 additional residues on the N-terminus. The U12-wt oligo was not visible in the electron density, even though the central  $\beta$ -sheet of RRM, which was expected to bind the RNA, was not blocked by any other protein element. The unique feature of this protein is an N-terminal extension: it folds into two  $\alpha$ -helices, a short  $3_{10}$ -helix and a long loop, which wrap around the protein (Figure 23 C). A search for structural homologs using the DALI-Server (Holm and Sander 1993) did not detect any other known RRM-structure, which possesses a similar N-terminal helical expansion.

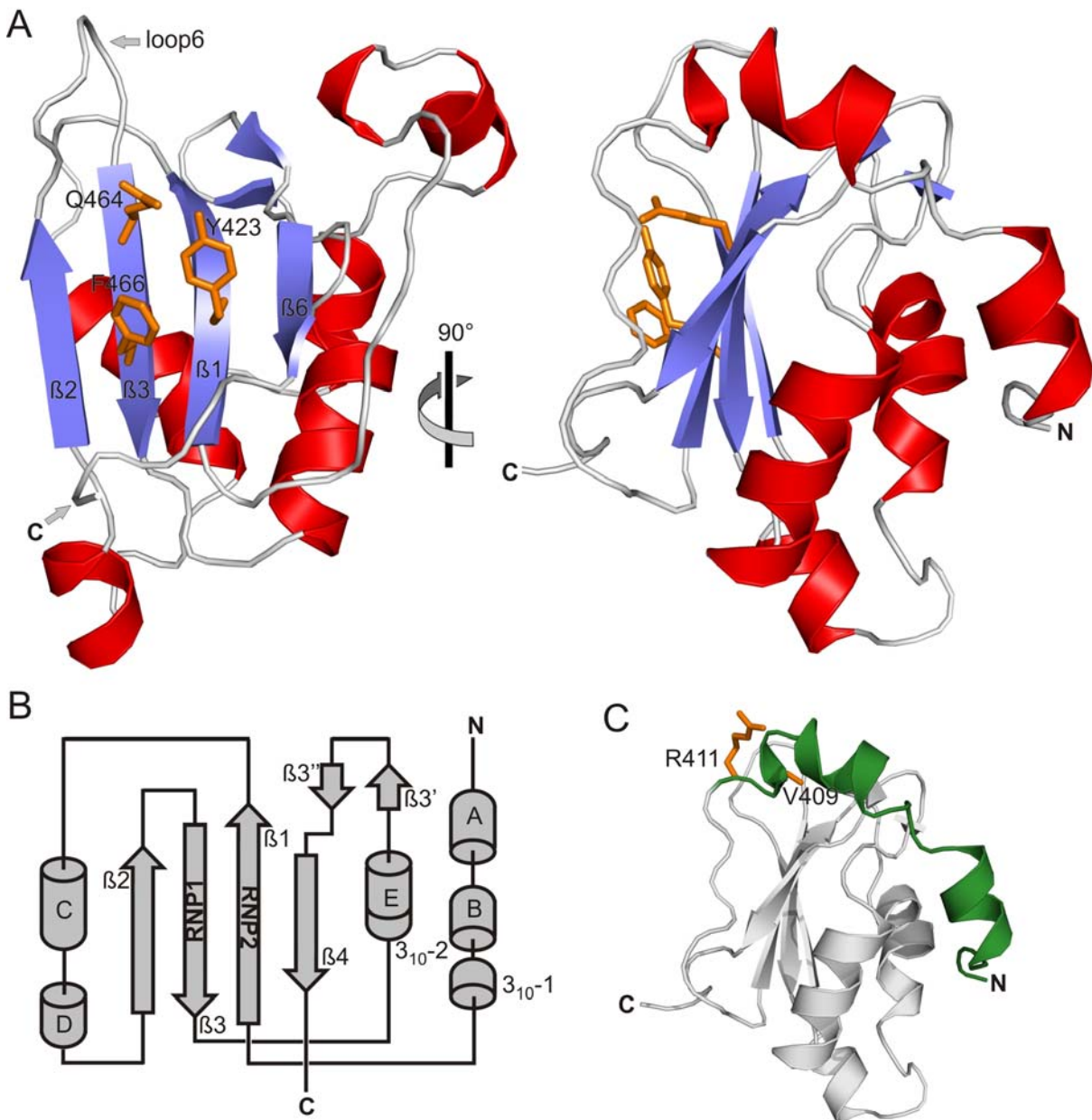
Apart from this unique element, the U11/U12 65K<sup>380-517</sup> structure closely resembles the U1A- and U2B''-nRRM structures. Root mean square deviation (rmsd)-values of the superimposed C $\alpha$ -atoms from U1A and U2B'' (residue range 5-92 in 1URN and 7-92 in 1A9N respectively) on the equivalent residues in the 65K<sup>380-517</sup> structure (see alignment in Figure 24) are in the

range of 1.6-1.8 Å. This emphasizes the structural resemblance of the three proteins. The overall structure and the topology of 65K<sup>380-517</sup> are depicted in Figure 23. Another difference of the core RRM in the 65K<sup>380-517</sup> structure in comparison to the one in U1A is the accommodation of an additional helix between  $\beta$ 1 and  $\beta$ 2, which comprises residues 449-454. A C-terminal helix as in U1A-nRRM could not be detected in the 65K<sup>380-517</sup> structure.

In homologous protein domains, the hydrophobic core is usually composed of conserved residues. In U1A-nRRM twelve amino acids, which are mainly located on the interface of the  $\beta$ -sheet and the  $\alpha$ -helices contribute to the hydrophobic core (Nagai, Oubridge et al. 1990; Kranz, Lu et al. 1996). Overall, this hydrophobic core is maintained in U2B''-nRRM and in the 65K<sup>380-517</sup> structure. A few of these twelve residues are slightly shifted in sequence or altered in the 65K<sup>380-517</sup> structure compared to U1A, but in general they engage in similar interactions and thereby form essentially the same hydrophobic core.

An intricate hydrogen bonding network in U1A- and U2B''-nRRM ensures that RNA-binding residues are positioned correctly. A central element of this hydrogen bonding network in U1A-nRRM is the interaction-triad between Tyr13, Gln54 and Phe56 (Tyr423, Gln468 and Phe470 in 65K-cRRM) (Kranz and Hall 1999; Law, Chambers et al. 2005). On the one hand this interaction-triad provides a platform for non-specific RNA-interaction on the other hand it contacts residues elsewhere in the protein, which make specific RNA contacts, and thereby ensures their proper orientation, enabling sequence-specific RNA-binding. Each residue of the triad is responsible for correct positioning of one other region in the RRM: Tyr13 correctly positions the C-terminal tail via its RNA-interaction, Gln54 orients loop 3 and Phe56 participates in a four-element stack including residues from the C-terminus and two RNA bases (A71 and C72 in U1A-RNA complex).

In the 65K<sup>380-517</sup> structure the hydrogen bonds between the interaction triad of Tyr423, Gln468 and Phe470 are conserved. Concerning conservation of the hydrogen bonds, which are established from the interaction triad to other parts in the protein, a clear statement is difficult. In the present structure contacts of Gln468 to the loop 6 residues Arg464 (Lys50) and Lys466 (Arg52) were not found. Instead they are replaced by side-chain and main-chain contacts to other loop 6 residues between Met460 and Glu462. Additionally, even though Phe470 establishes some of the expected contacts, the involvement in a four-element stack is unclear, as Asp92 of the stack in U1A is replaced by an arginine (Arg505). As the contacts between Tyr423 to the C-terminus should be RNA-mediated and RNA is absent from the 65K<sup>380-517</sup> structure, nothing can be said about this residue. In summary, central elements of the hydrogen bonding network seem to be conserved, but whether the fine-tuned contacts are established in the same way as in U1A and U2B''-nRRM remains unclear.



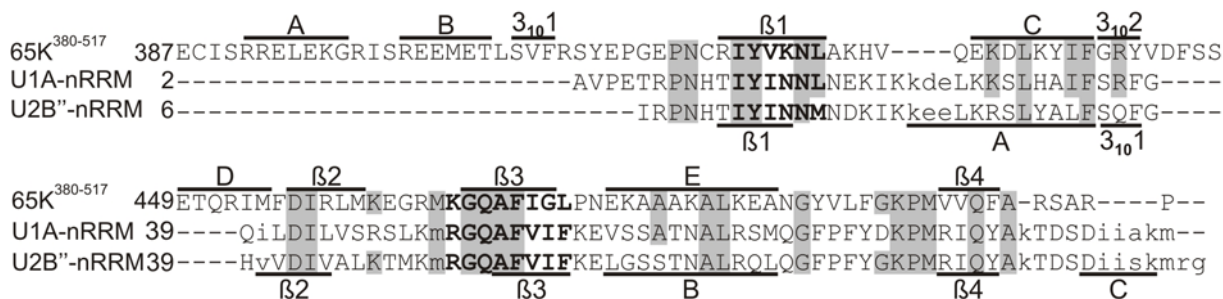
**Figure 23: Crystal Structure and Topology of 65K<sup>380-517</sup>**

The overall structure of the 65K<sup>380-517</sup> protein and the protein topology. **A** View on the  $\beta$ -sheet and 90° rotated side view. The conserved RNP residues are highlighted as orange sticks. The helices are shown in red, the  $\beta$ -sheets in blue. **B** Topology of the 65K<sup>380-517</sup> protein. Helices are depicted as cylinders, sheets as arrows. **C** Structure of the 65K<sup>380-517</sup> protein with the N-terminal helical extension (residues 387-411) highlighted in green. The two amino acids, which were tested for a potential involvement in RNA-binding (V409 and R411; see chapter 4.4.4), are shown as orange sticks.

Comparing the RNA-bound and -unbound states of U1A-nRRM (PDB-entry: 1URN and 1OIA/1NU4 respectively) reveals that RNA-binding induces a more rigid structure especially in loop 3 and the C-terminus (Oubridge, Ito et al. 1994), which are both involved in direct, sequence-specific interactions to the RNA. In case of the 65K<sup>380-517</sup> structure flexible regions, indicated by high B-factors, can be found in the termini, in loop 6 (loop 3 in U1A) and in the region of the additional helix D. Loop 6 (amino acid 461-466) was modeled with two different main chain conformations for residues 461-463 (0.5 occupancy each). Helical turns as in U1A cannot be found in loop 6 of the 65K<sup>380-517</sup> structure. None of the homologous structures adopts a similar loop-conformation as loop 6 in the 65K<sup>380-517</sup> structure. This is indicated by rmsd-values over 3 Å when the position of the C $\alpha$ -atoms in the region between residues 459-467 in the 65K<sup>380-517</sup> structure is compared to the aligning region in U1A and U2B'' (see alignment in Figure 24). In conclusion, the finding that loop 6 is flexible in the 65K<sup>380-517</sup> structure, may well be due to the absence of RNA. If RNA was bound, one would expect loop 6 to be more ordered at least when 65K applies a similar RNA-binding mechanism as U1A.

#### 4.4.4 RNA-Binding by the 65K C-Terminal RNA Recognition Motif

In the two-step RNA-binding mechanism, which has been proposed for U1A-nRRM, the RNA is first recruited to the protein by interaction of positively charged amino acids on the protein surface with the negatively charged RNA-backbone. In a second step the close range interactions are established and the sequence-specific contacts are formed (Katsamba, Myszka et al. 2001). These interactions are mainly mediated by residues from loop 3, the variable region preceding loop 3 and the C-terminus (Katsamba, Bayramyan et al. 2002). I attempted to find out, whether 65K has implemented the same strategy to bind its target RNA. For a detailed investigation of the RNA-binding mechanism, a closer look was taken at residues and regions, which are relevant for RNA-binding in U1A- and U2B''-nRRM. Mutations were introduced in the core 65K-cRRM on the basis of a structure based sequence alignment between the 65K<sup>380-517</sup> structure and the two homologous proteins (Figure 24). Subsequently, the RNA-affinities of the mutated 65K<sup>380-517</sup> GST-fusion proteins were monitored in gel shift assays with the U12-wt RNA.



**Figure 24: Structure Based Sequence Alignment of the Homologous RRM**s

Structure based sequence alignment of the 65K<sup>380-517</sup> structure, U1A-nRRM (chain A; PDB-entry: 1URN\_A) and U2B''-nRRM (chain B; PDB-entry: 1A9N\_B). Only those residues, which are visible in the crystal structures, are aligned. The secondary structure elements (black bars and a label) of 65K<sup>380-517</sup> are depicted above the three sequences, the secondary structure of U1A and U2B'' is shown below the three sequences. The labels of secondary structure elements are as follows:  $\alpha$ -helices are symbolized by a bold letter,  $3_{10}$ -helices are abbreviated with  $3_{10}$  and a number and  $\beta$ -sheets are named with  $\beta$  and a number. The RNP consensus sequences are shown in bold letters. Residues, which are identical to 65K<sup>380-517</sup> are shown in a grey box.

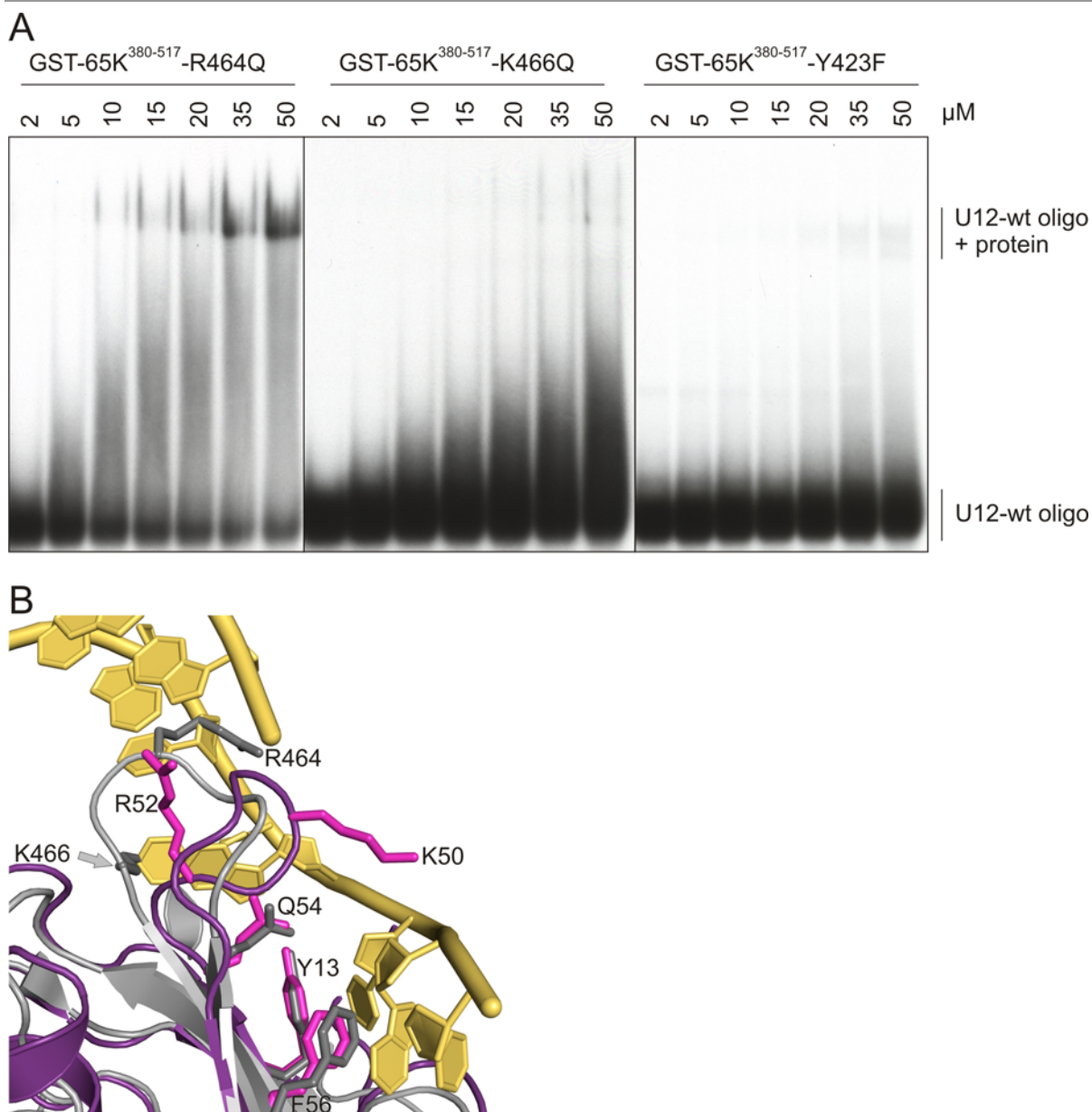
Loop 3 plays a special role in RNA-binding by U1A and U2B'', because it protrudes through the RNA-loop and thereby splays out the RNA-bases to promote their sequence specific recognition. Residues Leu49, Lys50 and Arg52 from loop 3 fulfill particular roles in this process. Leu49 contributes to the longevity of the U1A-RNA complex, because with its branched side chain it hooks loop 3 on the RNA (Laird-Offringa and Belasco 1995). In the 65K<sup>380-517</sup> structure, a glycine (Gly463) can be found at the corresponding position in loop 6 (loop 3 in U1A; Figure 25 B). Most likely this unbranched residue will not be able to hook the protein loop on the RNA-loop. Another important residue in U1A is Lys50: it participates in the initial RNA recruitment and more importantly by contacting Arg52 it holds this crucial residue in place. In 65K-cRRM, Lys50 is conservatively replaced by an arginine (Arg464). Mutation of Arg464 to a glutamine causes a 34 fold decrease in the affinity to U12-wt RNA as measured in gel shift assays with the 65K<sup>380-517</sup> GST-fusion protein (Figure 25 A). This is in accordance with U1A, where the RNA-affinity is decreased over 10 times by a glutamine at this position (Nagai, Oubridge et al. 1990; Katsamba, Myszka et al. 2001). In U1A, the most crucial loop 3-residue is Arg52. It contacts the loop-closing base-pair in U1 snRNA hairpin II in a sequence-specific manner: hydrogen bonds with N7 and O6 of G76 are formed (Oubridge, Ito et al. 1994) and N1 of A66 is contacted. In the 65K<sup>380-517</sup> structure a lysine, Lys466, instead of an arginine is found at the corresponding position. This should still allow formation of the contact to A66 and to O6 of G76. However, whether Lys466 could occupy the position necessary for this interaction remains unclear from the structure. Binding of GST-65K<sup>380-517</sup> to



the U12-wt oligo was almost completely abolished when Lys466 was mutated to a glutamine (Figure 25). This observation correlates well with the situation in U1A, where the Arg52Gln mutation abolishes RNA-binding completely, while it is restored partially by the conservative mutation to a lysine (Nagai, Oubridge et al. 1990; Oubridge, Ito et al. 1994). In conclusion, when Arg464 and Lys466 in the 65K-cRRM core are altered, RNA-binding by 65K<sup>380-517</sup> is abrogated. Therefore these residues could in principle carry out similar functions as in U1A and U2B", whereas Gly463 is unlikely to serve the same purpose as Leu49.

Comparisons to other RNA recognition motifs showed that not only the sequence of loop 3 contributes to target-specificity, but also the length of the loop is important: the loop length varies with the size of the target RNA-loop (Katsamba, Bayramyan et al. 2002). In the 65K<sup>380-517</sup> structure, loop 6 ranges from amino acid 461 to 466, comprising only six amino acids. This length is significantly shorter than in U1A and U2B", where the loop is composed of nine residues. Interestingly, the U12 snRNA hairpin III loop only comprises 7 instead of 10 nucleotides as in the U1 and U2 snRNA-targets. A shorter loop 6 could therefore suffice in order to protrude through the RNA.

A second set of residues is involved in the close range interactions between U1A and its RNA-target: the aromatic residues on the  $\beta$ -sheet, which provide a general binding platform (Allain, Howe et al. 1997; Zeng and Hall 1997; Maris, Dominguez et al. 2005). These residues stack on the RNA-bases and thereby form an "inter-molecular hydrophobic core" with the RNA (Kranz and Hall 1999). As discussed in the context of the interaction-triad (see chapter 4.4.3), the  $\beta$ -sheet aromatics also contribute indirectly to target-specificity, because they position other RNA-contacting residues (Kranz and Hall 1998; Kranz and Hall 1999). The most important residue in this respect in U1A is Tyr13 on position 2 of the RNP2. Even the conservative mutation Tyr13Phe abolishes RNA-binding completely (Oubridge, Ito et al. 1994). In the 65K-cRRM core, the corresponding Tyr423Phe mutation leads to a very dramatic loss in affinity of the 65K<sup>380-517</sup> GST-fusion protein to U12-wt RNA in gel shift assays (Figure 25). A Tyr–Ala mutation entirely abolishes RNA-binding (data not shown). This finding is in accordance with the observed effects in U1A.

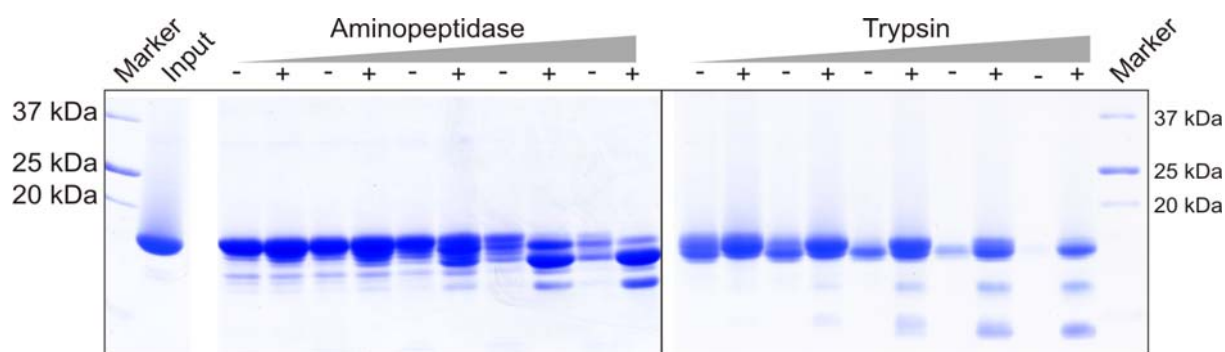


**Figure 25: RNA-Binding of Mutants in Loop 6 and of  $\beta$ -Sheet Aromatics**

Gel shift assay of mutants in loop 6 and Tyr423 and the superimposition of the structures of U1A in the RNA-bound state and 65K<sup>380-517</sup>. **A** Gel shift assay of GST-65K<sup>380-517</sup>-R464Q, GST-65K<sup>380-517</sup>-K466Q and GST-65K<sup>380-517</sup>-Y423F with the wildtype (wt) RNA-oligomer from U12 snRNA hairpin III (U12-wt RNA).  $K_{Dapp}$ 's were estimated to be  $\sim 550$   $\mu$ M for GST-65K<sup>380-517</sup>-R464Q. For GST-65K<sup>380-517</sup>-K466Q and GST-65K<sup>380-517</sup>-Y423F an estimation was impossible. **B** Superimposition of U1A-nRRM in violet in its RNA-bound state and the 65K<sup>380-517</sup> structure in grey in the area of loop 3/6 and the  $\beta$ -sheet. The RNA is shown in yellow. Important residues are shown as sticks in light-magenta for U1A and dark grey for 65K<sup>380-517</sup>.

Another similarity between U1A-nRRM and 65K-cRRM is the adoption of a more compact fold upon RNA-binding. In U1A loop 3 and the C-terminus become better ordered and folded in the presence of RNA. Limited proteolysis of 65K<sup>380-517</sup> in the presence and in the absence of U12-wt oligo revealed the same: the digestion pattern of the protein was different when the RNA-oligo was present and the protein was partially protected against proteolysis in the presence of RNA (Figure 26). Additionally, thermal unfolding experiments using circular dichroism showed that the melting temperature for 65K<sup>380-517</sup> in the presence of RNA is almost 10 °C higher than in the absence of RNA (64 °C instead of 55 °C respectively; data not shown).

Taken together, it seems that the sequence-specific and the close range interactions with RNA targets are very likely similar in U11/U1265K-cRRM when compared to U1A- and U2B" amino-terminal RRM. Additionally, all three proteins adopt a more compact fold upon RNA-binding. Therefore, the mechanism of RNA-binding seems to be conserved between these RNA recognition motifs, which is in accordance to their sequence similarity, the resemblance of their protein- and their RNA-target structures.



**Figure 26: Limited Proteolysis of 65K<sup>380-517</sup>**

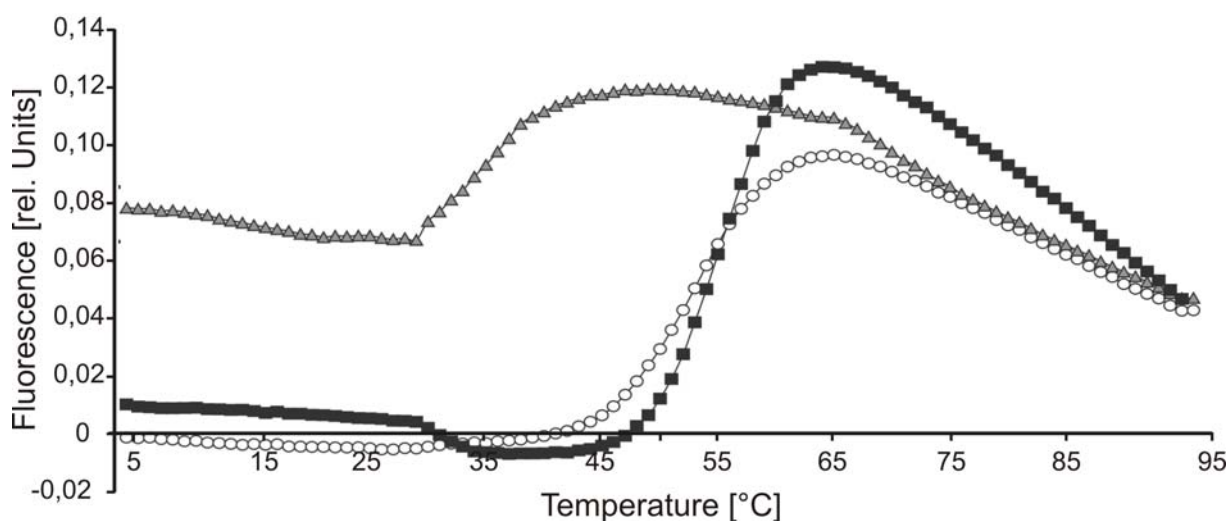
A comparative limited proteolysis between 65K<sup>380-517</sup> in the absence and in the presence of U12-wt RNA is shown. “-“ designates the lanes without RNA, “+” the lanes with RNA. the grey triangle symbolizes increasing concentrations of protease.

#### 4.4.5 The Role of the N-Terminus

The N-terminus runs like a clamp around the core RRM domain. As proven in gel shift assays with truncation mutants of U11/U12 65K, the absence of this N-terminal extension causes a complete loss of RNA-affinity (see chapter 4.4.1). This indicates a central role for the N-terminal helical extension in RNA-binding. I suspected that the N-terminus of 65K<sup>380-517</sup>

either directly participates in RNA-binding or is an integral part of the RRM domain. In the latter case removal of the N-terminal residues could cause instability of the domain structure and thus a loss of RNA-affinity.

To test whether the N-terminus directly participates in RNA-binding, potential RNA-contacting residues were mutated and the mutants were tested in gel shift assays. An overlay of the U1 hairpin II on the 65K<sup>380-517</sup> structure revealed two amino acids as candidates for a participation in RNA-binding: Val409 and Arg411. They are located in the loop after the second helix, neighboring  $\beta$  4 (see Figure 23 B). In the overlay the two residues are in close proximity to the nucleotide U68 of U1 hairpin II, which could allow van-der-Waals contacts of Val409 or sequence-specific contacts of the flexible Arg411 (overlay not shown). Mutating these residues to alanine and glutamine, respectively, caused a slight loss in binding (3.0 and 1.9 fold reduced-affinity when GST-fusion proteins were compared to GST-65K-cRRM; data not shown). In conclusion, RNA-binding with mutated N-terminal expansions remained strong. This observation contrasts with the complete loss of binding upon mutation of crucial RNA-binding residues such as Tyr423 or Lys466. Therefore the N-terminal extension engages best in superficial RNA contacts, which are not required for RNA-binding per se.



**Figure 27: Thermal Melting of Truncated 65K Proteins**

Thermal melting using real-time PCR is shown for the three truncation mutants of 65K. The grey rectangles resemble the melting curve of 65K<sup>411-505</sup>, the white circles show the 65K<sup>380-517</sup> curve and the black squares depict the curve for 65K<sup>380-506</sup>.

Next the idea was tested whether the N-terminal helical extension was important for the overall stability of the domain. Using a fluorescence-based thermal melting assay, the stabilities of 65K-cRRM constructs including or lacking the N-terminal expansion were compared. The following principle underlies these assays: the fluorescence of the amphiphilic dye sypro orange increases as soon as the dye enters a hydrophobic environment (Semisotnov, Rodionova et al. 1991). During the thermal unfolding of a protein, the hydrophobic residues of the core become exposed. The fluorescence of sypro orange increases substantially when it gains access to the hydrophobic core of a protein that unfolds in thermal melting. From the fluorescence change, melting temperatures ( $T_m$ ) can be estimated.  $T_m$  is the temperature at which half of the protein is unfolded. 65K<sup>380-505</sup> and 65K<sup>380-511</sup> exhibited melting temperatures of 54 °C and 52 °C, respectively (Figure 27). This demonstrates that the removal of the C-terminal expansion has no significant influence on the protein stability. In sharp contrast, 65K<sup>411-505</sup>, lacking additionally the N-terminal expansion, exhibited a dramatically reduced  $T_m$  of about 36 °C. Additionally, this construct showed an increased level of dye incorporation even at low temperatures (high intercept at the y-axis), which is an indication for a protein that is not properly folded *a priori* (Semisotnov, Rodionova et al. 1991). All these results are in good agreement with circular dichroism measurements, in which a small, but continuous slope of the 65K<sup>411-505</sup> melting curve indicated non-cooperative unfolding. The two other constructs showed sigmoidal curves indicating cooperative unfolding (data not shown). Taken together, the instability of the protein in the absence of the N-terminal extension is an indication that this part of 65K<sup>380-517</sup> may play a role in maintaining the overall stability and fold of the RRM domain.

## 5 Discussion

### 5.1 The Prp19-Complex

The Prp19-complex is essential for pre-mRNA splicing in diverse organisms. It plays a crucial role in spliceosome activation, where it may help to define the correct base pairing between the pre-mRNA, U5 and U6 snRNP. Possibly the complex is needed for spliceosome assembly in the early spliceosomal complexes as well. In the present study I managed to reveal some novel insights into the human Prp19-complex: it was demonstrated that the human Prp19 coiled-coil domain forms a homo-tetramer and that CDC5 interacts with this domain. In the *S. cerevisiae* nineteen complex, a known model for the overall architecture of the Prp19p protein could be confirmed by electron microscopy negative stain imaging. Moreover, some of the domain borders in Prp19p were mapped. They are in well agreement with previous studies, whereas the minimal Prp19p coiled-coil region, which is needed for tetramerization, was mapped slightly shorter on the C-terminus.

The interaction between human Prp19 and CDC5 resembles the known interaction between Prp19p and Cef1p in *S. cerevisiae*. Together with the tetramerization of the human Prp19 coiled-coil, these findings suggest that the model, which has been proposed for the interactions among the Prp19-complex proteins in *S. cerevisiae* (see chapter 2.3.3), is transferable to the human system at least as far as hPrp19 and CDC5 are concerned.

One point to discuss in this context is the role of Spf27 or its functional counterpart in yeast, Snt309p. It has been proposed that Snt309p first needs to bind to Prp19p in order to modulate the interaction properties of Prp19p (Chen, Tsao et al. 1999; Ohi and Gould 2002). Only after this binding, a stable nineteen complex can be formed. However, interactions between Prp19p and other proteins are in principle possible without Snt309p-involvement. The interactions will just not be as stable as in the presence of Snt309p. Therefore the detected interaction between the human Prp19 coiled-coil and CDC5 without the involvement of Spf27 is not unusual. It most likely is of the same nature as the interaction between Prp19p and Cef1p in *S. cerevisiae*.

Further insights into the Prp19-complex architecture in humans and *S. cerevisiae* were prohibited by difficulties to express the proteins in sufficient degree of purity and in sufficient amounts for structural studies. It is important to find solutions to the question where the difficulties in recombinant expression may come from and how they might be circumvented. Interesting insights in this respect come from expression trials, which have been performed in the different expression hosts.

Possible reasons for difficulties with the expression in *E. coli* might for example be that the required tRNAs are not present, that proper folding cannot be achieved or that necessary post-translational modifications may be missing. Expression in strains, which carry additional tRNAs, low expression temperatures, optimized media, slow induction, variation of the time point of induction and the length of the expression time provide some means to evade these problems. All these strategies were applied in the course of this work without success suggesting that the *E. coli* might not be suited for the production of these proteins.

A closer look at other types of expression hosts, which have been tested for the human Prp19-complex proteins, provides further information. Initial expression trials with human PRL1 have been performed in the amoeba *Dictyostelium discoideum* in collaboration with another department. In this expression host the protein was not expressed at a detectable level. I performed first tests with PRL1-WD40 in *Baculo* virus infected insect cells, but could not achieve expression either. Extensive experiments to establish stable HeLa S3 cell lines overexpressing the human Prp19-complex proteins have been carried out by Dr. I. Lemm from my department. For hPrp19, CDC5 and PRL1 these cell lines could not be created, which is an indication that overexpression is lethal for the cells (personal communication with Dr. I. Lemm).

This co-occurrence of difficulties in different expression hosts suggest a more general problem with the essential human Prp19-complex proteins: the expression levels of these proteins seem to be strictly regulated in higher eukaryotes, which prevents any overproduction. For the non-essential proteins AD002 and Spf27, however, the HeLa cell lines were vital. This indicates that the non-essential proteins are not as strictly regulated. In clear contrast to the expression in higher eukaryotes, at least small amounts of the essential human proteins hPrp19, CDC5 and PRL1 could be overexpressed in *E. coli*, even though they did not seem to be folded properly. The following conclusion can be drawn: expression of the three essential human proteins in lower organisms might in general be possible, because in these organisms overexpression of the proteins is not lethal. But lower organisms may in turn not be able to perform the correct folding or necessary modifications, so that the observed phenomena like insolubility, co-purification of chaperones or degradation occur. On the other side, for higher organisms, which potentially perform correct folding and modifications, overexpression of the essential proteins from the human Prp19-complex seems to be lethal. Therefore a screening for expression hosts should be undertaken in order to find an organism, which is able to perform proper folding and some modifications, but does not regulate the expressions of these proteins as strictly as human cells. Alternatively, cell free systems can be employed to produce the human Prp19-complex proteins. We already demonstrated that the proteins can be produced in these systems: a far smaller amount of

degradation was observed in the *in vitro*-translated CDC5 protein than in the CDC5 protein, which was produced in *E. coli*. Moreover, the hPrp19 and the CDC5 proteins seem to be functional, because they interact. Functionality and less degradation are signs of a more correct folding of the human proteins in the wheat germ *in vitro*-translation system, which therefore constitutes a potential option for protein expression.

Most likely the step, at which production of the human Prp19-complex proteins failed in *E. coli* was protein folding, because some of the essential proteins were expressed, but appeared to be insoluble, degraded or they co-purified with a chaperone. As an alternative to the suggested switch of the expression host, I thought to circumvent the problem by coexpressing the proteins. Coexpression of interaction partners might help to achieve correct folding and thereby yield a higher solubility. Several pair-wise combinations of the human Prp19-complex proteins were coexpressed (see Table 9). According to my results, however, coexpression did not occur. I tested whether the vectors are not compatible with each other. However, the same vector types worked in positive controls, which excluded vector-incompatibility. The reason for the absence of coexpression remains unclear.

Despite the difficulties in recombinant protein production, some proteins were obtained and subjected to crystallization. However, no crystals could be produced. The proteins or protein domains, which were subjected to crystallization, are the human Prp19 coiled-coil domain, the yeast full-length Prp19p protein and the three truncated Prp19p proteins containing the U-box and the coiled-coil domain. However, according to the proposed model of the *S. cerevisiae* Prp19p protein the WD40- and the U-box domains are flexibly attached to the central coiled-coil. This implies a high degree of flexibility, which might prevent crystallization at least in the full-length protein, but most probably also in the three Prp19p truncation mutants. The difficulties in crystallization of the human Prp19 coiled-coil could be due to the construct length: most probably the construct is too long in its N-terminus. It may therefore include some of the residues, which flexibly link the U-box domain to the coiled-coil. These residues in turn might cause the cessation of crystal growth and thereby prevent structure determination.

Lately, progress was made in the native purification approach: a HeLa S3 cell line, which stably expresses FLAG/HA-tagged AD002, has been established by Dr. I. Lemm. With this FLAG/HA-tagged AD002 M. Grote from my department managed to purify the Prp19-complex from HeLa nuclear extract with a 2-3 times higher yield compared to antibody purification methods. Electron microscopic negative stain images of the purified complexes were obtained, which clearly revealed distinct particles. It is unclear whether enough material for crystallographic studies can be obtained. However, the natively purified complexes can be subjected to limited proteolysis in order to define a minimal Prp19-complex. The mapped



protein domains might then be produced recombinantly in a suitable expression host, eventually even in *E. coli*.

Summarizing the studies on the human and *S. cerevisiae* Prp19-complex, a few pieces could be added to the puzzle, which turned out to be more intricate than previously assumed. All obstacles on the way, however, made the Prp19-complex even more interesting: the question how a protein complex, which may exert different functions in different stages of the spliceosomal cycle, finally promotes spliceosome activation, is not only interesting from a functional point of view. Different aspects of the Prp19-complex architecture will allow intriguing insights: the potential switch mechanism initiated by Snt309p, which may alter protein interactions, and the protein-protein interaction interfaces, which are suited for the design of small molecule inhibitors, provide interesting topics for future research.

## 5.2 The U11/U12 65K C-Terminal RNA Recognition Motif

The U11/U12 65K protein forms a molecular bridge between the U11 and the U12 snRNP in the minor spliceosome. It thereby achieves intron-bridging and establishes the close proximity between the 5'-splice site and the branch point, which is required for the first catalytic step in splicing. In the present investigation of the U11/U12 65K carboxy-terminal RRM it was shown that approximately 30 residues amino-terminally of the RRM core are essential for binding to the U12 snRNA hairpin III. The homologous N-terminal RRMs in U1A and U2B'' do not require a similar extension for RNA-binding. Apart from this, the RNA-binding by U11/U12 65K-cRRM seems to be similar to the homologous proteins U1A and U2B'': mutations of critical RNA-binding residues in 65K-cRRM evoked comparable effects. Additionally, all three proteins adopt a more compact fold upon RNA-binding. The crystal structure of the expanded 65K C-terminal RRM showed that the  $\alpha$ -helical N-terminal extension wraps around the protein. Removal of this N-terminal tail significantly decreased the thermal stability of the 65K-cRRM. These results demonstrate that the 65K C-terminal RRM is supplemented by an N-terminal extension, which confers stability on the domain and thereby preserves the RNA-binding capacity.

### 5.2.1 RNA-Binding and the Role of the C-Terminus

Despite the global similarities in RNA-binding, details may differ. These differences may be the reason for the lower affinity of 65K-cRRM to the U12 snRNA hairpin III ( $K_{Dapp} \sim 11 \mu\text{M}$ ) compared to U1A-nRRM ( $K_{Dapp} = 32 \text{ pM}$ ). The replacement of the critical arginine in loop 3 in U1A-nRRM by a lysine (Lys466) in loop 6 of 65K-cRRM might contribute to a lower affinity.

Additionally, Leu49 in U1A is replaced by Gly463 in 65K-cRRM. Therefore the RNA might not be anchored to the RRM as tightly in 65K-cRRM as in U1A, because the RNA-loop cannot be hooked to the  $\beta$ -sheet.

Even though it is only a small effect, the observed two fold increase in RNA-affinity upon deletion of C-terminal amino acids in 65K<sup>380-506</sup> also differs to U1A. This could suggest an inhibitory role of the C-terminus in 65K-cRRM. Such an inhibition of RNA-binding can occur when the  $\beta$ -sheet is blocked by C-terminal residues. Examples can be found in the cap binding protein 20 (CBP20) and the polypyrimidine tract binding protein (PTB) (Mazza, Segref et al. 2002; Maris, Dominguez et al. 2005). In both proteins the residue on RNP1 position 5 is blocked by residues from the C-terminus. In 65K-cRRM a phenylalanine (Phe470) is located at this position 5 of RNP1. However, the C-terminal residues, which could cause the potential inhibition (residues 507-517), are not visible in the crystal structure of 65K<sup>380-517</sup>. Therefore it is unclear whether a block of Phe470 causes the reduced affinity of 65K<sup>380-517</sup> compared to 65K<sup>380-506</sup>. At least the residues, which are visible in the crystal structure (aa 387-506) do not lay over the  $\beta$ -sheet, which makes a block of the  $\beta$ -sheet by these residues improbable.

Another aspect to discuss regarding the C-terminus is the potential relevance of Pro506 for RNA-binding. In gel shift assays 65K<sup>380-506</sup> binds U12-wt RNA, whereas 65K<sup>411-505</sup> does not. It was concluded that this is due to the absence of the N-terminal residues. Still the proline506 could also be (partially) responsible for the observed loss of affinity. Pro506 does not establish any contact to other residues in the protein. Its deletion will therefore most likely not cause instability of the domain, which might lead to the observed loss of RNA-binding. Still Pro506 could establish a stacking interaction to RNA. Single stacking interactions can contribute significantly to the stability of a protein-RNA complex (Blakaj, McConnell et al. 2001). Nonetheless, it is very improbable that a single stacking interaction causes a complete loss of RNA-affinity. Taken together, the absence of Pro506 most likely contributes little to the observed loss of RNA-binding in 65K<sup>411-505</sup>.

### 5.2.2 Functional Stabilization of the 65K C-Terminal RNA Recognition Motif

The  $\alpha$ -helical N-terminus of U11/U12 65K-cRRM contributes to the stabilization of the core RRM and in this way it influences the RNA-affinity. This protein region adopts a loose fold and it is predicted to be unfolded by sequence-based fold prediction programs. This raises the question whether the N-terminal helical extension is an extended part of the RRM domain or whether it should rather be envisioned as an autonomous element.

Fold prediction programs designate the N-terminus of 65K<sup>380-517</sup> to be unfolded. In the structure a loose fold with relatively high B-factors ( $\sim 37 \text{ \AA}^2$  for aa 387-411 compared to  $\sim 33 \text{ \AA}^2$  for aa 412-506) can be found. Despite features, the N-terminus seems to constitute an integral part of the RRM domain, as its removal causes instability of the domain. A similar mechanism, in which a canonical folding motif is stabilized by extensions to the domain, has been observed in the *S. cerevisiae* Prp8p protein (Pena, Liu et al. 2007). This protein contains a domain with a Jab1/MPN-like core known from deubiquitinating enzymes, onto which insertions and terminal appendices are grafted. The N- and C-terminal appendices and the two insertions are tightly wrapped around the Jab1/MPN-like core and contribute to its stability. One point, at which this stabilization became evident, was the purification behavior of Prp8: in contrast to the core alone, the protein with the extensions could easily be purified. The same is seen in 65K-cRRM: the deletion mutant 65K<sup>411-505</sup> was hard to purify compared to 65K<sup>380-517</sup> and 65K<sup>380-505</sup>.

A closer look at the N-terminus and the interface between the core RRM and the N-terminal helical extension reveals that the N-terminus of the protein exhibits some typical features of a so-called intrinsically unfolded protein (IUP) or intrinsically unfolded domain (IUD) (Meszaros, Tompa et al. 2007). IUPs undergo a disorder-order transition upon binding to their partner, which can be another protein, DNA or RNA, but do usually not adopt a compact globular fold. Numerous examples of IUPs have been noted in the areas of cell cycle control, transcriptional and translational regulation (Wright and Dyson 1999). Among other characteristics IUPs often use a high percentage of their total surface area to form the interface to their interaction partner. Moreover the amino acids in IUPs have a large surface area per residue, which is involved in the interface. These characteristics can be found for the N-terminus of 65K-cRRM as well: 30.2 % of the total surface of the N-terminus is used to build up the interface to the core RRM. Additionally, the surface area per residue, which is involved in the interface, is almost two times larger for the N-terminal amino acids than for amino acids from the core RRM domain.

It remains to be seen whether the N-terminal helical extension is permanently grafted onto the RRM core during the splicing cycle. Alternatively, its intermediate dissociation could be used to modulate the RNA-affinity of 65K-cRRM at certain stages. Interestingly, it has been shown in a previous finding that the residues between the proline rich region and the C-terminal RRM influence the C-terminal RRM as well as the N-terminal half of the 65K protein (Benecke 2004). Even though it is speculative a temporary dissociation of the N-terminal extension from the core RRM, which modulates the affinity of 65K to the U12 snRNA, could functionally be coupled to 59K binding in the N-terminus of the protein and thereby to the U11 snRNA-binding.

## 6 Outlook

### 6.1 The Prp19-Complex

Investigation of the human Prp19-protein complex needs to be based on a broad approach: material from native purifications is going to be used to determine an electron microscopic structure, from which the overall shape of the complex can be determined. To elucidate fine details, single protein domains and interacting domains solved by X-ray crystallography can be fitted in the electron microscopic structure. In order to define these domains, the natively purified complex should be subjected to limited proteolysis followed by mass spectrometric analyses. The defined globular protein domains can subsequently be produced by *in vitro*-translation or for example in yeast (e.g. in *Pichia pastoris*), in some cases perhaps also in *E. coli*. The hPrp19 coiled-coil domain, which was produced in the course of this study, should be analyzed with one-dimensional NMR to identify the flexible residues. Because of the small amounts obtained in recombinant production, labeling for NMR analysis might be difficult. A re-designed construct might serve to improve the crystals for structure determination.

In parallel, the effort to assemble the complex recombinantly should be continued. In this way smaller assembly intermediates can be isolated and their structures, functions and interactions can be studied. An assembly pathway for the complex can be proposed as it has been done for other multi-protein complexes like the nuclear pore complex (Lutzmann, Kunze et al. 2002). Eventually the entire complex might be gained in sufficient amounts to subject it to crystallization. Native purification or *in vitro*-translation of several proteins might not yield the required amounts. Crystallographic structure determination of the entire Prp19-complex could be possible, because the flexibility of individual components in the fully assembled core complex might well be reduced to a great extent. Alternatively, a recombinantly assembled complex can be subjected to limited proteolysis and crystallized afterwards.

For the *S. cerevisiae* nineteen complex the following is suggested: the core proteins should first be re-cloned into standard expression vectors for individual expressions in *E. coli*. In a first approach, proteins that interact with Prp19p should be investigated. Electron microscopic negative stain imaging provides a good basis to monitor changes in the Prp19p-structure, which might be introduced by interaction partners. In combination with limited proteolysis, protein truncation and interaction assays, constructs suitable for crystallization may be produced. Additionally, the existing Prp19p-constructs from this study should further be investigated: for the Prp19p U-box coiled-coil proteins, structure determination by NMR is

possible. Alternatively, the Prp19p coiled-coil and the WD40-domain can be investigated individually with NMR or X-ray crystallography and can later on be modeled according to electron microscopic images. The other interaction worth to be studied is the interaction between Cef1p and Prp46p. In this case, the minimal interacting domains, which were mapped to residues 427-590 in Cef1p and amino acids 106-450 in Prp46p in earlier studies (Ohi and Gould 2002) should be cloned and either coexpressed or expressed individually. Alternatively, the full-length proteins can be expressed from standard expression vectors. Co-purification and limited proteolysis or production of truncated proteins combined with interaction studies may lead to a minimal interacting complex, which is suited for structural studies.

The Prp19-complex in humans and in *S. cerevisiae* seems to play a role in several steps of the spliceosomal cycle. Most importantly, it functions in spliceosome activation directly prior to the first transesterification reaction. The ongoing structural and functional investigations on the Prp19-complex will therefore shed light on one of the most important steps during pre-mRNA splicing and thereby broaden the knowledge about a central step in protein biosynthesis.

## 6.2 The U11/U12 65K C-Terminal RNA Recognition Motif

The amino-terminal extension of the U11/U12 65K C-terminal RNA recognition motif is an interesting object for ongoing studies. This region of the protein constitutes an element, which might be used to couple RNA-binding to other functional events like the binding of 59K. To confirm this hypothesis, binding and structural studies with the full-length 65K protein, the U12-wt RNA and the 59K protein should be undertaken. In addition to its possible functional role, the N-terminal helical extension constitutes an element, which is not conserved among the homologous proteins U1A, U2B" and U11/U12 65K. Therefore it may provide insights into evolutionary relationships between the three proteins and on a higher level between the major and the minor spliceosome.

To reveal details of the mechanism of RNA-recognition, a crystal structure of U11/U12 65K c-RRM in complex with RNA may be determined. Crystallization trials of the 65K<sup>380-506</sup> protein, which binds the U12-wt RNA with a higher affinity, are underway. It may be necessary, though, to apply surface engineering as it has been performed in U1A (Oubridge, Ito et al. 1995). This could be achieved by the introduction of mutations, which alter the crystal interface in the RNA-free protein. This surface engineering might be best performed on the C-terminal truncation mutant.

A detailed picture of the 65K-cRRM RNA-binding mechanism will allow conclusion on the relationship of U1A, U2B” and U11/U12 65K. Moreover, the interdependence of U12 snRNA-binding and 59K-binding will provide in depth insights into the molecular bridging function. Together these results may help to unravel major differences between the two types of spliceosomes.

## 7 References

- Achsel, T., H. Brahms, et al. (1999). "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**(20): 5789-802.
- Ajuh, P., B. Kuster, et al. (2000). "Functional analysis of the human CDC5L complex and identification of its components by mass spectrometry." *Embo J* **19**(23): 6569-81.
- Ajuh, P. and A. I. Lamond (2003). "Identification of peptide inhibitors of pre-mRNA splicing derived from the essential interaction domains of CDC5L and PLRG1." *Nucleic Acids Res* **31**(21): 6104-16.
- Ajuh, P., J. Sleeman, et al. (2001). "A direct interaction between the carboxyl-terminal region of CDC5L and the WD40 domain of PLRG1 is essential for pre-mRNA splicing." *J Biol Chem* **276**(45): 42370-81.
- Albers, M., A. Diment, et al. (2003). "Identification and characterization of Prp45p and Prp46p, essential pre-mRNA splicing factors." *RNA* **9**(1): 138-50.
- Allain, F. H., P. W. Howe, et al. (1997). "Structural basis of the RNA-binding specificity of human U1A protein." *Embo J* **16**(18): 5764-72.
- Ban, N., P. Nissen, et al. (2000). "The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution." *Science* **289**(5481): 905-20.
- Bandziulis, R. J., M. S. Swanson, et al. (1989). "RNA-binding proteins as developmental regulators." *Genes Dev* **3**(4): 431-7.
- Behzadnia, N., M. M. Golas, et al. (2007). "Composition and three-dimensional EM structure of double affinity-purified, human prespliceosomal A complexes." *EMBO J* **26**(6): 1737-48.
- Ben-Yehuda, S., I. Dix, et al. (2000). "Genetic and physical interactions between factors involved in both cell cycle progression and pre-mRNA splicing in *Saccharomyces cerevisiae*." *Genetics* **156**(4): 1503-17.
- Benecke, H. (2004). RNA/Protein & Protein/Protein-Interaktionen innerhalb des U11/U12 di-snRNP Partikels des U12-abhängigen Spleißosoms aus HeLa-Zellen. *Doktorarbeit*.
- Benecke, H., R. Luhrmann, et al. (2005). "The U11/U12 snRNP 65K protein acts as a molecular bridge, binding the U12 snRNA and U11-59K protein." *EMBO J* **24**(17): 3057-69.
- Bentley, R. C. and J. D. Keene (1991). "Recognition of U1 and U2 small nuclear RNAs can be altered by a 5-amino-acid segment in the U2 small nuclear ribonucleoprotein particle (snRNP) B" protein and through interactions with U2 snRNP-A' protein." *Mol Cell Biol* **11**(4): 1829-39.
- Berget, S. M. (1995). "Exon recognition in vertebrate splicing." *J Biol Chem* **270**(6): 2411-4.
- Berglund, J. A., N. Abovich, et al. (1998). "A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition." *Genes Dev* **12**(6): 858-67.
- Bernstein, H. S. and S. R. Coughlin (1997). "Pombe Cdc5-related protein. A putative human transcription factor implicated in mitogen-activated signaling." *J Biol Chem* **272**(9): 5833-7.
- Bessonov, S., M. Anokhina, et al. (2008). "Isolation of an active step I spliceosome and composition of its RNP core." *Nature* **452**(7189): 846-50.
- Blakaj, D. M., K. J. McConnell, et al. (2001). "Molecular dynamics and thermodynamics of protein-RNA interactions: mutation of a conserved aromatic residue modifies stacking interactions and structural adaptation in the U1A-stem loop 2 RNA complex." *J Am Chem Soc* **123**(11): 2548-51.
- Blum, H., H. Beier, et al. (1987). "Improved silver staining of plant proteins, RNA and DNA polyacrylamid gels." *Elektrophoresis* **8**: 93-99.

- Boelens, W., D. Scherly, et al. (1991). "A weak interaction between the U2A' protein and U2 snRNA helps to stabilize their complex with the U2B" protein." Nucleic Acids Res **19**(3): 455-60.
- Boelens, W., D. Scherly, et al. (1991). "Analysis of in vitro binding of U1-A protein mutants to U1 snRNA." Nucleic Acids Res **19**(17): 4611-8.
- Boelens, W. C., E. J. Jansen, et al. (1993). "The human U1 snRNP-specific U1A protein inhibits polyadenylation of its own pre-mRNA." Cell **72**(6): 881-92.
- Boudrez, A., M. Beullens, et al. (2000). "NIPP1-mediated interaction of protein phosphatase-1 with CDC5L, a regulator of pre-mRNA splicing and mitotic entry." J Biol Chem **275**(33): 25411-7.
- Branlant, C., A. Krol, et al. (1982). "U2 RNA shares a structural domain with U1, U4, and U5 RNAs." EMBO J **1**(10): 1259-65.
- Brosi, R., H. P. Hauri, et al. (1993). "Separation of splicing factor SF3 into two components and purification of SF3a activity." J Biol Chem **268**(23): 17640-6.
- Brow, D. A. (2002). "Allosteric cascade of spliceosome activation." Annu Rev Genet **36**: 333-60.
- Brow, D. A. and C. Guthrie (1988). "Spliceosomal RNA U6 is remarkably conserved from yeast to mammals." Nature **334**(6179): 213-8.
- Brunger, A. T., P. D. Adams, et al. (1998). "Recent developments for the efficient crystallographic refinement of macromolecular structures." Curr Opin Struct Biol **8**(5): 606-11.
- Burge, C. B., R. A. Padgett, et al. (1998). "Evolutionary fates and origins of U12-type introns." Mol Cell **2**(6): 773-85.
- Burge, C. B., T. Tuschl, et al. (1999). Splicing of Precursors to mRNAs by the Spliceosome. The RNA world II. Gesteland, Cech and Atkins. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Pres: 525-560.
- Burns, C. G., R. Ohi, et al. (1999). "Evidence that Myb-related CDC5 proteins are required for pre-mRNA splicing." Proc Natl Acad Sci U S A **96**(24): 13789-94.
- Caspary, F. and B. Seraphin (1998). "The yeast U2A/U2B complex is required for pre-spliceosome formation." EMBO J **17**(21): 6348-58.
- CCP4-program-suite, T. (1994). "The CCP4 suite: programs for protein crystallography." Acta Crystallogr D Biol Crystallogr **50**(Pt 5): 760-3.
- Cech, T. R. (1990). "Nobel lecture. Self-splicing and enzymatic activity of an intervening sequence RNA from Tetrahymena." Biosci Rep **10**(3): 239-61.
- Cech, T. R. (1990). "Self-splicing of group I introns." Annu Rev Biochem **59**: 543-68.
- Chan, S. P. and S. C. Cheng (2005). "The Prp19-associated complex is required for specifying interactions of U5 and U6 with pre-mRNA during spliceosome activation." J Biol Chem **280**(35): 31190-9.
- Chan, S. P., D. I. Kao, et al. (2003). "The Prp19p-associated complex in spliceosome activation." Science **302**(5643): 279-82.
- Chen, C. H., D. I. Kao, et al. (2006). "Functional links between the Prp19-associated complex, U4/U6 biogenesis, and spliceosome recycling." RNA **12**(5): 765-74.
- Chen, C. H., W. C. Yu, et al. (2002). "Functional and physical interactions between components of the Prp19p-associated complex." Nucleic Acids Res **30**(4): 1029-37.
- Chen, H. R., S. P. Jan, et al. (1998). "Snt309p, a component of the Prp19p-associated complex that interacts with Prp19p and associates with the spliceosome simultaneously with or immediately after dissociation of U4 in the same manner as Prp19p." Mol Cell Biol **18**(4): 2196-204.
- Chen, H. R., T. Y. Tsao, et al. (1999). "Snt309p modulates interactions of Prp19p with its associated components to stabilize the Prp19p-associated complex essential for pre-mRNA splicing." Proc Natl Acad Sci U S A **96**(10): 5406-11.
- Chou, C. C., F. Forouhar, et al. (2003). "Crystal structure of the C-terminal 10-kDa subdomain of Hsc70." J Biol Chem **278**(32): 30311-6.



- Cornvik, T., S. L. Dahlroth, et al. (2006). "An efficient and generic strategy for producing soluble human proteins and domains in *E. coli* by screening construct libraries." Proteins **65**(2): 266-73.
- Cramer, P., D. A. Bushnell, et al. (2001). "Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution." Science **292**(5523): 1863-76.
- Cremer, T., G. Kreth, et al. (2000). "Chromosome territories, interchromatin domain compartment, and nuclear matrix: an integrated view of the functional nuclear architecture." Crit Rev Eukaryot Gene Expr **10**(2): 179-212.
- De Guzman, R. N., R. B. Turner, et al. (1998). "Protein-RNA recognition." Biopolymers **48**(2-3): 181-95.
- Deckert, J., K. Hartmuth, et al. (2006). "Protein composition and electron microscopy structure of affinity-purified human spliceosomal B complexes isolated under physiological conditions." Mol Cell Biol **26**(14): 5528-43.
- DeLucas, L. J., T. L. Bray, et al. (2003). "Efficient protein crystallization." J Struct Biol **142**(1): 188-206.
- Derewenda, Z. S. and P. G. Vekilov (2006). "Entropy and surface engineering in protein crystallization." Acta Crystallogr D Biol Crystallogr **62**(Pt 1): 116-24.
- Dietrich, R. C., R. Inorvaia, et al. (1997). "Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns." Mol Cell **1**(1): 151-60.
- Dominguez, C. and F. H. Allain (2006). "NMR structure of the three quasi RNA recognition motifs (qRRMs) of human hnRNP F and interaction studies with Bcl-x G-tract RNA: a novel mode of RNA recognition." Nucleic Acids Res **34**(13): 3634-45.
- Du, H. and M. Rosbash (2002). "The U1 snRNP protein U1C recognizes the 5' splice site in the absence of base pairing." Nature **419**(6902): 86-90.
- Emsley, P. and K. Cowtan (2004). "Coot: model-building tools for molecular graphics." Acta Crystallogr D Biol Crystallogr **60**(Pt 12 Pt 1): 2126-32.
- Fischer, U. and R. Luhrmann (1990). "An essential signaling role for the m3G cap in the transport of U1 snRNP to the nucleus." Science **249**(4970): 786-90.
- Frilander, M. J. and J. A. Steitz (1999). "Initial recognition of U12-dependent introns requires both U11/5' splice-site and U12/branchpoint interactions." Genes Dev **13**(7): 851-63.
- Frilander, M. J. and J. A. Steitz (2001). "Dynamic exchanges of RNA interactions leading to catalytic core formation in the U12-dependent spliceosome." Mol Cell **7**(1): 217-26.
- Fu, X. D. (1993). "Specific commitment of different pre-mRNAs to splicing by single SR proteins." Nature **365**(6441): 82-5.
- Gilliland, G. L., M. Tung, et al. (1994). "Biological Macromolecule Crystallization Database, Version 3.0: new features, data and the NASA archive for protein crystal growth data." Acta Crystallogr D Biol Crystallogr **50**(Pt 4): 408-13.
- Gilliland, G. L., M. Tung, et al. (2002). "The Biological Macromolecule Crystallization Database: crystallization procedures and strategies." Acta Crystallogr D Biol Crystallogr **58**(Pt 6 No 1): 916-20.
- Golas, M. M., B. Sander, et al. (2003). "Molecular architecture of the multiprotein splicing factor SF3b." Science **300**(5621): 980-4.
- Gotzmann, J., C. Gerner, et al. (2000). "hnRNP 200: a novel human common nuclear matrix protein combining structural and regulatory functions." Exp Cell Res **261**(1): 166-79.
- Gozani, O., R. Feld, et al. (1996). "Evidence that sequence-independent binding of highly conserved U2 snRNP proteins upstream of the branch site is required for assembly of spliceosomal complex A." Genes Dev **10**(2): 233-43.
- Grillari, J., P. Ajuh, et al. (2005). "SNEV is an evolutionarily conserved splicing factor whose oligomerization is necessary for spliceosome assembly." Nucleic Acids Res **33**(21): 6868-83.
- Gunderson, S. I., K. Beyer, et al. (1994). "The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A) polymerase." Cell **76**(3): 531-41.
- Guthrie, C. and B. Patterson (1988). "Spliceosomal snRNAs." Annu Rev Genet **22**: 387-419.

- Hall, S. L. and R. A. Padgett (1994). "Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites." *J Mol Biol* **239**(3): 357-65.
- Hall, S. L. and R. A. Padgett (1996). "Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns." *Science* **271**(5256): 1716-8.
- Hamm, J., N. A. Dathan, et al. (1990). "Multiple domains of U1 snRNA, including U1 specific protein binding sites, are required for splicing." *EMBO J* **9**(4): 1237-44.
- Hart, D. J. and F. Tarendeau (2006). "Combinatorial library approaches for improving soluble protein expression in Escherichia coli." *Acta Crystallogr D Biol Crystallogr* **62**(Pt 1): 19-26.
- Hartmuth, K., H. Urlaub, et al. (2002). "Protein composition of human prespliceosomes isolated by a tobramycin affinity-selection method." *Proc Natl Acad Sci U S A* **99**(26): 16719-24.
- Hastings, M. L. and A. R. Krainer (2001). "Functions of SR proteins in the U12-dependent AT-AC pre-mRNA splicing pathway." *Rna* **7**(3): 471-82.
- Heinrichs, V., M. Bach, et al. (1990). "U1-specific protein C is required for efficient complex formation of U1 snRNP with a 5' splice site." *Mol Biol Rep* **14**(2-3): 165.
- Hermann, H., P. Fabrizio, et al. (1995). "snRNP Sm proteins share two evolutionarily conserved sequence motifs which are involved in Sm protein-protein interactions." *EMBO J* **14**(9): 2076-88.
- Hirayama, T. and K. Shinozaki (1996). "A cdc5+ homolog of a higher plant, Arabidopsis thaliana." *Proc Natl Acad Sci U S A* **93**(23): 13371-6.
- Holm, L. and C. Sander (1993). "Protein structure comparison by alignment of distance matrices." *J Mol Biol* **233**(1): 123-38.
- Howard, P. K., K. G. Ahern, et al. (1988). "Establishment of a transient expression system for Dictyostelium discoideum." *Nucleic Acids Res* **16**(6): 2613-23.
- Huang, Y. and J. A. Steitz (2005). "SRprises along a messenger's journey." *Mol Cell* **17**(5): 613-5.
- Jankowsky, E., M. E. Fairman, et al. (2005). "RNA helicases: versatile ATP-driven nanomotors." *J Nanosci Nanotechnol* **5**(12): 1983-9.
- Jankowsky, E., C. H. Gross, et al. (2001). "Active disruption of an RNA-protein interaction by a DEXH/D RNA helicase." *Science* **291**(5501): 121-5.
- Jiang, J., K. Prasad, et al. (2005). "Structural basis of interdomain communication in the Hsc70 chaperone." *Mol Cell* **20**(4): 513-24.
- Jurica, M. S., D. Sousa, et al. (2004). "Three-dimensional structure of C complex spliceosomes by electron microscopy." *Nat Struct Mol Biol* **11**(3): 265-9.
- Kambach, C., S. Walke, et al. (1999). "Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs." *Cell* **96**(3): 375-87.
- Kastner, B. (1998). Purification and Electron Microscopy of spliceosomal snRNPs. *RNP particles, Splicing and Autoimmune Diseases*. J. Schenkel. Berlin Heidelberg, Springer Verlag: 95-140.
- Kastner, B., N. Fischer, et al. (2008). "GraFix: sample preparation for single-particle electron cryomicroscopy." *Nat Methods* **5**(1): 53-5.
- Katsamba, P. S., M. Bayramyan, et al. (2002). "Complex role of the beta 2-beta 3 loop in the interaction of U1A with U1 hairpin II RNA." *J Biol Chem* **277**(36): 33267-74.
- Katsamba, P. S., D. G. Myszka, et al. (2001). "Two functionally distinct steps mediate high affinity binding of U1A protein to U1 hairpin II RNA." *J Biol Chem* **276**(24): 21476-81.
- Kohtz, J. D., S. F. Jamison, et al. (1994). "Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors." *Nature* **368**(6467): 119-24.
- Kolossova, I. and R. A. Padgett (1997). "U11 snRNA interacts in vivo with the 5' splice site of U12-dependent (AU-AC) pre-mRNA introns." *Rna* **3**(3): 227-33.
- Kramer, A. (1996). "The structure and function of proteins involved in mammalian pre-mRNA splicing." *Annu Rev Biochem* **65**: 367-409.

- Kranz, J. K. and K. B. Hall (1998). "RNA binding mediates the local cooperativity between the beta-sheet and the C-terminal tail of the human U1A RBD1 protein." J Mol Biol **275**(3): 465-81.
- Kranz, J. K. and K. B. Hall (1999). "RNA recognition by the human U1A protein is mediated by a network of local cooperative interactions that create the optimal binding surface." J Mol Biol **285**(1): 215-31.
- Kranz, J. K., J. Lu, et al. (1996). "Contribution of the tyrosines to the structure and function of the human U1A N-terminal RNA binding domain." Protein Sci **5**(8): 1567-83.
- Kundrot, C. E. (2004). "Which strategy for a protein crystallization project?" Cell Mol Life Sci **61**(5): 525-36.
- Kunkel, G. R., R. L. Maser, et al. (1986). "U6 small nuclear RNA is transcribed by RNA polymerase III." Proc Natl Acad Sci U S A **83**(22): 8575-9.
- Laemmli, U. K. (1970). "Cleavage of structural proteins during the assembly of the head of bacteriophage T4." Nature **227**(5259): 680-5.
- Laird-Offringa, I. A. and J. G. Belasco (1995). "Analysis of RNA-binding proteins by in vitro genetic selection: identification of an amino acid residue important for locking U1A onto its RNA target." Proc Natl Acad Sci U S A **92**(25): 11859-63.
- Lamond, A. I., M. M. Konarska, et al. (1988). "Spliceosome assembly involves the binding and release of U4 small nuclear ribonucleoprotein." Proc Natl Acad Sci U S A **85**(2): 411-5.
- Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.
- Law, M. J., E. J. Chambers, et al. (2005). "Kinetic analysis of the role of the tyrosine 13, phenylalanine 56 and glutamine 54 network in the U1A/U1 hairpin II interaction." Nucleic Acids Res **33**(9): 2917-28.
- Law, M. J., M. E. Linde, et al. (2006). "The role of positively charged amino acids and electrostatic interactions in the complex of U1A protein and U1 hairpin II RNA." Nucleic Acids Res **34**(1): 275-85.
- Lei, X. H., X. Shen, et al. (2000). "Human Cdc5, a regulator of mitotic entry, can act as a site-specific DNA binding protein." J Cell Sci **113 Pt 24**: 4523-31.
- Levine, A. and R. Durbin (2001). "A computational scan for U12-dependent introns in the human genome sequence." Nucleic Acids Res **29**(19): 4006-13.
- Lopez, P. J. and B. Seraphin (2000). "YIDB: the Yeast Intron DataBase." Nucleic Acids Res **28**(1): 85-6.
- Lu, X. and R. J. Legerski (2007). "The Prp19/Pso4 core complex undergoes ubiquitylation and structural alterations in response to DNA damage." Biochem Biophys Res Commun **354**(4): 968-74.
- Luhrmann, R., B. Kastner, et al. (1990). "Structure of spliceosomal snRNPs and their role in pre-mRNA splicing." Biochim Biophys Acta **1087**(3): 265-92.
- Lutzmann, M., R. Kunze, et al. (2002). "Modular self-assembly of a Y-shaped multiprotein complex from seven nucleoporins." EMBO J **21**(3): 387-97.
- Mahajan, K. N. and B. S. Mitchell (2003). "Role of human Pso4 in mammalian DNA repair and association with terminal deoxynucleotidyl transferase." Proc Natl Acad Sci U S A **100**(19): 10746-51.
- Makarov, E. M., O. V. Makarova, et al. (2002). "Small nuclear ribonucleoprotein remodeling during catalytic activation of the spliceosome." Science **298**(5601): 2205-8.
- Makarova, O. V., E. M. Makarov, et al. (2004). "A subset of human 35S U5 proteins, including Prp19, function prior to catalytic step 1 of splicing." Embo J **23**(12): 2381-91.
- Maniatis, T. and R. Reed (2002). "An extensive network of coupling among gene expression machines." Nature **416**(6880): 499-506.
- Maris, C., C. Dominguez, et al. (2005). "The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression." Febs J **272**(9): 2118-31.

- Maschhoff, K. L. and R. A. Padgett (1993). "The stereochemical course of the first step of pre-mRNA splicing." Nucleic Acids Res **21**(23): 5456-62.
- Massenet, S., A. Mougin, et al. (1998). Posttranscriptional modifications in the U small nuclear RNAs. The Modification and Editing of RNA. H. Grosjean and R. Benner. Washington DC, ASM Press: 201-227.
- Mattaj, I. W. (1986). "Cap trimethylation of U snRNA is cytoplasmic and dependent on U snRNP protein binding." Cell **46**(6): 905-11.
- Mattaj, I. W. and E. M. De Robertis (1985). "Nuclear segregation of U2 snRNA requires binding of specific snRNP proteins." Cell **40**(1): 111-8.
- Mazza, C., A. Segref, et al. (2002). "Large-scale induced fit recognition of an m(7)GpppG cap analogue by the human nuclear cap-binding complex." EMBO J **21**(20): 5548-57.
- McDonald, W. H., R. Ohi, et al. (1999). "Myb-related fission yeast cdc5p is a component of a 40S snRNP-containing complex and is essential for pre-mRNA splicing." Mol Cell Biol **19**(8): 5352-62.
- McPherson, A. (1999). Crystallization of Biological Macromolecules. New York, Cold Spring Harbor Laboratory Press.
- Meszáros, B., P. Tompa, et al. (2007). "Molecular principles of the interactions of disordered proteins." J Mol Biol **372**(2): 549-61.
- Michel, F. and A. Jacquier (1987). "Long-range intron-exon and intron-intron pairings involved in self-splicing of class II catalytic introns." Cold Spring Harb Symp Quant Biol **52**: 201-12.
- Montzka, K. A. and J. A. Steitz (1988). "Additional low-abundance human small nuclear ribonucleoproteins: U11, U12, etc." Proc Natl Acad Sci U S A **85**(23): 8885-9.
- Moore, M. J., C. C. Query, et al. (1993). Splicing of precursors to mRNA by the spliceosome. The RNA World. A. Gesteland. Cold Spring Harbor, New York, Cold Spring Harbor Laboratory Press: 303-357.
- Moore, M. J. and P. A. Sharp (1993). "Evidence for two active sites in the spliceosome provided by stereochemistry of pre-mRNA splicing." Nature **365**(6444): 364-8.
- Murshudov, G. N., A. A. Vagin, et al. (1997). "Refinement of macromolecular structures by the maximum-likelihood method." Acta Crystallogr D Biol Crystallogr **53**(Pt 3): 240-55.
- Nagai, K. and I. W. Mattaj (1994). RNA-Protein interactions. K. Nagai and I. W. Mattaj. Oxford, Oxford University Press: 150-177.
- Nagai, K., C. Oubridge, et al. (1990). "Crystal structure of the RNA-binding domain of the U1 small nuclear ribonucleoprotein A." Nature **348**(6301): 515-20.
- Nemeth, K., K. Salchert, et al. (1998). "Pleiotropic control of glucose and hormone responses by PRL1, a nuclear WD protein, in Arabidopsis." Genes Dev **12**(19): 3059-73.
- Neubauer, G., A. King, et al. (1998). "Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex." Nat Genet **20**(1): 46-50.
- Newman, A. J. (1997). "The role of U5 snRNP in pre-mRNA splicing." EMBO J **16**(19): 5797-800.
- Nurse, P., P. Thuriaux, et al. (1976). "Genetic control of the cell division cycle in the fission yeast *Schizosaccharomyces pombe*." Mol Gen Genet **146**(2): 167-78.
- Ohi, M. D. and K. L. Gould (2002). "Characterization of interactions among the Cef1p-Prp19p-associated splicing complex." RNA **8**(6): 798-815.
- Ohi, M. D., A. J. Link, et al. (2002). "Proteomics analysis reveals stable multiprotein complexes in both fission and budding yeasts containing Myb-related Cdc5p/Cef1p, novel pre-mRNA splicing factors, and snRNAs." Mol Cell Biol **22**(7): 2011-24.
- Ohi, M. D., L. Ren, et al. (2007). "Structural characterization of the fission yeast U5.U2/U6 spliceosome complex." Proc Natl Acad Sci U S A **104**(9): 3195-200.
- Ohi, M. D., C. W. Vander Kooi, et al. (2003). "Structural insights into the U-box, a domain associated with multi-ubiquitination." Nat Struct Biol **10**(4): 250-5.
- Ohi, M. D., C. W. Vander Kooi, et al. (2005). "Structural and functional analysis of essential pre-mRNA splicing factor Prp19p." Mol Cell Biol **25**(1): 451-60.

- Ohi, R., A. Feoktistova, et al. (1998). "Myb-related *Schizosaccharomyces pombe* cdc5p is structurally and functionally conserved in eukaryotes." *Mol Cell Biol* **18**(7): 4097-108.
- Ohi, R., D. McCollum, et al. (1994). "The *Schizosaccharomyces pombe* cdc5+ gene encodes an essential protein with homology to c-Myb." *EMBO J* **13**(2): 471-83.
- Oubridge, C., N. Ito, et al. (1994). "Crystal structure at 1.92 Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin." *Nature* **372**(6505): 432-8.
- Oubridge, C., N. Ito, et al. (1995). "Crystallisation of RNA-protein complexes. II. The application of protein engineering for crystallisation of the U1A protein-RNA complex." *J Mol Biol* **249**(2): 409-23.
- Padgett, R. A., M. Podar, et al. (1994). "The stereochemical course of group II intron self-splicing." *Science* **266**(5191): 1685-8.
- Padgett, R. A. and G. C. Shukla (2002). "A revised model for U4atac/U6atac snRNA base pairing." *Rna* **8**(2): 125-8.
- Patel, A. A., M. McCarthy, et al. (2002). "The splicing of U12-type introns can be a rate-limiting step in gene expression." *EMBO J* **21**(14): 3804-15.
- Patel, A. A. and J. A. Steitz (2003). "Splicing double: insights from the second spliceosome." *Nat Rev Mol Cell Biol* **4**(12): 960-70.
- Pedelacq, J. D., E. Piltch, et al. (2002). "Engineering soluble proteins for structural genomics." *Nat Biotechnol* **20**(9): 927-32.
- Pena, V., S. Liu, et al. (2007). "Structure of a multipartite protein-protein interaction domain in splicing factor prp8 and its link to retinitis pigmentosa." *Mol Cell* **25**(4): 615-24.
- Perez Canadillas, J. M. and G. Varani (2003). "Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein." *EMBO J* **22**(11): 2821-30.
- Potashkin, J., D. Kim, et al. (1998). "Cell-division-cycle defects associated with fission yeast pre-mRNA splicing mutants." *Curr Genet* **34**(3): 153-63.
- Price, S. R., P. R. Evans, et al. (1998). "Crystal structure of the spliceosomal U2B"-U2A" protein complex bound to a fragment of U2 small nuclear RNA." *Nature* **394**(6694): 645-50.
- Query, C. C., M. J. Moore, et al. (1994). "Branch nucleophile selection in pre-mRNA splicing: evidence for the bulged duplex model." *Genes Dev* **8**(5): 587-97.
- Query, C. C., S. A. Strobel, et al. (1996). "Three recognition events at the branch-site adenine." *EMBO J* **15**(6): 1392-402.
- Raker, V. A., K. Hartmuth, et al. (1999). "Spliceosomal U snRNP core assembly: Sm proteins assemble onto an Sm site RNA nonanucleotide in a specific and thermodynamically stable manner." *Mol Cell Biol* **19**(10): 6554-65.
- Reddy, R., D. Henning, et al. (1987). "The capped U6 small nuclear RNA is transcribed by RNA polymerase III." *J Biol Chem* **262**(1): 75-81.
- Reed, R. (1989). "The organization of 3' splice-site sequences in mammalian introns." *Genes Dev* **3**(12B): 2113-23.
- Reed, R. (1996). "Initial splice-site recognition and pairing during pre-mRNA splicing." *Curr Opin Genet Dev* **6**(2): 215-20.
- Reed, R. and L. Palandjian (1997). Spliceosome assembly. *Eukaryotic mRNA processing*. Oxford, Oxford IRL Press: 1-37 and 103-129.
- Rinke, J., B. Appel, et al. (1985). "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**(4): 721-31.
- Ruby, S. W. and J. Abelson (1991). "Pre-mRNA splicing in yeast." *Trends Genet* **7**(3): 79-85.
- Sambrook, J., E. F. Fritsch, et al. (1989). *Molecular cloning-A Laboratory Manual*. Cold Spring Harbour, New York, Cold Spring Harbour Laboratory Press.
- Scherly, D., W. Boelens, et al. (1990). "Binding specificity determinants of U1A and U2B" proteins." *Mol Biol Rep* **14**(2-3): 181-2.

- Scherly, D., W. Boelens, et al. (1990). "Major determinants of the specificity of interaction between small nuclear ribonucleoproteins U1A and U2B" and their cognate RNAs." Nature **345**(6275): 502-6.
- Schneider, C., R. A. Newman, et al. (1982). "A one-step purification of membrane proteins using a high efficiency immunomatrix." J Biol Chem **257**(18): 10766-9.
- Schneider, C., C. L. Will, et al. (2002). "Human U4/U6.U5 and U4atac/U6atac.U5 tri-snRNPs exhibit similar protein compositions." Mol Cell Biol **22**(10): 3219-29.
- Schuwirth, B. S., M. A. Borovinskaya, et al. (2005). "Structures of the bacterial ribosome at 3.5 Å resolution." Science **310**(5749): 827-34.
- Semisotnov, G. V., N. A. Rodionova, et al. (1991). "Study of the "molten globule" intermediate state in protein folding by a hydrophobic fluorescent probe." Biopolymers **31**(1): 119-28.
- Seraphin, B. (1995). "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**(9): 2089-98.
- Sharp, P. A. (1987). "Splicing of messenger RNA precursors." Science **235**(4790): 766-71.
- Shen, V. and M. Kiledjian (2006). "A view to a kill: structure of the RNA exosome." Cell **127**(6): 1093-5.
- Shin, C. and J. L. Manley (2002). "The SR protein SRp38 represses splicing in M phase cells." Cell **111**(3): 407-17.
- Singh, R. and R. Reddy (1989). "Gamma-monomethyl phosphate: a cap structure in spliceosomal U6 small nuclear RNA." Proc Natl Acad Sci U S A **86**(21): 8280-3.
- Skrisovska, L., C. F. Bourgeois, et al. (2007). "The testis-specific human protein RBMY recognizes RNA through a novel mode of interaction." EMBO Rep **8**(4): 372-9.
- Slabinski, L., L. Jaroszewski, et al. (2007). "The challenge of protein structure determination--lessons from structural genomics." Protein Sci **16**(11): 2472-82.
- Slabinski, L., L. Jaroszewski, et al. (2007). "XtalPred: a web server for prediction of protein crystallizability." Bioinformatics **23**(24): 3403-5.
- Smith, T. F., C. Gaitatzes, et al. (1999). "The WD repeat: a common architecture for diverse functions." Trends Biochem Sci **24**(5): 181-5.
- Sontheimer, E. J., P. M. Gordon, et al. (1999). "Metal ion catalysis during group II intron self-splicing: parallels with the spliceosome." Genes Dev **13**(13): 1729-41.
- Sontheimer, E. J. and J. A. Steitz (1993). "The U5 and U6 small nuclear RNAs as active site components of the spliceosome." Science **262**(5142): 1989-96.
- Sontheimer, E. J., S. Sun, et al. (1997). "Metal ion catalysis during splicing of premessenger RNA." Nature **388**(6644): 801-5.
- Studier, F. W. (2005). "Protein production by auto-induction in high density shaking cultures." Protein Expr Purif **41**(1): 207-34.
- Stukenberg, P. T., K. D. Lustig, et al. (1997). "Systematic identification of mitotic phosphoproteins." Curr Biol **7**(5): 338-48.
- Tang, J., N. Abovich, et al. (1996). "Identification and characterization of a yeast gene encoding the U2 small nuclear ribonucleoprotein particle B" protein." Mol Cell Biol **16**(6): 2787-95.
- Tardiff, D. F. and M. Rosbash (2006). "Arrested yeast splicing complexes indicate stepwise snRNP recruitment during in vivo spliceosome assembly." RNA **12**(6): 968-79.
- Tarn, W. Y., C. H. Hsu, et al. (1994). "Functional association of essential splicing factor(s) with PRP19 in a protein complex." EMBO J **13**(10): 2421-31.
- Tarn, W. Y., K. R. Lee, et al. (1993). "Yeast precursor mRNA processing protein PRP19 associates with the spliceosome concomitant with or just after dissociation of U4 small nuclear RNA." Proc Natl Acad Sci U S A **90**(22): 10821-5.
- Tarn, W. Y. and J. A. Steitz (1996). "Highly diverged U4 and U6 small nuclear RNAs required for splicing rare AT-AC introns." Science **273**(5283): 1824-32.

- Tarn, W. Y. and J. A. Steitz (1996). "A novel spliceosome containing U11, U12, and U5 snRNPs excises a minor class (AT-AC) intron in vitro." *Cell* **84**(5): 801-11.
- Tarn, W. Y. and J. A. Steitz (1997). "Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge." *Trends Biochem Sci* **22**(4): 132-7.
- Toro, I., S. Thore, et al. (2001). "RNA binding in an Sm core domain: X-ray structure and functional analysis of an archaeal Sm protein complex." *EMBO J* **20**(9): 2293-303.
- Tsai, R. T., R. H. Fu, et al. (2005). "Spliceosome disassembly catalyzed by Prp43 and its associated components Ntr1 and Ntr2." *Genes Dev* **19**(24): 2991-3003.
- Tsai, W. Y., Y. T. Chow, et al. (1999). "Cef1p is a component of the Prp19p-associated complex and essential for pre-mRNA splicing." *J Biol Chem* **274**(14): 9455-62.
- Urlaub, H., V. A. Raker, et al. (2001). "Sm protein-Sm site RNA interactions within the inner ring of the spliceosomal snRNP core structure." *EMBO J* **20**(1-2): 187-96.
- Vander Kooi, C. W., M. D. Ohi, et al. (2006). "The Prp19 U-box crystal structure suggests a common dimeric architecture for a class of oligomeric E3 ubiquitin ligases." *Biochemistry* **45**(1): 121-30.
- Varani, G. and K. Nagai (1998). "RNA recognition by RNP proteins during RNA processing." *Annu Rev Biophys Biomol Struct* **27**: 407-45.
- Vidal, V. P., L. Verdone, et al. (1999). "Characterization of U6 snRNA-protein interactions." *RNA* **5**(11): 1470-81.
- Villa, T. and C. Guthrie (2005). "The Isy1p component of the NineTeen complex interacts with the ATPase Prp16p to regulate the fidelity of pre-mRNA splicing." *Genes Dev* **19**(16): 1894-904.
- Walter, T. S., C. Meier, et al. (2006). "Lysine methylation as a routine rescue strategy for protein crystallization." *Structure* **14**(11): 1617-22.
- Wang, Q., K. Hobbs, et al. (2003). "The Clf1p splicing factor promotes spliceosome assembly through N-terminal tetratricopeptide repeat contacts." *J Biol Chem* **278**(10): 7875-83.
- Will, C. L. and R. Luhrmann (2005). "Splicing of a rare class of introns by the U12-dependent spliceosome." *Biol Chem* **386**(8): 713-24.
- Will, C. L. and R. Luhrmann (2006). Spliceosome Structure and Function. *The RNA World*. R. F. Gesteland, T. R. Cech and J. F. Atkins. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press: 369-400.
- Will, C. L., C. Schneider, et al. (2004). "The human 18S U11/U12 snRNP contains a set of novel proteins not found in the U2-dependent spliceosome." *RNA* **10**(6): 929-41.
- Will, C. L., C. Schneider, et al. (2001). "A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site." *EMBO J* **20**(16): 4536-46.
- Will, C. L., C. Schneider, et al. (1999). "Identification of both shared and distinct proteins in the major and minor spliceosomes." *Science* **284**(5422): 2003-5.
- Wright, P. E. and H. J. Dyson (1999). "Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm." *J Mol Biol* **293**(2): 321-31.
- Wu, Q. and A. R. Krainer (1999). "AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes." *Mol Cell Biol* **19**(5): 3225-36.
- Wu, S., C. M. Romfo, et al. (1999). "Functional recognition of the 3' splice site AG by the splicing factor U2AF35." *Nature* **402**(6763): 832-5.
- Yang, Q. and E. Jankowsky (2005). "ATP- and ADP-dependent modulation of RNA unwinding and strand annealing activities by the DEAD-box protein DED1." *Biochemistry* **44**(41): 13591-601.
- Yean, S. L., G. Wuenschell, et al. (2000). "Metal-ion coordination by U6 small nuclear RNA contributes to catalysis in the spliceosome." *Nature* **408**(6814): 881-4.
- Young, J. C., J. M. Barral, et al. (2003). "More than folding: localized functions of cytosolic chaperones." *Trends Biochem Sci* **28**(10): 541-7.
- Yu, Y. T., M. D. Shu, et al. (1998). "Modifications of U2 snRNA are required for snRNP assembly and pre-mRNA splicing." *Embo J* **17**(19): 5783-95.

- Yu, Y. T. and J. A. Steitz (1997). "Site-specific crosslinking of mammalian U11 and u6atac to the 5' splice site of an AT-AC intron." Proc Natl Acad Sci U S A **94**(12): 6030-5.
- Zeng, Q. and K. B. Hall (1997). "Contribution of the C-terminal tail of U1A RBD1 to RNA recognition and protein stability." Rna **3**(3): 303-14.
- Zhang, N., R. Kaur, et al. (2005). "The Pso4 mRNA splicing and DNA repair complex interacts with WRN for processing of DNA interstrand cross-links." J Biol Chem **280**(49): 40559-67.
- Zieve, G. W. and R. A. Sauterer (1990). "Cell biology of the snRNP particles." Crit Rev Biochem Mol Biol **25**(1): 1-46.



## 8 Appendices

### 8.1 Details on Protein Expression

protein	amino acids	source organism	<i>E. coli</i> strain	vector	fusion tag	tested conditions
PRL1	1-514	human	Rosetta	pET-M11	N-His	LB-medium conditions
AD002	1-229	human	Rosetta	pET-M11	N-His	LB-medium, OD <sub>600</sub> 0.7, 30 °C, little bit soluble
AD002	1-229	human	Rosetta	pGEX6p_1	N-GST	LB-medium, OD <sub>600</sub> 0.7, 30 °C
Spf 27	1-225	human	Rosetta	pGEX6p_1	N-GST	LB-medium conditions
Spf 27	1-225	human	Rosetta	pET-M11	N-His	LB-medium conditions
PRL1 WD40	193-514	human	Rosetta/BL21	pGEX6p_1	N-GST	LB-medium conditions
PRL1 WD40	173-490	human	Rosetta/BL21	pET-M13	C-His	LB, OD <sub>600</sub> 1.7, auto-inducing medium
PRL1 WD40	173-498	human	Rosetta/BL21	pET-M13	C-His	LB-/auto-inducing Medium, 16 °C
PRL1 WD40	187-490	human	Rosetta/BL21	pET-M13	C-His	LB-/auto-inducing medium, 16 °C
PRL1 WD40	187-498	human	Rosetta/BL21	pET-M13	C-His	LB-/auto-inducing medium, 16 °C
PRL1 WD40	195-490	human	Rosetta/BL21	pET-M13	C-His	LB-/auto-inducing medium, 16 °C
PRL1 WD40	195-498	human	Rosetta/BL21	pET-M13	C-His	LB-/auto-inducing medium, 16 °C
Prp19p	1-503	<i>S. cerevisiae</i>	BL21	pRSF-duet	N-His	LB, small amount soluble, co-purifications
Prp19p	1-503	<i>S. cerevisiae</i>	BL21	pCDF-duet	N-His	LB, small amount soluble, co-purifications
Cef1p	1-590	<i>S. cerevisiae</i>	Rosetta/BL21	pET-/pCDF-duet	N-His/none	LB-medium, OD <sub>600</sub> 0.7
Prp46p	1-451	<i>S. cerevisiae</i>	Rosetta/BL21	pET-/pCDF-duet	N-His/none	LB-medium, OD <sub>600</sub> 0.7
Snt309p	1-175	<i>S. cerevisiae</i>	Rosetta/BL21	pET-/pCDF-duet	N-His/none	auto-inducing medium

**Table 15: Insoluble Proteins and Tested Conditions**

All proteins, which could not be obtained soluble, and the tested expression conditions are listed. Details for the expression conditions can be found in Table 5 and Table 6.

protein	amino acids	source organism	<i>E. coli</i> strain	vector	fusion tag	media-type	IPTG [mM]	OD <sub>600</sub> at induction	temperature [°C]	expression time [h]	comment
Prp19	1-504	human	Rosetta	pGEX6p_1	N-GST	LB	1	0.7-0.8	18	ON	low amount, co-purifications
Prp19	1-504	human	Rosetta	pET-28	N-His	LB	1	0.7-0.8	18	ON	lower amount, co-purifications
Prp19 coiled-coil	58-138	human	Rosetta	pGEX6p_1	N-GST	LB	1	0.7	18	ON	purified, crystallized
Prp19 coiled-coil	58-138	human	Rosetta	pET-M20	N-TrxAV N-His	LB	1	0.7	18	ON	low amount, tag not cleavable
Prp19 coiled-coil	58-138	human	Rosetta	pCDF-duet	N-His	LB	1	0.7	18	ON	low amount
CDC5	1-802	human	Rosetta	pET-M60	N-NusA	LB	1	0.7	18	ON	degradation
CDC5	1-802	human	Rosetta	pGEX6p_1	N-GST	LB	0.5	0.7	18	ON	low amount, degradation
CDC5	1-802	human	Rosetta	pET-M11	N-His	LB	1	1.7	18	ON	low amount, degradation
PRL1	1-514	human	Rosetta	pGEX6p_1	N-GST	LB	1	0.7	18	ON	low amount, contamination
Hsp73	1-646	human	Rosetta	pGEX6p_1	N-GST	LB	0.1-1	1.7	30	3	homolog structure known
AD002	1-229	human	Rosetta	pET-MZZ	N-ZZ/ N-His	LB	1	0.7	30	3	only ~10 % soluble
CDC5 C-term domain	702-802	human	Rosetta	pGEX6p_1	N-GST	LB	1	0.8	18	ON	soluble, tag cleavable
CDC5 C-term domain	702-802	human	Rosetta	pET-M11	N-His	LB	1	0.8	18	ON	only ~30 % soluble
CDC5 C-term domain	702-802	human	Rosetta	pET-MZZ	N-ZZ/ N-His	LB	1	0.8	18	ON	only ~30 % soluble

**Table 16: Soluble Proteins and Expression Conditions**

All proteins and protein domains, which could be obtained soluble, are listed. The expression condition that resulted in the highest protein yields is specified.

Table 16 (continued)

protein	amino acids	source organism	<i>E. coli</i> strain	vector	fusion tag	media-type	IPTG [mM]	OD <sub>600</sub> at induction	temperature [°C]	expression time [h]	comment
Prp19p	1-503	<i>S. cerevisiae</i>	Rosetta/BL21	pGEX6p_1	N-GST	auto-inducing	-	-	16	24	purified, crystallization
Prp19p U-box coiled-coil	1-132	<i>S. cerevisiae</i>	Rosetta2	pGEX6p_1	N-GST	auto-inducing	-	-	16	24	purified, crystallization
Prp19p U-box coiled-coil	1-141	<i>S. cerevisiae</i>	Rosetta2	pGEX6p_1	N-GST	auto-inducing	-	-	16	24	purified, crystallization
Prp19p U-box coiled-coil	1-149	<i>S. cerevisiae</i>	Rosetta2	pGEX6p_1	N-GST	auto-inducing	-	-	16	24	purified, crystallization

## 8.2 DNA-Oligonucleotides

protein name	Vector (series)	cleavage sites	forward primer	reverse primer
CDC5	pGEX6p_1	SmaI/NotI	TCCCCGGGGATGCCTCGAATTATGATCAAGGGGGGGGATG G	ATAAGAATCGGGCCGCTCAGAATTTTGACTTTAAAGTCTCT TTCTCC
CDC5	pET-M	BsmBI/SalI	GATCCGTCGCCATGCCTCGAATTATGATCAAGG	ACGGCTGACTCGAGATTTTGACTTTAAAGCTCTTTCTCC
CDC5	duet	AscI/PacI	TTGGCGGCTGACTACCGAAACCCTGTTATGTACAAAGGATGCG CTCGAATTATGATCAAG	CCTTAATTAATGATGGTGGATGTTAGAAATTTGACTTTA AAGTCTCTTTCTC
CDC5_C-term	pGEX6p_1	BamHI/EcoRI	CGCCGGATCCATGAGGGTCCACATGACGACAGAAGC	CCGGAATTCACAGAATTTTGACTTTAAAGTC
CDC5_C-term	pET-M	NcoI/SalI	CATGCCATGGAAGCCAAAGAGGGCTGCAAGATGG	ACGGCTGACTCGAGATTTTGACTTTAAAGCTCTTTCTCC
Hsp73	pGEX6p_1	BamHI/XhoI	CGCGGATCCATCTCCAAGGGACCTGCAAGTTGGTATTGATCTT GG	CCGGCTCGAGTTAATCAACCCTCTTCAATGGTGGGCC
PRL1	pGEX6p_1	BamHI/SmaI	CGCGGATCCATGTCCAAGGGACCTGCAAGTTGGTATTGATCTTG G	TCCCCCGGGTTAAAATCTCTTTCTCTTTGATAATTTCTGG
PRL1	pET-M	NcoI/Acc65I	CATGCCATGGTCGAGGGTACAGAAAACATTC	GATCGGTACCTTAAAATCTCTTTCTCTTTGATAATTTCTGG
PRL1	duet	SalI/NotI	ACGGCTCGACACTACCGAAACCCTGTTATGTACAAAGGATGGTC GAGGAAGGTACAGAAAAC	ATAAGAATCGGGCCGCTAAAATCTCTTTCTCTTTGATAAT TC
PRL1_WD40 <sup>173-490</sup>	pET-M	BspHI/Acc65I	CCATTTCATGAAGAACTCTGCCTGATGGCTAA	CTGGGTACCTTAGTGTGATGGTGGTGGTGTCTGTATACTT TAATGGTTTTATCAGCTTCA
PRL1_WD40 <sup>173-498</sup>	pET-M	BspHI/Acc65I	CCATTTCATGAAGAACTCTGCCTGATGGCTAA	CTGGGTACCTTAGTGTGATGGTGGTGGTGTCTTTCTGTG GCTGTGTCATCCT
PRL1_WD40 <sup>187-498</sup>	pET-M	BspHI/Acc65I	CCATTTCATGAAGAACTCTGCCTGATGGCTAA	CTGGGTACCTTAGTGTGATGGTGGTGGTGTCTGTATACTT TAATGGTTTTATCAGCTTCA
PRL1_WD40 <sup>187-498</sup>	pET-M	BspHI/Acc65I	CCATTTCATGAAGAACTCTGCCTGATGGCTAA	CTGGGTACCTTAGTGTGATGGTGGTGGTGTCTTTCTGTG GCTGTGTCATCCT
PRL1_WD40 <sup>195-490</sup>	pET-M	BspHI/Acc65I	CCATTTCATGAAGAACTCTGCCTGATGGCTAA	CTGGGTACCTTAGTGTGATGGTGGTGGTGTCTGTATACTT TAATGGTTTTATCAGCTTCA

Table 17: DNA-Oligonucleotides

DNA-Oligonucleotides, which were used for cloning of all constructs used in the course of this work. DNA-oligonucleotides for mutagenesis and sequencing are not listed.

Table 17 (continued)

protein name	Vector (series)	cleavage sites	forward primer	reverse primer
PRL1_WD40 <sup>198-498</sup>	pET-M	BspHI/Acc65I	CCATTTCATGAAACTCTACAGGGTTATCAGTGGCA	CTGGGTACCTTAGTGTGATGGTATGTTCTTCTGTG GCTGTGTCATCCT
PRL1_WD40 <sup>193-514</sup>	pGEX6p_1	BamHI/EcoRI	CGCGGATCCATGCCGTGGAAACTCTACAGGGTTATCAG	CCGGAATTCTTTAAAATCTCTTTCTTGTATAAATTTCTGG
Prp19	pGEX6p_1	EcoRI/XhoI	CCGGAATTCATGTCCTTAATCTGCTCCATCTCTAACG	CCGCTCGAGCTACAGGCTGTAGAACTTGAAGCTTCTGTCC
Prp19	pET-M	BspHI/XhoI	CATGTCATGAATGAGCCTAATCTGCTCCATCTCTAACG	ACGTACCTCGAGCTACAGGCTGTAGAACTTGAAGCTTCT
Prp19	duet	EcoRI/NotI	CCGGAATTCGACTACCCGAAACCGGTTATGTACAAGGTATGTCC CTAATCTGCTCCATCTCTAACG	AAGGAAAAAGCGGCCCTTACAGGCTGTAHAAC TTGAGG CTTCTG
Prp19_cc	pGEX6p_1	BamHI/EcoRI	CGCGGATCCATGGCTCACCCAATCCGGGCCCAAG	CCGGAATTCCTAAAGCCTGTGTTTCAAGGTAGCCAGAG
Prp19_cc	pET-M	NcoI/Acc65I	CATGCCATGGCTCACCCAATCCGGGCCCAAG	ACGTACGGTACCCTAAAGCCTGTGGTTTCAAGGTAGCCAGA G
Prp19_cc	duet	EcoRI/NotI	CGGAATTCGACTACCCGAAACCGGTTATCTAC AAGGTATGGCTC ACCCAATCCGGGCCCAAGCCTCCCTC	ATAAGAATCGGGCCGCTTAAAGCCTGTGGTTTCAAGGGTAGC CAGAG
AD002	pET-M	BsmBI/Acc65I	GTACCGTCTCCCATGACAACAGCAGCCAGGCCAACC	ACGTACGGTACCCTATTTAATATATTTCTCCATGAAC
Spt27	pET-M	NcoI/Acc65I	CATGCCATGGCGGGCACAGGTTTGGTGG	GATCGGTACCTCAGAAGTCTTGCCGGATGTTTTTC
Cef1p	duet	BamHI/HindIII	CGCGGATCCATGCCCGCCCGTACCAATATACGTGA	CCCAAGCTTTTATATGGAAAGAAAGAAATTTAGCATGGCTT G
Prp46p	duet	BamHI/HindIII	CGCGGATCCATGGACGGGAAATGATCACAAAATC	CCCAAGCTTTTAAAATCTTTTGGCCGCTTAAGTTGG
Prp19p	pGEX6p_1	BamHI/XhoI	CGCGGATCCATGCTTTGTGCTATTAGTGGGAAAAGTTCCCTAG	CCGCTCGAGTTAGGGTGTCAATGC AACAATATTGAAAC
Prp19p	duet	BamHI/HindIII	CGCGGATCCATGCTTTGTGCTATTAGTGGGAAAAGTTCCCTAG	CCCAAGCTTTTAGGGTGTCAATGC AACAATATTGAAAC

Table 17 (continued)

protein name	Vector (series)	cleavage sites	forward primer	reverse primer
P <sub>rip19p</sub> -U <sub>box-cc<sup>1-132</sup></sub>	pGEX6p_1	BamHI/XhoI	CGCGGATCCATGCTTTGTGCTATTAGTGGGAAA	CCGCTCGAGTTATTCGTTTTTCCATCAATAGTTGC
P <sub>rip19p</sub> -U <sub>box-cc<sup>1-141</sup></sub>	pGEX6p_1	BamHI/XhoI	CGCGGATCCATGCTTTGTGCTATTAGTGGGAAA	CCGCTCGAGTTATGAGGATTTGGGTAAATCCTTCG
P <sub>rip19p</sub> -U <sub>box-cc<sup>1-149</sup></sub>	pGEX6p_1	BamHI/XhoI	CGCGGATCCATGCTTTGTGCTATTAGTGGGAAA	CCGCTCGAGTTATCTCGTAATACTACCCGCTTGC
Snt309p	duet	NdeI/PacI	GGGAATTCATATGGACGGCCCTTAGCCTTTGTTG	CCTTAATTAATTATAACCTCTTTCTGTATATTTCCAAATCG
65K <sup>300-517</sup>	pGEX6p_1	BamHI/XhoI	GGATCCGACTCTGATGAAATGCC	GCCCTCGAGTCACAGCCCATGGTTTGGTG
65K <sup>404-505</sup>	pGEX6p_1	BamHI/XhoI	CGCGGATCCATGGAAACACTTTCAGTTTTTCAGAA	CCGCTCGAGTTATCTAGCAGATCGAGCAAACCTGAAC
65K <sup>411-505</sup>	pGEX6p_1	BamHI/XhoI	CGCGGATCCAGAAAGTTATGAACCCGGGTGAACC	CCGCTCGAGTTATCTAGCAGATCGAGCAAACCTGAAC
65K <sup>417-505</sup>	pGEX6p_1	BamHI/XhoI	CGCGGATCCGAACCAACTGTAGAATTTATGTAAGA	CCGCTCGAGTTATCTAGCAGATCGAGCAAACCTGAAC
65K <sup>300-505</sup>	pGEX6p_1	BamHI/XhoI	CGCGGATCCGACTCTGATGAAATGCCCTTCAGAATG	CCGCTCGAGTTATGGTCTAGCAGATCGAGCAAAC
65K <sup>300-501</sup>	pGEX6p_1	BamHI/XhoI	CGCGGATCCGACTCTGATGAAATGCCCTTCAGAATG	CCGCTCGAGTTAAGCAAACCTGAACCCACCATG

### 8.3 Protein and DNA Sequences

protein name	source organism	NCBI protein-GI	NCBI gene-ID	Swiss-Prot number
CDC5	human	11067747	NM_001253	Q99459
Prp19	human	7657381	NM_014502	Q9UMS4
Hsp73	human	5729877	NM_006597	P11142
PRL1	human	4505895	NM_002669	O43660
AD002	human	7705475	NM_016403	Q9P013
Spf27	human	5031653	NM_005872	O75934
Cef1p	<i>S. cerevisiae</i>	6323869	NC_001145	Q03654
Prp19p	<i>S. cerevisiae</i>	6322992	NC_001144	P32523
Prp46p	<i>S. cerevisiae</i>	6325106	NC_001148	Q12417
Snt309p	<i>S. cerevisiae</i>	6325358	NC_001148	Q06091
U11/U12 65K	human	16553747	AK057799	Q96LT9

**Table 18: Identification Numbers for Protein and DNA Sequences**

The protein identification-numbers (GI) and the gene identification-numbers (gene-ID) according to the NCBI-database are listed. Additionally the identification-numbers from the Swiss-Prot database are given.

## 8.4 Refinement Statistics for the 65K-cRRM Crystal Structure

Refinement	
Resolution (Å)	30.0-2.5 (2.56-2.50) <sup>a</sup>
Reflections	
Number	5283
Completeness (%)	99.7(99.5)
Test Set (%)	10.7
R <sub>work</sub> <sup>e</sup>	19.4 (23.6)
R <sub>free</sub> <sup>e</sup>	25.2 (35.8)
ESU (Å) <sup>f</sup>	0.184
Contents of A.U. <sup>g</sup>	
Protein Molecules/Residues/Atoms	1/120/982
Water Oxygens	84
Mean B-Factors (Å <sup>2</sup> )	
Wilson	33.8
Protein	32.8
Water	34.6
Ramachandran Plot <sup>g</sup>	
Preferred	96.6
Allowed	3.4
Disallowed	0
Rmsd <sup>h</sup> from Target Geometry	
Bond Lengths (Å)	0.008
Bond Angles (°)	1.180
Rmsd B-Factors (Å <sup>2</sup> )	
Main Chain Bonds	0.402
Main Chain Angles	0.654
Side Chain Bonds	1.118
Side Chain Angles	1.896

**Table 19: Statistics for the Refinement of the 65K-cRRM Crystal Structure**

Refinement Statistics for the 65K-cRRM crystal structure. <sup>a</sup> Data for the highest resolution shell in parentheses; <sup>b</sup>  $R_{\text{sym}}(I) = \sum_{\text{hkl}} \sum_i |I_i(\text{hkl}) - \langle I(\text{hkl}) \rangle| / \sum_{\text{hkl}} \sum_i |I_i(\text{hkl})|$ ; for  $n$  independent reflections and  $i$  observations of a given reflection;  $\langle I(\text{hkl}) \rangle$  – average intensity of the  $i$  observations; <sup>c</sup>  $R = \sum_{\text{hkl}} ||F_{\text{obs}}| - |F_{\text{calc}}|| / \sum_{\text{hkl}} |F_{\text{obs}}|$ ; R<sub>work</sub> – hkl ∉ T; R<sub>free</sub> – hkl ∈ T; T – test set; <sup>d</sup> ESU – estimated overall coordinate error based on maximum likelihood; <sup>e</sup> A.U. – asymmetric unit; <sup>f</sup> Calculated with Molprobit (http://molprobit.biochem.duke.edu/); <sup>g</sup> Rmsd – root-mean-square deviation.



## 8.5 Designing Targets for Structural Analyses

A structure of a protein like U11/12 65K or of a protein complex such as the Prp19-complex always provides valuable functional insights. However, obtaining such structures is not yet a routine process. It becomes even more difficult when many components build a dynamic macromolecular assembly as in the spliceosome or other macromolecular assemblies. Often a combination of electron microscopy, X-ray crystallography and nuclear magnetic resonance (NMR) helps to obtain a more complete picture of such macromolecular assemblies. To apply these methods, the target proteins or protein-complexes have to be obtained in sufficient amounts and a high degree of purity. They can either be purified from a native source or they can be produced recombinantly. Recombinant protein production is often used for crystallographic studies, because it usually yields higher amounts of protein (at lower costs). Especially for crystallographic purposes, the definition of globular domains is of great importance for the initial target design. The following chapter provides an overview on domain definition and target design techniques. Parameters influencing protein crystallization are explained and the proteins, which were investigated in the course of this thesis, are analyzed with respect to these parameters.

### Domain Definition and Target Design

Flexible regions often impede structural investigations, especially in the case of X-ray crystallography. Because they prohibit any regular contact in a crystal lattice and they do not contribute favorably to the entropy of crystallization, flexible protein regions often prevent crystal growth. In cases, in which a protein with flexible regions can be crystallized, the flexible parts may be missing in the electron density due to disorder. Therefore, the main focus in protein target design for crystallographic purposes is to define compactly folded domains, which can be produced in sufficient amounts and purity.

Putatively ordered and disordered regions in proteins can be predicted by bioinformatic approaches. Two principles underlie most of these disorder prediction programs: either they are based on sequence analysis or on energy calculations. In the first case, the occurrence of certain amino acids, which tend to be more frequent in disordered parts of a protein, is evaluated. Longer and charged amino acids (e.g. lysine or glutamate) as well as prolines or glycines can often be found in disordered regions. The second type of prediction programs tries to estimate the energetic contribution of certain residues to intra-molecular contacts. Mostly contacts with residues in direct neighborhood of the evaluated residue are evaluated. If the energy contribution reaches a negative value, the residue is likely to be involved in

contacts with its surrounding sequence and therefore it is likely to be part of a folded domain. Because the error rate of both type of programs is still fairly high, the most reliable predictions are obtained by applying a number of programs followed by manual inspection.

The standard way to detect domain borders experimentally is limited proteolysis. This method is based on the principle that proteases can only recognize and cleave their target sequences if these are not buried in a domain, but exposed in a loop or linker region. Typically, a protein of interest is digested with a set of proteases and the fragments are analyzed by mass spectrometry. The required amounts of protein are low, which permits application of this technique to characterize proteins that are available only in low yield. Defined domains can often be re-cloned and expressed with a higher yield.

Another experimental technique, the so called library-approach is mainly focused on the production of soluble protein, but also gives some information on domain structure. In library-approaches, the DNA-construct is randomly separated into smaller pieces and re-cloned into vectors, which carry a certain reporter system. Subsequently, the proteins are expressed and only the soluble proteins, which likely correspond to folded domains, are detected by the reporter (Cornvik, Dahlroth et al. 2006; Hart and Tarendeau 2006). Entire genomes have already been sampled in this way to detect protein domains, which are suitable for X-ray crystallography.

Apart from the definition of domain borders, the target protein can be further streamlined to increase the chances for crystallization. Novel methods in the target design or target modification are directed evolution, surface residue engineering and chemical modifications. Directed evolution is a library approach to design soluble constructs, which may have a higher likelihood of crystallization because of altered sequence properties (Pedelacq, Piltch et al. 2002). Typical surface engineering is aimed at reducing the surface entropy of a protein (Derewenda and Vekilov 2006). For this purpose large flexible surface residues (like lysine or glutamate) can be exchanged for small polar amino acids or disulfide bridges can be introduced. Contrary to surface engineering, chemical modifications are applied to the already purified sample often in a kind of "shotgun"-approach. Chemical modifications include lysine-actetylation and -methylation, deglycosylation or protein-protein and protein-ligand crosslinking. Lysine-methylation is performed routinely by structural genomics factories (Walter, Meier et al. 2006). The more hydrophobic nature of a methylated lysine may favor protein-protein interactions. Additionally the isoelectric point (pI) is decreased and therefore the chances for crystallization are increased, because proteins with a moderately acidic pI are more likely to crystallize (Slabinski, Jaroszewski et al. 2007).

A very good source of information on protein properties, which influence protein production and crystallization, was made available by structural genomics projects. The advantage of

these projects is that both successful as well as failed trials to produce and crystallize diverse proteins are listed in public databases. A summary of protein features, which influence either protein production or crystallization is given in Slabinski et al. (Slabinski, Jaroszewski et al. 2007). According to these authors, proteins with a medium sequence length (70-200 amino acids), a medium hydrophobicity (hydrophobicity index  $\sim 0-0.2$ ) and a moderately acidic character (pI 5-6) are the easiest to produce. The highest chances for crystallization are given with slightly acidic (pI  $\sim 6$ ) proteins of medium lengths and hydrophobicity indexes (also called Gravy indexes) of  $\sim 0.1$ . These values can be explained in the following way: a too high hydrophobicity would promote aggregation, whereas very low hydrophobicity or high charges would impede protein-protein interactions, which are necessary for crystal nucleation. In addition to this information, the authors found that the length of disordered stretches is more important for crystallization than the percentage of disorder. According to these parameters a server has very recently been implemented, which predicts the likelihood of a protein to crystallize (Slabinski, Jaroszewski et al. 2007). The most important factors, which influence crystallization and the crystallization likelihood prediction for the proteins, which were investigated in the course of this thesis, are given in Table 20. Only Prp19p and Snt309p have been worked on in structural genomics projects. Prp19p has been crystallized, but the work was stopped (most probably no diffraction-quality crystals were obtained). Snt309p was purified and is currently analyzed with 2D-NMR.

protein name	source	aa	amino acid range	MW [kDa]	pI	gravy index	longest disorder range[aa]	crystallization probability
CDC5	human	802	1-802	92.3	8.2	-0.96	104-213(110aa)	very difficult
CDC5_C-term	human	97	706-802	11.7	8.7	-1.05	758-769(12aa)	very difficult
Hsp73	human	646	1-646	70.9	5.4	-0.46	498-544(47aa)	very difficult
PRL1	human	514	1-514	57.2	9.2	-0.49	99-159(61aa)	very difficult
PRL1_WD40	human	319	173-490	35.5	9.4	-0.28	173(1aa)	average
PRL1_WD40	human	326	173-498	36.3	9.1	-0.31	308-311(5aa)	suboptimal
PRL1_WD40	human	305	187-490	34.0	9.2	-0.27	187(1aa)	average
PRL1_WD40	human	312	187-498	34.8	8.7	-0.30	308-311(5aa)	suboptimal
PRL1_WD40	human	322	193-514	36.0	9.1	-0.30	512-513(2aa)	suboptimal
PRL1_WD40	human	296	195-490	32.9	9.2	-0.21	0	suboptimal
PRL1_WD40	human	304	195-498	33.7	8.5	-0.26	494-498(5aa)	optimal
Prp19	human	504	1-504	55.2	6.1	-0.17	119-164(46aa)	very difficult
Prp19_cc	human	81	58-138	9.0	9.5	-0.29	66-74(9aa)	very difficult
AD002	human	229	1-229	26.6	5.6	-1.43	44-185(142aa)	very difficult
Spf27	human	225	1-225	26.1	5.5	-0.70	202-225(24aa)	very difficult
Cef1p	<i>S. cer.</i>	590	1-590	67.7	8.3	-0.93	227-394(168aa)	very difficult
Prp46p	<i>S. cer.</i>	451	1-451	50.7	6.8	-0.42	38-52(15aa)	suboptimal
Prp19p	<i>S. cer.</i>	503	1-503	56.6	5.0	-0.39	126-177(52aa)	very difficult
Prp19p_Ubox-cc	<i>S. cer.</i>	132	1-132	14.7	5.3	-0.26	54-79(25aa)	difficult
Prp19p_Ubox-cc	<i>S. cer.</i>	141	1-141	15.7	5.3	-0.35	54-79(25aa)	difficult
Prp19p_Ubox-cc	<i>S. cer.</i>	149	1-149	16.6	5.6	-0.33	54-79(25aa)	difficult
Snt309p	<i>S. cer.</i>	175	1-175	20.7	8.5	-0.90	71-76(6aa)	difficult
U11/U12 65K	human	517	1-517	58.6	7.6	-0.72	213-420(207aa)	very difficult
65K <sup>380-517</sup>	human	138	380-517	16.0	9.4	-0.79	8-41(34aa)	very difficult
65K <sup>404-505</sup>	human	102	404-505	11.8	9.7	-0.39	1-17(17aa)	difficult
65K <sup>411-505</sup>	human	95	411-505	10.9	9.8	-0.5	1-5(5aa)	difficult
65K <sup>417-505</sup>	human	89	417-505	10.3	9.9	-0.39	86-89(4aa)	difficult
65K <sup>380-506</sup>	human	127	380-506	14.7	9.1	-0.62	8-42(34aa)	very difficult
65K <sup>380-501</sup>	human	122	380-501	14.1	8.6	-0.57	8-42(34aa)	difficult

**Table 20: Protein sequence analysis**

Protein properties, which may influence crystallization, are shown for the proteins, which were investigated in the course of this thesis. *S. cerevisiae* is abbreviated as *S. cer.* The factors, which may have an adverse impact on crystallization, are displayed with a red background. The hydrophobicity index is called gravy index.

## 8.6 List of Abbreviations

aa	amino acid(s)
APS	ammoniumperoxodisulfate
ATP	adenosintriphosphate
bp	branchpoint
BSA	bovine serum albumin
C	cytosine
°C	degree celsius
cDNA	complementary DNA
<i>C. elegans</i>	<i>Caenorhabditis elegans</i>
Ci	curie
C-terminus	carboxy-terminus
d	desoxy
Da	dalton
<i>D. melanogaster</i>	<i>Drosophila melanogaster</i>
DMSO	dimethylsulfoxide
DNA	desoxyribonucleic acid
DNase	desoxyribonuclease
DTT	Dithiothreitol
<i>E. coli</i>	<i>Escherichia coli</i>
<i>e. g.</i>	<i>exempli gratia</i> (for the sake of example)
EM	electron microscopy
<i>et al.</i>	<i>et alii</i> (lateinisch für: und andere)
G	guanosine
g	gramm/centrifugal force
h	hour(s)
His-tag	hexa His-tag
hn	heterogenous nuclear
<i>i. e.</i>	<i>id est</i> (that is)
kDa	kilodalton
$K_{Dapp}$	apparent binding constant
l	liter
LB	Luria-Bertani-broth
M	molar [mol/l]
m <sup>3</sup> G	2,2,7-trimethylguanosine cap

---

m <sup>7</sup> G	N7-monomethylguanosine cap
min	minutes
mM	millimolar
MR	molrep
mRNA	messenger RNA
MW	molecular weight
nt	nukleotide
NTC	nineteen complex
N-terminus	amino-terminus
NTPs	nukleosite-5'-triphosphate
OD	optic density
oligo	oligonucleotide
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate buffered saline
Pfu	<i>Pyrococcus furiosus</i>
pH	inverse logarithmic representation of hydrogen proton (H <sup>+</sup> )-concentration
pre-mRNA	precursor-messenger RNA
Prp	pre-mRNA processing
R	purine-base
RBD	RNA-binding domain
rmsd	root mean square deviation
RNA	ribonucleic acid
RNAi	RNA-interference
RNasin	ribonuclease-inhibitor
RNP	ribonucleoprotein
rpm	rounds per minute
RRM	RNA recognition motif
RT	room temperature
S	Svedberg-Einheit [10 <sup>-13</sup> s]
s	second(s)
<i>S. cerevisiae</i>	<i>Saccharomyces cerevisiae</i>
SDS	sodium-dodecyl-sulfate
SDS-PAGE	SDS-polyacrylamide gel electrophoresis
sn	small nuclear
sno	small nucleolar

---

---

SR	serine-arginine-rich
ss	splice-site
TBE	tris-borate-EDTA-solution
TEMED	N,N,N',N'-tetramethylethylenediamine
TEV	tobacco etch virus
T <sub>m</sub>	melting temperature
Tris/HCl	tris-(hydroxymethyl)-aminomethan-hydrochloride
tRNA	transfer RNA
U	uracile
UK	United Kingdom
USA	United States of America
U snRNA	uridine-rich small nuclear RNA
U snRNP	uridine-rich small nuclear ribonucleoprotein
UTR	untranslated region
UV	ultra-violett
V	Volt
v/v	volume per volume
w/v	weight per volume
wt	wildtype
Y	pyrimidine-base

Units from the SI-system as well as SI-prefixes are not listed. Amino acids are abbreviated according to the standard three-letter or single-letter code. Nucleotides are abbreviated according to the standard-single letter code.

## 8.7 Acknowledgements

Most sincerely I would like to thank my supervisor Prof. Dr. Markus Wahl. His supervision was great during the whole PhD project. While giving me freedom in pursuing research, he always contributed inspiring ideas to promote the project. Markus door was never closed to discuss scientific as well as personal concerns.

Furthermore I would like to express my gratitude to Prof. Dr. Reinhard Lührman for assistance and financing.

Thanks to the Boehringer Ingelheim Fonds for a great PhD fellowship including hiking tours with most impressive scenery, support for several conferences and, most importantly, many nice people. In particular I would like to acknowledge to Dr. Herrmann Fröhlich, Monika Beutelspacher and Claudia Walther for their continuous engagement.

I am indebted to Prof. Dr. Wolfgang Wintermeyer not only for accepting to be my referee, but also for the guidance during my scientific career.

A special thanks goes to my colleagues Gert Weber and Simon Trowitzsch. At any time of the day, they had an open ear for scientific problems and they always tried to find a solution. Additionally, they often made me cheer up and they definitely formed my life during the time in Göttingen. Thanks for the great time!

I deeply acknowledge the support by all former and present members of my lab. Fruitful discussions and generous help in each and every question regarding science and organizational aspects helped to create a great working environment and a good atmosphere. I am especially grateful to Elke Penka for all her help with cloning and many nice talks. Without Christian Stegmann's help with computational problems, life would have been much harder. I am deeply thankful for that.

I am much obliged to Dr. Ralf Jauch, who always appreciated my work. Thanks for initiating the work in the PhD-committee and thanks for the cool car!



I would like to acknowledge all members from the Lührmann lab for offering their help, wherever it was needed. In this respect a big thanks goes to Nastaran Behzadnia, Heike Benecke, Berkhan Aikyldiz, Marieke van Santen, Michael Grote, Marc Schneider and Irina Häcker. I am grateful to Dr. Dmitry Agafonov for his collaboration. Thanks to Dr. Björn Sander and Dr. Monika Golas for the electron microscopy work. Furthermore I would like to acknowledge Cindy Will and Reinhard Rauhut for answering so many questions.

I am indebted for the great technical support by Marion Kilian, Irene Öchsner, Gaby Heine, Monika Raabe and Uwe Plessmann. But not only this, Marion Kilian, Irene Öchsner and Gaby Heine were always there for personal encouragement.

I would like to thank Juliane Moses for her kindful and helpful assistance in all organizational questions.

Thanks to my family and friends for their continuous support and encouragement. A special thanks goes to my mum for giving me personal strength and enthusiasm. I am also very grateful for Trudie's and Rolf's warm welcome. Thanks to my uncle and aunt: your sustained interest in my scientific education always motivated me to carry on. Thanks to Mike Turra for being my true best friend at all times!

My most heartfelt thanks goes to Steffen for his unconditional love. He always gave me joy and structure.

## 8.8 *Curriculum Vitae*

### Personal Details

---

Name: Catharina Netter  
Date of Birth: 26.09.1979  
Place of Birth: Hannover, Germany  
Marital Status: single

### Academic Education

---

2004 – present      PhD-thesis in Biochemistry, X-ray crystallography group,  
Max-Planck-Institute for Biophysical Chemistry,  
Göttingen, Germany  
Topic: “ Structural Analyses of Spliceosomal Proteins:  
The Prp19-Complex and the U11/U12 65K Protein”

2002 –2003      Diploma-thesis in Biochemistry, Division for Biological  
Chemistry and Molecular Microbiology, Wellcome Trust  
Biocentre, University of Dundee, United Kingdom  
Topic: “Chitin and Chitinases – a study on substrate  
binding, processivity and inhibitor design”

2000 – 2002      Graduate Studies in Biochemistry,  
University of Witten/Herdecke, Germany  
Degree: Diplom Biochemiker (MSc)

1998 – 2000      Undergraduate Studies in Biochemistry,  
University of Hannover, Germany

### School Education

---

1994 - 1998      Richarda-Huch-Gymnasium (Secondary School),  
Hannover, Germany

1990 - 1994      Gymnasium Sophienschule, Hannover, Germany

1988 - 1990      Orientierungsstufe Lüerstrasse, Hannover, Germany

1984 - 1988      Hinrich-Wilhelm-Kopf Grundschule (Primary School),  
Hannover, Germany

---

**Internships**

---

Jul 2001 – Oct 2001	Internship at the Burnham Institute, La Jolla, USA Topic: “Biocomputational Studies on the Addiction Module MazE/MazF”
Aug 1998	Internship at the Institute for Pharmacology and Toxicology, Phillips-Universität Marburg, Germany

---

**Stipends**

---

Jan 2005 – Dec 2006	PhD-Stipend, Boehringer-Ingelheim Fonds
Mar 2003 – Aug 2003	Stipend for Travel Expenses for the stay in UK, Heinrich-Hertz Stiftung
Dec 2002 – Feb 2003	Short Time Stipend for the stay in UK, DAAD
Jul 2001 – Sep 2001	Stipend for Travel Expenses for the stay in USA Daniela und Jürgen Westphal Stiftung

---

**Publications**

---

Vaaje-Kolstad G, Vasella A, Peter MG, Netter C, Houston DR, Westereng B, Eijsink VG, van Aalten DM (2004), Interactions of a family 18 chitinase with the designed inhibitor HM508 and its degradation product, chitobiono-delta lactone, *J. Biol. Chem.* 279:3612-3619

Jauch R, Cho MK, Jakel S, Netter C, Schreiter K, Aicher B, Zweckstetter M, Jackle H, Wahl MC. Mitogen-activated protein kinases interacting kinases are autoinhibited by a reprogrammed activation segment, *EMBO J.* 2006; 25(17):4020-32

Jauch R, Jakel S, Netter C, Schreiter K, Aicher B, Jackle H, Wahl MC. Crystal structures of the Mnk2 kinase domain reveal an inhibitory conformation and a zinc binding site, *Structure*, 2005 (10):1559-68

Netter, C, Weber, G, Luhrmann, R, Wahl, MC. Crystal Structure of an Expanded RNA Recognition Motif of the Minor Spliceosomal U11/U12-65K Protein. In preparation, 2008.