INTRODUCTION

1. Splicing an essential mechanism of gene expression

Since the discovery of genetic traits by an Austrian monk growing peas in his cloister garden (Mendel, 1866), scientists have sought to understand how genes are inherited and how they give rise to a specific phenotype in an organism. The discovery of DNA as the carrier molecule of genetic information and the principle of its structure (Avery *et al*., 1944; Watson and Crick, 1953) represented the beginning of the molecular understanding of genes. The transmission of genetic information into the appearance of an organism was described by the "central dogma of molecular biology" (Crick, 1970). According to this dogma genetic information encoded in DNA is transcribed into RNA which is translated into protein. Although this pathway has become much more detailed and complicated since then, and exceptions from the direction of information flow have been found (for example in viruses), the general rule of the dogma still holds true. The genetic code that relates the base sequence of the DNA to the amino acid sequence of a protein is nearly universal for all organisms. As the physical structure of genes was first established by work in bacteria it soon became evident that transcription of DNA into RNA and translation of RNA into protein occurred simultaneously (Brenner *et al*., 1961; Imamoto and Ito, 1968). In eukaryotes the gene structure was assumed to be the same although the synthesized RNA has to be transported from the nucleus to the cytoplasm where it can be translated into protein.

However, several findings suggested that the pathway from DNA to protein in eukaryotes proceeds differently than in bacteria. Most importantly, a species of RNA called heterogeneous RNA (hnRNA) was detected in the nucleus that was longer than the mRNA present in the cytoplasm and seemed to have a shorter half-life (Bachenheimer and Darnell, 1975). The most important step towards understanding of the pathway from the primary RNA transcript to the RNA which was translated into protein was made by hybridization experiments between the cytoplasmic mRNA expressed during the late stages of adenovirus infection and the viral DNA from which it was transcribed. In the electron micrographs of these experiments hybrids could be identified that contained long stretches of out-looping sequences of the viral DNA that seemed to have no complementary sequences in the mRNA found in the cytoplasm (Berget *et al*., 1977; Chow *et al*., 1977). This led to the proposition of "split" genes that undergo processing prior to being translated into protein. The process was termed "splicing" because it removed intervening sequences (introns) and brought the remaining sequences (exons) together. The discovery of splicing also explained the appearance of hnRNA that could be envisioned as a precursor of the spliced mRNA.

Shortly thereafter a number of cellular genes were identified which contained intervening sequences that had to be removed before translation (for example: Jeffreys and Flavell, 1977; Breathnach *et al*., 1977). Soon it became clear that most eukaryotic genes contain several introns and that the primary transcript is much longer than the mature mRNA which is translated into protein.

1.1. Classes of introns

Several classes of introns have been identified in the last 20 years which differ in their sequence, secondary structure and/or in the mechanism by which they are spliced. The main classes are: tRNA introns, Group I and II introns (including self-splicing introns) and spliceosomal introns.

Introns are present in many tRNA genes throughout all kingdoms (61 out of 274 tRNA genes in yeast, Lopez and Séraphin, 2000) and are usually short in size (14-60 nt in yeast). Their sequences show no obvious conservation, but their location is fixed immediately 3' to the anticodon in the anticodon loop (Ogden *et al*., 1984). The splicing mechanism of tRNA introns involves an endonuclease which recognizes the splice sites and removes the intron. Afterwards, a tRNA ligase joins the two generated ends in a three step ATP- and GTP-dependent reaction. Finally, a phosphotransferase removes a 2' phosphate generated by the ligation and transfers it to NAD (reviewed in Abelson *et al*., 1998). In yeast, a pre-mRNA (*HAC1*) has been identified that is spliced by a similar mechanism (see 2.7.).

Group I and group II introns are found in several protozoan rRNAs (for example Zaug *et al*., 1983), but also in mitochondria and chloroplasts of fungi and plants and in eubacterial genomes. Group I introns have been identified in bacteriophages (reviewed in Cech, 1993). Many group I and group II introns have the ability to catalyze the splicing process without any additional protein and are therefore also called self-splicing introns. However, under physiological conditions proteins are often needed to provide structural support, for example maturases in the case of yeast mitochondrial group II introns (reviewed in Lambowitz and Perlman, 1990). Both group I and group II introns have a particular secondary and tertiary structure that forms the catalytic core for the splicing reaction (reviewed in Michel and Ferat, 1995; Cate *et al*., 1996; Golden *et al*., 1998). They are spliced by a two step transesterification mechanism that shares similarities with spliceosomal splicing (see 2.6.). Another minor class of introns (Group III) is present in *Euglena* chloroplasts (reviewed in Woolford and Peebles, 1992). The exact mechanism of splicing of these introns is not clear. In addition, imbricated combinations of group I, group

II and group III introns have been found that were named twintrons (Copertino and Hallick, 1991; Doetsch *et al*., 1998).

The largest class of introns is spliced by a complex protein-RNA machinery called the spliceosome. Spliceosomal splicing occurs only in eukaryotes and involves a two step transesterification reaction (see 2.1.). Most spliceosomal introns identified so far are present in pre-mRNAs, but also snoRNAs and snRNAs contain spliceosomal introns.

1.2. Why genes in pieces?

Looking at the gene organization of an average vertebrate gene it becomes clear that most of its length is occupied by introns. An enormous amount of energy is required to maintain these intronic sequences in the genome and express them: they have to be replicated, packed into chromatin, repaired, transcribed, spliced out and finally degraded. Why do genes contain introns and what is the reason that a mechanism as complex as splicing evolved to remove them? Why are introns present in higher species, but are only found in small numbers in bacteria or lower eukaryotes like yeast? Two main directions of questions have to be distinguished: first, questions regarding the evolutionary acquisition, maintenance or loss of introns and second questions regarding the selective advantage or disadvantage offered to an organism by the presence of introns and the splicing process. The debate about "intron early" or "intron late" models reflects the first direction (for example Gilbert, 1978; Crick, 1979). Proponents of the "intron early" or exon theory claim that introns were present very early in evolution, but have been lost in organisms that were "streamlined" for fast growth like bacteria or yeast (Gilbert and Glynias, 1993). An argument in favor of this theory is the notion that exons sometimes correspond to functional domains in proteins (de Souza *et al*., 1996, for an opposing view see Stoltzfus *et al*., 1994). According to the "exon shuffling" hypothesis early existing protein modules where exchanged and combined to create a variety of new proteins which left introns as spacers behind (Gilbert and Glynias, 1993). In contrast, the "intron late" theory states that introns were acquired late in evolution (Palmer and Logsdon, 1991) and were inserted into preexisting genes. Phylogenetic analysis is supporting the late insertion of many introns, but can not exclude the existence of a few "ancient" introns (reviewed in Logsdon *et al*., 1998). Both models have arguments in favor, but definitive evidence for one or the other is lacking (reviewed in Mattick, 1994). It is important to keep in mind that self-splicing introns might have predated the modern spliceosomal introns (see 2.6.) so that the splicing process in an ancient organism might not have depended on the presence of additional factors.

Possible mechanisms for the acquisition, mobility and loss of introns by an organism have been described (e. g. reverse splicing, self splicing introns as mobile elements, transposable elements; Giroux *et al*., 1994; reviewed in Lambowitz and Belfort, 1993). However, the loss of introns at the DNA level is not an easy task since the splice junction has to be restored without destroying the reading frame of the harboring gene. Therefore this process most likely proceeds through a spliced RNA intermediate. The removal of an intron has to occur in the germline to be heritable, which could explain why intron loss in vertebrates, where the germline cells are separated early in development, is a rare phenomenon (Logsdon *et al*., 1998). In contrast, changing the length or sequence (except for the consensus splice sites) of an intron in many cases does not interfere with its splicing and is a rather common phenomenon that can be explained by transposition, recombination events or replication errors occurring in cells.

If it was indeed difficult to get rid of introns why should the organism not profit from this additional genetic material? Once present in genes, introns were flexible regions without strong evolutionary pressure on their sequence that could evolve new functions without harming the surrounding gene. Several examples that support this hypothesis of late gain of function for introns are known by now. Especially in higher eukaryotes many genes can be spliced in a variable manner giving rise to several protein isoforms (reviewed in Smith *et al*., 1989a; Lopez, 1998). This process, called alternative splicing, can be regulated by additional factors that interact with sequence signals in the intron or exon (for example Wu and Maniatis, 1993). One of the best described examples for alternative splicing is involved in the sex determination pathway of *Drosophila melanogaster*. A cascade of alternatively spliced factors regulates the decision about the sexual fate of the entire organism (reviewed in Burtis, 1993; Moore *et al*., 1993; MacDougall *et al*., 1995). In yeast only a few cases of regulated splicing have been reported: these include the splicing of the meiosis dependent *MER2* (Engebrecht *et al*., 1991) and *MER3* (Nakagawa and Ogawa, 1999) genes and the splicing of the *HAC1* pre-mRNA (Sidrauski *et al*., 1996). It is important to note, however, that the splicing of *HAC1* proceeds through a mechanism that is different from the one seen for most pre-mRNAs and is similar to the splicing of tRNAs (see 2.7.).

Other examples for intron function include the presence of a number of snoRNAs inside introns that are cleaved out after the intron has been removed from the pre-mRNA (Moore, 1996) and the presence of maturase or transposase genes in many group I and II introns (see 2.6.).

2. Spliceosomal splicing

2.1. Mechanism of the splicing reaction

In the beginning of the 1980s conserved sequences were identified in pre-mRNAs that flanked the intron sequences and seemed to act as signals for splicing (reviewed in: Breathnach *et al*., 1978; Mount, 1982). These sequences forming consensus splice sites (5' splice site and 3' splice site) were found to be highly conserved in eukaryotic cells (Padgett *et al*., 1986) pointing to a common mechanism of splicing in these organisms.

The development of an *in vitro* system utilizing cell-free extract containing soluble components necessary for splicing was an important step toward the understanding of the nuclear splicing process (Hernandez and Keller, 1983; Padgett *et al*., 1983; Krainer *et al*., 1984; Lin *et al*., 1985). Further analysis of the intermediates of this reaction led to the proposition of the chemical mechanisms of splicing (Ruskin *et al*., 1984; Padgett *et al*., 1984).

Splicing takes place after or during transcription when the pre-mRNA is still in the nucleus and involves two consecutive transesterification reactions (Figure 1). In the first step, the 2' hydroxyl group of an adenosine at the so called branch site (BS) attacks the phosphodiester bond of the first nucleotide at the 5' splice site (5'SS). This reaction generates a free 5' exon and a lariat intermediate that contains a branch structure: $^{...G5'p2'}{}_{p}A_{3'p5'N...}$. In the second step the free 3' hydroxyl group of the 5' exon attacks the 3' splice site (3'SS) generating the joined exons and the free intron-lariat (Grabowski *et al*., 1984; Ruskin *et al*., 1984; Padgett *et al*., 1984).

Soon after the discovery of splicing it became evident that this process also occurred in lower eukaryotes like yeast (Ng and Abelson, 1980). Because the chemical mechanism of splicing and many factors involved in the splicing process are conserved from yeast to man, the budding yeast, *Saccharomyces cerevisiae*, has become one of the most important model systems for the analysis of splicing (see 3.; reviewed in Rymond and Rosbash, 1993).

While most pre-mRNA splicing events occur in *cis*, which means the same molecule provides donor and acceptor of the splicing reaction, some organisms (like nematodes and trypanosomes) show a different splicing mechanism that involves two independent RNA molecules which are joined by *trans*-splicing (reviewed in Nilsen, 1997). Through this mechanism a spliced leader (SL) RNA is attached to the 5' end of mRNAs thereby generating a uniform 5' end and also allowing for the generation of several mRNAs from a single transcript (Figure 1).

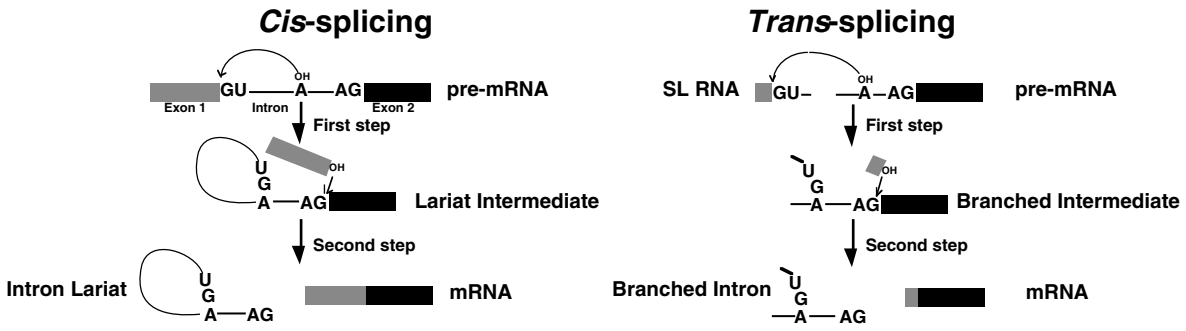

Figure 1. Mechanism of nuclear pre-mRNA splicing

Schematic representation of the *cis*- and *trans*-splicing reaction. Some conserved nucleotides and the hydroxyl groups involved in the chemical reaction are shown.

The precursor RNA contains clearly defined elements in its intron that determine the proper sites for the splicing reaction (Figure 5). The 5' splice site marks the beginning of the intron. It contains a highly conserved consensus sequence which is R/GUAUGU for yeast and AG/GURAGU for mammals (R= purine, / denotes the exon/intron boundary). The branch site lies between 100 and 18 bases upstream of the 3' splice site and has the consensus: UACUAAC for yeast and CURAY for vertebrates (Y= pyrimidine, N= any nucleotide, A= branching nucleotide). In higher eukaryotes, a polypyrimidine tract variable in length is often located between the branch site and the 3' splice site. The 3' splice site has the consensus: YAG/N for yeast and YAG/R for mammals (Senapathy *et al*., 1990; Burge *et al*., 1998a; Lopez and Séraphin, 1999).

2.2. The spliceosome

The nuclear machinery responsible for the excision of introns and the joining of exons is a large protein-RNA complex (50-60S) dubbed the spliceosome. It was first analyzed by separation in glycerol gradients (Grabowski *et al*., 1985; Brody and Abelson, 1985; Frendewey and Keller, 1985). Later studies concerning the assembly and rearrangement of the spliceosome were done using non-denaturing gel-electrophoresis (Konarska and Sharp, 1986; Konarska and Sharp, 1987; Pikielny *et al*., 1986; Cheng and Abelson, 1987; Séraphin and Rosbash, 1989).

Spliceosomes are formed around the pre-mRNA substrate by the successive assembly of five small nuclear ribonucleoprotein-particles (snRNPs): U1, U2, U4, U5 and U6 which are composed each of a small nuclear RNA (snRNA), seven Sm core proteins common to all snRNPs (except for U6 snRNP, which contains a related set of seven proteins, the Sm-like proteins, see 2.4.) and several snRNP-specific proteins (reviewed in Krämer, 1996; Burge *et al*., 1998a). The snRNPs play a central role in the process of splicing. They are responsible for recognition of the splice sites and definition of exon/intron boundaries. In addition, the snRNAs build the framework of the spliceosome by interacting with each other and with the pre-mRNA. These interactions are partially mediated through base pairing and are dynamic so that the spliceosome complex changes during the process of splicing (Figure 2).

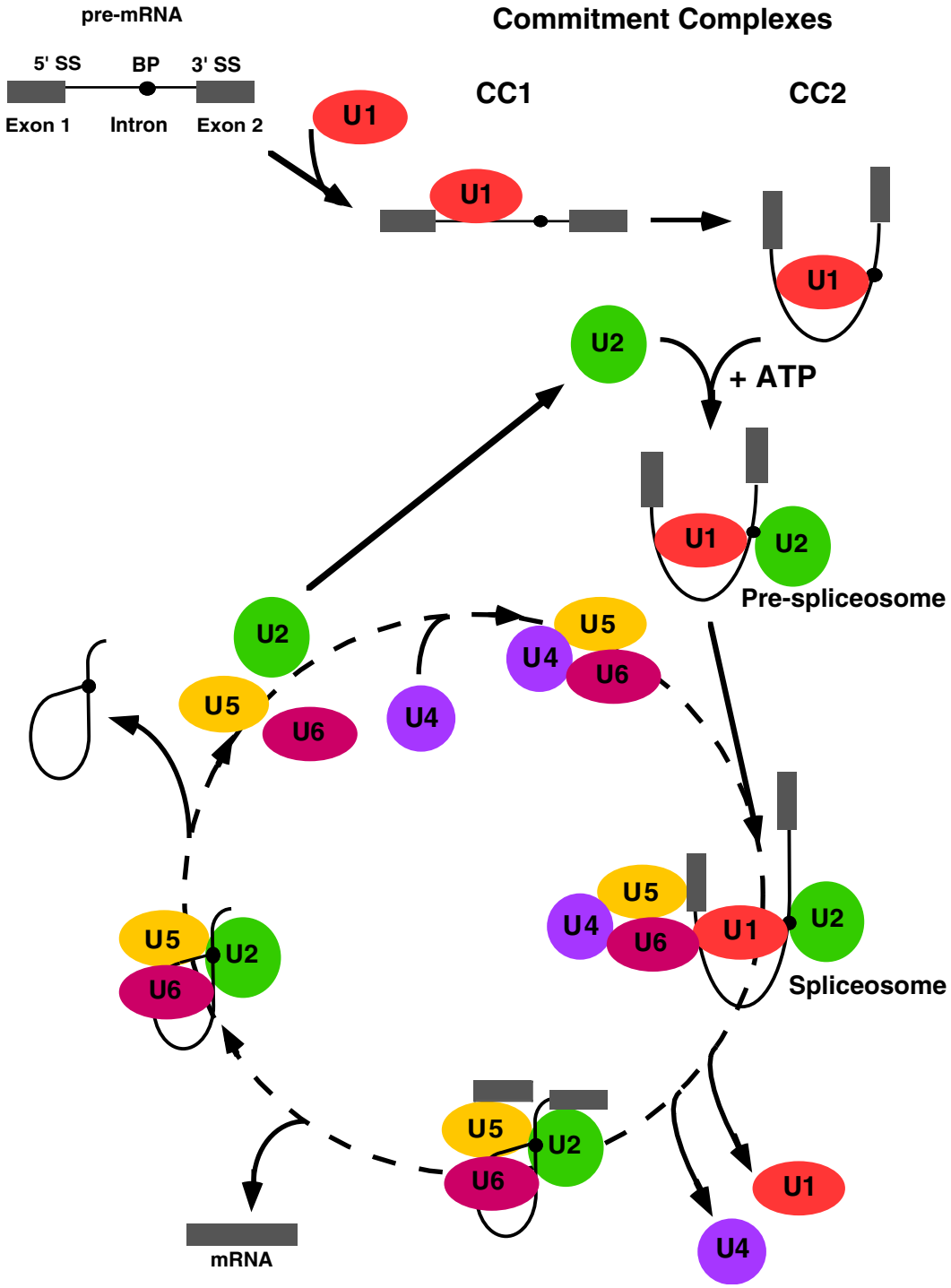

Figure 2. Spliceosome assembly

The stepwise assembly of snRNPs (shown as coloured spheres) on the pre-mRNA leads to the formation of the functional spliceosome. First, the 5' splice site (5'SS) is recognized by U1 snRNP (commitment complex 1, CC1 in yeast). Then a bridging interaction between the 5'SS and the branchpoint forms commitment complex 2 (CC2 in yeast, complex E in mammals). In the first ATP dependent step U2 snRNP is added to form the pre-spliceosome. The addition of the tri-snRNP U4/5/6 yields the mature spliceosome. The snRNPs are thought to recycle after the completion of the splicing reaction. Non-snRNP splicing factors are omitted for simplicity.

2.3. Dynamic interactions of snRNAs and the pre-mRNA

a) Building an active spliceosome

The following picture of the dynamic interactions in the spliceosome has emerged (Figure 3). The first step of spliceosome assembly involves the recognition of the 5' splice site by the U1 snRNP that is mediated in part by base pairing interactions (Figure 3A). This occurs in the commitment complexes in yeast (Séraphin and Rosbash, 1989) and in complex E in mammals (Michaud and Reed, 1991). Subsequently the ATP-dependent pre-spliceosome or complex A is formed in which U2 snRNA binds to the branch site region by base pairing and protein-RNA interactions (Figure 3A). A sequence in U2 snRNA (5'GUAGUA3') base pairs with a conserved sequence in the branch site forming a short duplex UACUAAC:GUAGUA in which the branch nucleotide is bulged out (Wu and Manley, 1989; Zhuang and Weiner, 1989; Query *et al*., 1994). In yeast, U2 snRNA has been proposed and shown to recognize the conserved branchpoint region by base pairing *in vivo* (Ares, 1986; Parker *et al*., 1987). This complex is transformed into the spliceosome or complex B upon U4/5/6 triple snRNP binding (Pikielny *et al*., 1986; Konarska and Sharp, 1987; Cheng and Abelson, 1987). In order for U6 snRNA to base pair to an overlapping site at the 5' splice site, U1 snRNA interaction with the 5'SS has to be disrupted (Figure 3). U1 base pairing to the 5'SS seems to be no longer needed for the proper assembly of the spliceosome (Konforti *et al*., 1993; Crispino *et al*., 1994; Tarn and Steitz, 1994). Once the base pairing between U1 snRNA and the 5'SS is destabilized, U6 snRNA binds to intron positions 4, 5 and 6 of the 5'SS. This interaction involves the ACA of the highly conserved ACAGAG sequence of U6 snRNA (Kandels-Lewis and Séraphin, 1993; Lesser and Guthrie, 1993). More recently, a genetic interaction between U6 snRNA and the first intron nucleotide has been shown (Luukkonen and Séraphin, 1998). Results from *in vitro* splicing studies of pre-mRNAs with an extended complementarity to U6 snRNA suggest that these interactions may stabilize the binding of the 5'SS by the spliceosome (Crispino and Sharp, 1995). The replacement of U1 snRNA by U6 snRNA interaction with the 5'SS occurs before the first step of splicing (Kandels-Lewis and Séraphin, 1993; Lesser and Guthrie, 1993; Wassarman and Steitz, 1993a).

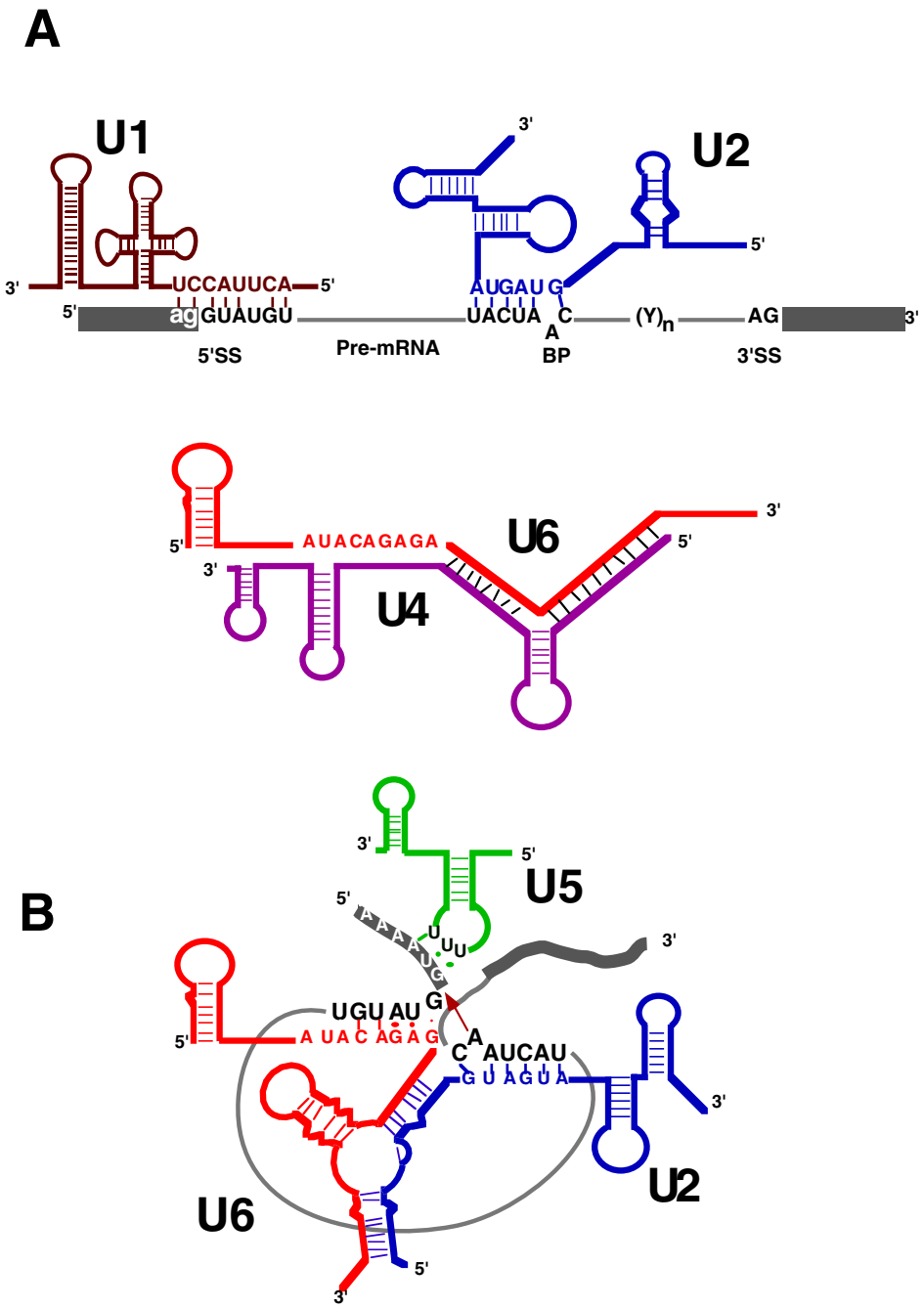

Figure 3. Dynamic interactions in spliceosome assembly

A network of base pairing interactions forms the framework for the catalytic center of the spliceosome. This is brought about by the rearrangement of several snRNA-snRNA and snRNA-pre-mRNA interactions. First, U1 snRNA base pairs with the 5' splice site and U2 snRNA base pairs with the branchpoint region, where the branchpoint adenosine is bulged out. U4 snRNA and U6 snRNA interact with each other through two extended helices. For the formation of the active spliceosome U6 snRNA has to replace U1 snRNA at the 5'SS by the formation of at least three base pairing interactions. This requires the disruption of the U4-U6 snRNA interaction. The formation of two or three new helices between the U2 and the U6 snRNA brings the two reaction partners for the first step of splicing (5'SS and branchpoint) into close proximity. The U5 snRNA interacts with bases in the first exon. The rearrangement of the spliceosome for the second step of splicing is not shown.

The function of U5 snRNA has long been discussed in this context. In a series of genetic and crosslinking studies its binding to exon sequences through a conserved loop that contains several uridines has been established (Newman and Norman, 1991; Newman and Norman, 1992; Wyatt *et al*., 1992; Sontheimer and Steitz, 1993; Newman *et al*., 1995). However, *in vitro* reconstitution experiments showed that the conserved loop is dispensable for the first step of splicing in yeast (O'Keefe *et al*., 1996). Prior to the first step of splicing, a conformational change occurs that destabilizes the association of U4 snRNP with the complex (Pikielny and Rosbash, 1986; Cheng and Abelson, 1987; Konarska and Sharp, 1987). U6 snRNA is initially base paired to U4 snRNA through two extended helices that have to be disrupted to allow U6 snRNA to bind to U2 snRNA through two (Madhani and Guthrie, 1992) or three (Sun and Manley, 1995) new helices (Figure 3). This is mediated, in part, by base pairing between the 3' end of U6 snRNA and the 5' end of U2 snRNA (Datta and Weiner, 1991; Wu and Manley, 1991). Since U2 snRNA binds to the branch site and U6 snRNA binds to the 5' splice site this interaction brings the reaction partners for the first transesterification into close proximity (Figure 3B).

A secondary structure model for the interaction of the snRNAs leading to the actual splicing reaction has been proposed (Madhani and Guthrie, 1992; McPheeters and Abelson, 1992). The branch site adenosine is thought to be bulged out of the helix formed by the branch site and U2 snRNA. The 2' hydroxyl group of this adenosine is positioned to attack the 5' splice site. In this reaction a free 5' exon and a lariat of the intron attached to the 3' exon is formed. U5 snRNA has been shown to make contact to both splice sites in their exon portions and thus could provide a structural element to bring both exons together (Newman and Norman, 1991; Newman and Norman, 1992; Wyatt *et al*., 1992; Sontheimer and Steitz, 1993; Newman *et al*., 1995). This contact has to be supported by additional RNA-protein interactions because the exon sequences are not strongly conserved and extended base pairings are not possible.

b) 3' splice site recognition

Two separate steps for the recognition of the 3' splice site have to be distinguished. First, an interaction with factors prior to the formation of pre-spliceosome that facilitates U2 snRNP binding to the branchpoint region, and second, the interaction of factors prior to the second transesterification reaction which defines the exact position for the 3' splice site. The first recognition step seems not to be required in all organisms or for all introns.

In *Schizosaccharomyces pombe*, the 3' splice site AG is required already for the first step of splicing. Genetic experiments could show that the residues of the U1 snRNA adjacent to

the 5' splice site interaction region recognize the 3' splice site AG by base pairing (Reich *et al*., 1992). These experiment also indicated that the 3' splice site is recognized by additional factor(s) before the second step of splicing. In mammals, introns can be divided into AG-independent introns that can undergo the first step of splicing without containing a 3' splice site and AG-dependent introns that do not assemble functional spliceosomes when the AG is mutated (Reed, 1989). Apparently the requirement for a 3' splice site AG can be compensated by a long polypyrimidine tract (Smith and Nadal Ginard, 1989). Very recently, the small subunit of the U2 auxiliary factor (U2AF), $U2AF^{35}$, has been shown to specifically recognize the AG dinucleotide at the 3' splice site (Merendino *et al*., 1999; Wu *et al*., 1999; Zorio and Blumenthal, 1999). These findings led to the proposal that for AG-dependent introns 3' splice site recognition is mediated through $U2AF^{35}$ binding to the 3' splice site AG and $U2AF^{65}$ binding to the polypyrimidine tract, while in AG-independent introns only the recognition of the strong polypyrimidine tract by $U2AF^{65}$ is sufficient for the first assembly steps (reviewed in Moore, 2000).

In contrast, in *Saccharomyces cerevisiae*, introns with a mutated AG can undergo the first step of splicing and base pairing of the 3' splice site with the 5' end of the U1 snRNA seems not to be required (Séraphin and Kandels-Lewis, 1993). Consistent with the model that involves $U2AF^{35}$ in early 3' splice site recognition, no homologue of $U2AF^{35}$ is present in *S. cerevisiae*. Other interactions like the binding of BBP/ScSF1 to the well conserved branchpoint region (see 4.2.) are likely to relieve the necessity for an early recognition of the 3' splice site in *S. cerevisiae*.

Recognition of the 3' splice site depends on its position downstream of the branchpoint sequence which is often followed by a polypyrimidine tract (Reed, 1989). A scanning mechanism has been postulated to determine the first YAG (Y is pyrimdine) after the branchpoint (Smith *et al*., 1989b; Smith *et al*., 1993). However, this hypothesis has been challenged by experiments in yeast that argue against a simple scanning mechanism (Deshler and Rossi, 1991; Patterson and Guthrie, 1991). Further studies revealed that the spacing between the branchpoint and the YAG is important and that two closely spaced AGs can compete with each other for the second step of splicing (Luukkonen and Séraphin, 1997). Recent results from *in vitro* experiments using a bimolecular splicing system have reinforced the evidence for a linear search in 3' splice site AG selection for mammalian introns (Chen *et al*., 2000). The factors recognizing the 3' splice site prior to the second transesterification remain to be identified, but several candidates including U5 snRNA and the proteins Prp8 and Slu7 have been suggested (reviewed in Umen and Guthrie, 1995b; Chiara *et al*., 1997; Chua and Reed, 1999).

In the second transesterification reaction the free 3' hydroxyl group of the 5' exon attacks the 3' splice site producing the joined exons and the free intron in a lariat form.
The spliced mRNA leaves the spliceosome and can be transported to the cytoplasm while the lariat intron is debranched and degraded in the nucleus.

2.4. snRNPs

snRNPs, the major components of the spliceosome, undergo a complex process of assembly and maturation before they can function in the splicing process. This process has been analyzed in great detail in vertebrates, but much less is known about the maturation pathway in yeast. U1, U2, U4 and U5 snRNP are synthesized by RNA polymerase II and acquire a mono-methyl-7-guanosine cap (m7G) in the nucleus. In addition, some of the snRNAs have been shown to be trimmed at the 3' end after transcription (Hernandez, 1985; Yuo *et al*., 1985). The cap binding complex (consisting of CBP20 and CBP80) and other still unidentified features serve as a nuclear export signal, which ensures transport to the cytoplasm (Jarmolowski *et al*., 1994; Izaurralde *et al*., 1995). There, the snRNAs associate with seven Sm proteins that bind to a conserved U-rich site in the RNA, the Sm site (reviewed in Guthrie and Patterson, 1988). The crystal structure of two Sm heterodimers (Kambach *et al*., 1999) and interaction studies for all yeast Sm proteins (Camasses *et al*., 1998) led to the proposition of a heteromeric ring structure containing all seven Sm proteins. The question how this ring could bind the RNA remains to be solved. The protein SMN (Survival of Motor Neurons) responsible for Spinal Muscular Atrophy (SMA), one of the most common human genetic diseases, has been implicated in the association of Sm proteins with the snRNA (Liu *et al*., 1997; Fischer *et al*., 1997). *In vitro* experiments showed that the purified snRNP components can self-assemble on *in vitro* transcribed snRNAs to give functional snRNPs (Segault *et al*., 1995). However, the exact mechanism of this assembly is still unclear.
After the Sm proteins have associated with the snRNA the cap is modified to a tri-methyl guanosine structure (TMG, Plessel *et al*., 1994). This cap structure and the associated Sm proteins serve as import signals for the snRNP particle (Hamm *et al*., 1990; Marshallsay and Lührmann, 1994). Recently, it has been shown that 3' end trimming is required for the nuclear import of U2 snRNA (Huang and Pederson, 1999). In the nucleus the snRNAs are further modified (sugar and base modification, for example Patton, 1994; Yu *et al*., 1998; Hartmuth *et al*., 1999) and associate with additional snRNP specific proteins which are transported independently into the nucleus (e.g. U1A, U2B", Hetzer and Mattaj, 2000).

U6 snRNP differs from the other spliceosomal snRNPs in several respects: first, it is transcribed by RNA polymerase III, second, it contains a 5' γ-mono-methyl end, third, it does not leave the nucleus, but seems to contain a nuclear retention signal, fourth, it does not contain an Sm binding site and does not interact with the canonical Sm proteins. However, recent studies have identified a second set of Sm-like (Lsm) proteins that bind to U6 snRNA and could form a similar structure like the canonical Sm proteins on the other snRNPs (Séraphin, 1995; Cooper *et al*., 1995; Salgado-Garrido *et al*., 1999; Mayes *et al*., 1999).
Very recently, some of these Lsm proteins were detected in a second complex that seems to be involved in mRNA degradation in the cytoplasm (Boeck *et al*., 1998; Bouveret *et al*., 2000).

2.5. Proteins in pre-mRNA splicing

Assembly and functioning of the spliceosome requires approximately 100 proteins. Many of them are integral parts of the snRNPs, like the Sm-core proteins, which are common to all snRNPs (except for U6), and several snRNP specific proteins (reviewed in Burge *et al*., 1998a). A number of proteins involved in the splicing process possess RNA binding motifs (RRM), which bind single stranded RNA, or serine-argine rich (RS) domains that have been shown to facilitate RNA-RNA annealing. In addition, a large number of enzymatically active proteins have been identified, the largest group being the DExD/H box ATPases (reviewed in Staley and Guthrie, 1998). These proteins share several conserved sequence features and are thought to function in unwinding or remodeling RNA-RNA interactions. Very recently, a viral member of this family has been shown to have processive and directed RNA unwinding activity (Jankowsky *et al*., 2000). Examples for those activities in the splicing process are the Prp24 protein which mediates base pairing between the U4 and U6 snRNAs (Raghunathan and Guthrie, 1998) and the Prp28 protein which is involved in the replacement of U1 snRNP from the 5' splice site (Staley and Guthrie, 1999). Other ATPases are responsible for the transition from commitment complex to pre-spliceosome (Prp5, UAP56), function immediately before the first or second step of splicing (Prp2 and Prp16 respectively) or release the mRNA after splicing (Prp22, reviewed in Staley and Guthrie, 1998). Recently, a protein, U5-116 kDa (Snu114 in *S. cerevisiae*), which is similar to the ribosomal GTPase EF-2 has been identified as a component of the U5 snRNP (Fabrizio *et al*., 1997). This opens the question if translocation activities like the ones observed in translation are also functioning in splicing. Another protein component of U5 snRNP, U5-20 kDa, shows sequence similarity to a peptidyl-prolyl *cis*/*trans*-isomerase and

exhibits isomerase activity *in vitro*, but its function in splicing has not yet been elucidated (Teigelkamp *et al*., 1998).

The largest protein component of the spliceosome identified so far is the Prp8 protein. It shows very strong conservation through evolution (Hodges *et al*., 1995) and has been implicated in many different functions, ranging from the recognition of the 5' splice site GU dinucleotide (Reyes *et al*., 1996; Reyes *et al*., 1999) over the close association with the branch site in the active spliceosome (MacMillan *et al*., 1994) to the recognition of the 3' splice site before the second catalytic step of splicing (Teigelkamp *et al*., 1995; Umen and Guthrie, 1995a). In addition, it is required for the association of the U5 snRNP with the U4/U6 snRNP and interacts with a number of splicing factors (reviewed in Beggs *et al*., 1995; Newman, 1997).

Given the numerous interactions and proposed functions of this protein and in light of its strong evolutionary conservation Prp8p has been proposed to contribute to the catalytic center of the spliceosome (Reyes *et al*., 1999).

2.6. Evolution of nuclear pre-mRNA splicing

By comparison of spliceosomal splicing with self-splicing, a catalytic activity of snRNAs has been proposed (reviewed in Moore *et al*., 1993; Nilsen, 1998). The first step of group II self-splicing closely resembles the first step of spliceosomal pre-mRNA splicing in the involvement of a nucleoside at the branch site (predominantly an adenosine) whose 2' hydroxyl group attacks the 5'SS resulting in the formation of a lariat intermediate (Figure 4; Peebles *et al*., 1986; van der Veen *et al*., 1986). In addition, structural resemblance between domain 5 in group II introns and the U6/U2 helix in the spliceosome have been pointed out (Madhani and Guthrie, 1992). Group I self-splicing differs from the two other splicing mechanisms in its need for a free guanosine as a cofactor (Figure 4). The guanosine provides a hydroxyl group for the first transesterification step resulting in a linear intron-3' exon and a 5' exon intermediate. The second step in group I self-splicing involves the nucleophilic attack of the free 3' hydroxyl on the 5' exon at the 3' intron-exon boundary thereby creating the joined exons and a free intron (reviewed in Cech, 1986).

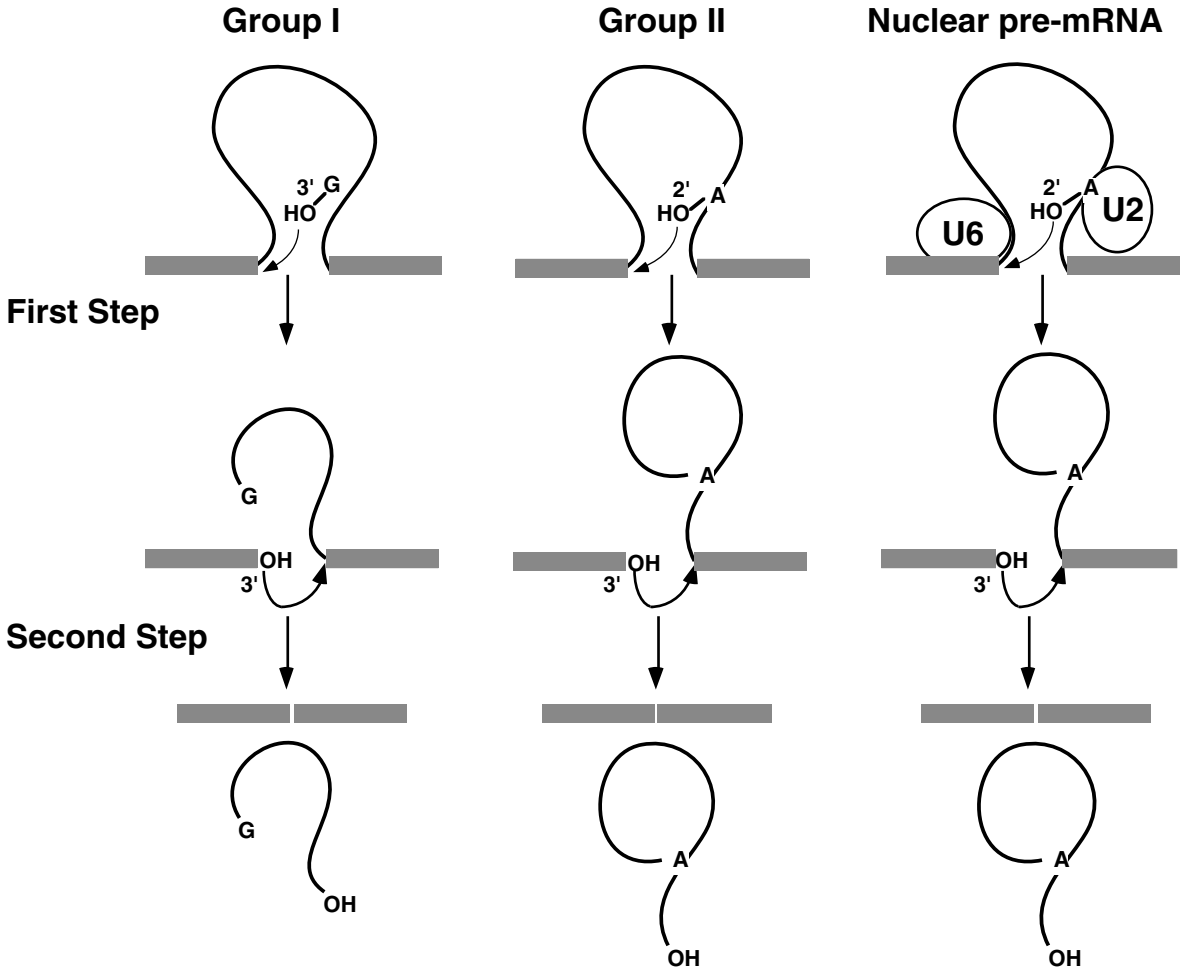

Figure 4. Comparison between Group I, Group II and spliceosomal splicing

Schematic representation of the splicing of Group I, Group II and nuclear pre-mRNA introns. The nucleophilic hydroxyl group (OH) and its position on the nucleotide (2' or 3') are shown for both steps of the transesterification reaction. Exons are drawn as gray boxes and introns are shown as black lines.

While the primary sequences of self-splicing introns are not conserved, their secondary and tertiary structures forming a set of domains are highly conserved (reviewed in Cech, 1993). The comparison of structures of group II introns and snRNAs involved in spliceosomal pre-mRNA splicing led to the hypothesis that snRNAs may also be catalytically active. It has been proposed that spliceosomal pre-mRNA splicing evolved from self-splicing by the separation of individual, catalytically active RNA domains of introns that became independent, trans-acting snRNAs (Sharp, 1985; Cech, 1986).

Focusing on the chemical mechanism involved in the different splicing mechanisms, and especially on the stereochemistry of the reactions, similarities between group I self-splicing and the second step of spliceosomal splicing have been pointed out. Both involve in line

transesterifications by a S_N2 mechanism which correlates with a change of the chirality of the group involved (Moore and Sharp, 1993).

2.7. Minor classes of pre-mRNA introns

Several introns in metazoans contain splice sites that deviate from the consensus sequence (Jackson, 1991). Instead of being flanked by the usual dinucleotides /GU-AG/ they contain /AU-AC/ (/ indicates the splice junction). Because these introns still seemed to be spliced efficiently and reliable the presence of a second type of spliceosome responsible for the removal of this minor class of introns was proposed (Hall and Padgett, 1994). The search for the components of this novel spliceosome led to the identification of the already known U11 and U12 snRNPs that replace U1 and U2 snRNP, respectively. In addition, snRNPs related to U4 and U6 were found to be present in the minor class of spliceosomes and were therefore called U4-ATAC and U6-ATAC. The U5 snRNP is the only snRNP shared by the two spliceosomes (Hall and Padgett, 1996; Tarn and Steitz, 1996b; Tarn and Steitz, 1996a). The minor class of introns and the associated spliceosome have been detected so far in vertebrates, insects and plants, but are absent from yeast (Wu *et al*., 1996).

Despite the original connection of the minor spliceosome to AU-AC containing introns, discovery of additional introns with non-consensus splice sites and the development of *in vitro* and *in vivo* systems to analyze splicing by the minor spliceosome showed that both types of spliceosomes are capable of splicing introns flanked by AU-AC and GU-AG (reviewed in Kreivi and Lamond, 1996; Burge *et al*., 1998b). It turned out that more characteristic determinants for the type of spliceosome activated are the branchpoint sequence that is recognized by U2 or U12 and the 3' part of the 5' splice site recognized by U1 or U11 (Figure 5). Therefore introns were classified as U2- or U12-dependent. The purification of the U11/U12 snRNP led to the identification of some common proteins with the U2 snRNP (the SF3b subunit, see 4.2.) and a protein homologous to the U1 snRNP 70K protein (Will *et al*., 1999). In addition, all Sm core proteins present in the snRNPs of the major spliceosome were found associated. Different models for the evolution of the two types of spliceosomes have been proposed (Burge *et al*., 1998b).

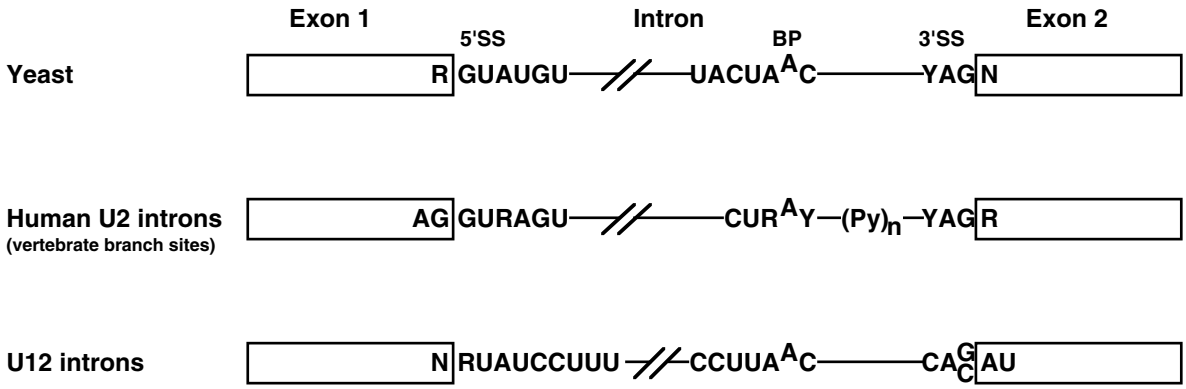

Figure 5. Splice site consensus sequences

Comparison of splice site consensus sequences for yeast, human U2 dependent and U12 dependent introns (according to Burge *et al*., 1998a). The most conserved regions, 5' splice site (5'SS), branchpoint (BP), 3' splice site (3'SS), are shown with their consensus sequences (R=purine, Y=pyrimidine). The polypyrimidine tract often present in metazoan U2 dependent introns is indicated as $(Py)_n$.

A completely different type of pre-mRNA splicing has recently been discovered for the yeast gene *HAC1*. This transcription factor mediates the unfolded protein response (UPR) in the endoplasmatic reticulum (ER) whereby several protein folding components are upregulated when the concentration of unfolded polypeptides reaches a certain threshold. The synthesis of Hac1p was shown to be regulated by splicing (Cox and Walter, 1996). However, this splicing event is independent of the spliceosome, but is instead brought about by the action of a tRNA ligase and the endoribonculease Ire1p (Sidrauski *et al*., 1996; Sidrauski and Walter, 1997. Therefore it closely resembles the splicing of tRNA precursors (Gonzalez *et al*., 1999). Very recently the yeast *HAC1* pre-mRNA has been shown to be correctly spliced in human cells and an ER stress element can be activated in human cells by yeast Hac1p (Niwa *et al*., 1999; Foti *et al*., 1999). It therefore seems likely that the UPR is conserved in mammalian cells and that a splicing mechanism similar to the one observed for *HAC1* exists also in higher eukaryotes.

2.8. Exon definition and intron bridging

The structure of mammalian genes, which often contain very long introns (while exons are rather short), raises the question how the splicing machinery can distinguish exon from intron sequences and how splice sites are brought together over large distances of intervening sequence. Moreover, if a gene contains many introns (as it is the case for most mammalian genes) a given 5' splice site has to be joined with the right 3' splice site to prevent the loss of coding sequence. This question is especially intriguing because in higher eukaryotes splice site consensus sequences are rather loosely conserved (Figure 5). A model

has been proposed where splice sites are defined by an interaction between U1 snRNP binding to the 5' splice site and U2AF^{65} binding to the polypyrimidine stretch of the upstream intron (Nasim *et al*., 1990; Robberson *et al*., 1990). This "exon definition" model is consistent with many of the frequently seen phenotypes associated with splice site mutations, namely "exon skipping" (the loss of an entire exon) or the activation of cryptic splice sites close to the mutated original splice site. However, for the terminal introns other interactions have to ensure the proper recognition of splice sites. In this case, interactions of U1 snRNP with the cap-binding complex bound to the 5' end of the RNA and with the poly-A site at the 3' end of the RNA are thought to provide the necessary information (Izaurralde *et al*., 1994; Ohno *et al*., 1987; Niwa and Berget, 1991; Wassarman and Steitz, 1993b; Lutz and Alwine, 1994; Gunderson *et al*., 1997; Vagner *et al*., 2000a). Very recently a physical interaction of U2AF^{65} with the poly(A) polymerase was demonstrated which stimulates U2AF^{65} binding to the polypyrimidine tract and splicing of the intron (Vagner *et al*., 2000b). Other models for the definition of exons and introns based on the recognition of introns or both introns and exons have been proposed alternatively (reviewed in Black, 1995).

A class of proteins that contains one or more RNP-type RNA-binding domains and a domain rich in arginine and serine residues (hence called SR proteins) has been implicated in bridging interactions in higher eukaryotes (reviewed in Fu, 1995). Members of this family can interact with U2AF^{35}, which itself contains an RS domain, (Wu and Maniatis, 1993) and with components of the U1 snRNP (Kohtz *et al*., 1994) or directly with the 5' splice site (Zuo and Manley, 1994). Therefore it has been proposed that these proteins might mediate intron and exon bridging interactions (Fu and Maniatis, 1992; Wu and Maniatis, 1993; Staknis and Reed, 1994). In addition, SR proteins influence splice site choice and different SR proteins are required for the splicing of specific introns. This indicates a role of SR proteins in splicing regulation (reviewed in Fu, 1995). Consistent with this hypothesis, the SR family of proteins is not present in *S. cerevisiae* and also other splicing factors that contain SR domains in higher eukaryotes are lacking those in yeast (like U2AF^{65} or U170K).

For short introns it is believed that bridging occurs rather across the intron than the exon (see for example Kennedy *et al*., 1998). In yeast, where introns are small and often located at the 5' end of the transcript intron bridging seems to be the rule, but an interaction with the cap-binding complex facilitates spliceosome formation (Lewis *et al*., 1996; Colot *et al*., 1996).

3. *Saccharomyces cerevisiae*, a model organism for pre-mRNA splicing

Soon after the discovery of splicing, the already well characterized budding yeast, *Saccharomyces cerevisiae*, became the organism of choice for genetic studies on splicing. Many mutants characterized previously in a screen for synthesis of RNA (therefore called *rna*; Hartwell, 1967; Hutchison *et al*., 1969) were found to be affected in RNA splicing and where therefore renamed "precursor of RNA processing" (*prp*) mutants (Vijayraghavan *et al*., 1989). In the past years a vast number of *PRP* genes and other splicing factors have been described in yeast and the list seems to be still growing (reviewed in Burge *et al*., 1998a). The analysis of the splicing mechanism remained not restricted to genetic methods, but also *in vitro* splicing assays were established (Newman *et al*., 1985) and new gel systems allowed for the detection of different splicing complexes (Pikielny *et al*., 1986; Cheng and Abelson, 1987; Séraphin and Rosbash, 1989). These studies showed that the general mechanism of splicing is conserved from yeast to man.

More recently the entire genome of *S. cerevisiae*, as the first eukaryotic organism, was sequenced (Goffeau *et al*., 1996) thus enabling scientists to search for putative homologues of splicing factors known from other organisms, and also to employ bioinformatical methods to detect all introns present in yeast genes. Although the latter analysis is not complete, as some introns escaped detection (for example *MER3*, see 1.2.), it is clear that in yeast only a limited set of genes harbor an intron. So far among the about 6000 genes identified, only 245 pre-mRNAs contain one intron and five pre-mRNAs contain two introns (Spingola *et al*., 1999; Lopez and Séraphin, 2000). However, as shown recently, many pre-mRNA introns reside in highly expressed genes, so that in average every third message present at a given time in a yeast cell contains an intron and has to be spliced (Lopez and Séraphin, 1999; Ares *et al*., 1999). In addition, 61 tRNA introns and the *HAC1* intron which are spliced by protein enzymes are present in the yeast genome (see 1.1. and 2.7.). There is a high degree of conservation not only in the general pathway of pre-mRNA splicing, but also among the RNA and protein factors involved in this process (Krämer, 1996; Burge *et al*., 1998a).

In addition, more practical reasons like easy handling, the presence of a large number of selectable markers, the existence of a haploid and a diploid state and the relative genetic stability have made yeast one of the favored model organisms of modern biology. The effectiveness and preciseness of homologous recombination in *S. cerevisiae* allows for the targeted disruption or modification of yeast genes. This enables large scale genomic and proteomic studies, but also detailed functional analysis of individual genes. *S. cerevisae* has proven to be a powerful model organism for the analysis of pre-mRNA splicing.

4. The importance of being early

4.1. Commitment to splicing

The decision to splice a given pre-mRNA and the choice of the appropriate splice sites take place at early stages of spliceosome assembly. Indeed, the first splicing complexes detectable *in vitro* already commit the associated pre-mRNA to the splicing pathway (Legrain *et al*., 1988; Séraphin and Rosbash, 1989; Michaud and Reed, 1991). These early steps are therefore important for the definition of splice sites and for the regulation of splicing. Intron recognition is initiated by the binding of the U1 snRNP to the 5' splice site (reviewed in Rosbash and Séraphin, 1991). It involves a U1 snRNA-pre-mRNA base pairing interaction that is strengthened by interactions of some snRNP proteins with neighboring pre-mRNA regions (Puig *et al*., 1999; Zhang and Rosbash, 1999) and a bridging interaction of the cap-binding complex with the methyl-7-guanosine cap of the pre-mRNA (Lewis *et al*., 1996; Colot *et al*., 1996). This complex can be detected in native gels following assembly in yeast extracts and is called commitment complex 1 (CC1) because it commits the pre-mRNA to the splicing pathway (Séraphin and Rosbash, 1989). It requires only an intact 5' splice site for its assembly. The formation of commitment complex 2 (CC2) depends, in addition, on the presence of a branchpoint sequence (Séraphin and Rosbash, 1991). Both commitment complexes contain the U1 snRNP (Séraphin and Rosbash, 1989) and require the presence of the cap-binding complex for efficient formation (Colot *et al*., 1996; Lewis *et al*., 1996; Fortes *et al*., 1999; Puig *et al*., 1999). These complexes are formed in the absence of ATP and can be chased quantitatively into spliceosomes (Séraphin and Rosbash, 1989). An “early” complex with similar properties has also been identified in mammalian nuclear extracts and was termed complex E (Michaud and Reed, 1991). Given the sequence requirements for CC2 formation it has been speculated for a long time that the transition from CC1 to CC2 would involve the recognition of the branchpoint by factors interacting directly or indirectly with the U1 snRNP. This interaction would bridge the intron and bring 5' and 3' splice sites into close proximity (Figure 6).

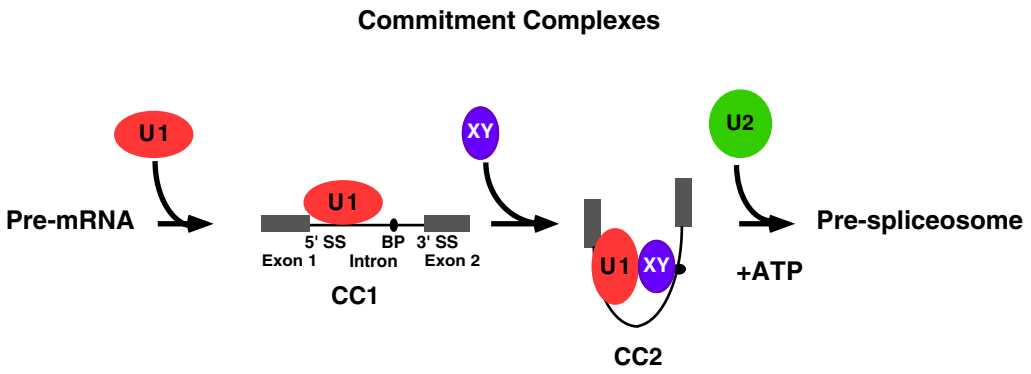

Figure 6. Commitment complexes and intron bridging

4.2. SF1/BBP and U2AF^{65}/Mud2p recognize the branchpoint region

The yeast gene *MUD2* was identified in a screen for mutants synthetic lethal with a U1 snRNA mutation. The Mud2 protein was shown to be involved in the formation and to be a component of CC2 (Abovich *et al*., 1994). *MUD2* was found to be inessential for viability. However, the observation that mutations synthetic lethal with a *MUD2* deletion affect components of the U1 and U2 snRNP (Abovich *et al*., 1994) supported a role for the Mud2 protein in intron bridging. In addition, Mud2p requires an intact branchpoint region for binding (Abovich *et al*., 1994; Rain and Legrain, 1997) and has been suggested to recognize the nucleotide immediately preceding the conserved branchpoint sequence (Rain and Legrain, 1997). The metazoan homologue of Mud2p appears to be the U2AF^{65} splicing factor. This protein interacts with the pyrimidine-rich sequence that often follows metazoan branchpoints (Zamore and Green, 1989; Zamore *et al*., 1992) and is required for U2 snRNP addition (Ruskin *et al*., 1988) by facilitating the base pairing of U2 snRNA with the branch site (Valcárcel *et al*., 1996). U2AF^{65} has been shown to be present, together with the U1 snRNP and several SR proteins, in the metazoan E complex (Bennett *et al*., 1992a; Staknis and Reed, 1994). The involvement of the U2AF^{65}-pre-mRNA interaction in intron recognition is demonstrated by its implication in alternative splicing regulation (Valcárcel *et al*., 1993). Beside U2AF^{65}, a second metazoan splicing factor has been shown to be required for pre-spliceosome formation (Krämer and Utans, 1991). This factor is also required for pre-mRNA splicing. It has been biochemically purified and the corresponding polypeptide was named SF1. Interestingly, database searches revealed the presence of a related protein, ScSF1, encoded by the yeast genome (Arning *et al*., 1996). The gene encoding ScSF1 was independently identified as *MSL5* in a screen for mutants synthetic lethal with a *MUD2* truncation (Abovich and Rosbash, 1997). The corresponding protein was shown to be implicated in the formation of CC2 and to physically interact with Mud2p (Abovich and

Rosbash, 1997; Fromont-Racine *et al*., 1997; Rain *et al*., 1998). In addition, it has been proposed to be present in CC2 since it co-precipitates U1 snRNA in a pre-mRNA dependent but U2 snRNP independent manner (Abovich and Rosbash, 1997). A bridging interaction between the Mud2p/ScSF1 complex binding to the branchpoint region and the U1 snRNP bound to the 5' splice site has been proposed (Abovich and Rosbash, 1997). The exact nature of this bridging is not clear, but proteins of the U1 snRNP have been shown to interact with ScSF1 (Prp40p and Prp39p; Abovich and Rosbash, 1997; Fromont-Racine *et al*., 1997). It was therefore named Branchpoint Bridging Protein (BBP) (Abovich and Rosbash, 1997). An interaction similar to the one in yeast has been proposed between mammalian SF1 and the formin binding proteins FBP11 and FBP21 which share a conserved WW motif with Prp40p (Abovich and Rosbash, 1997; Bedford *et al*., 1997; Bedford *et al*., 1998).

MSL5 is essential in yeast (Abovich and Rosbash, 1997) and, more recently, the homologue of SF1 in C. *elegans* was also found to be required for viability (Mazroui *et al*., 1999). Characterization of yeast and human BBP/SF1 revealed that they interact specifically with the branchpoint sequence (Berglund *et al*., 1997). Additional studies demonstrated that human BBP interacts with $U2AF^{65}$ (Abovich and Rosbash, 1997; Rain *et al*., 1998) and that both proteins bind cooperatively to the branchpoint/polypyrimidine tract region (Berglund *et al*., 1998a). The protein was therefore renamed Branchpoint Binding Protein (BBP). A model has been proposed to explain the differences in the recognition of the branchpoint in yeast and mammals. While the branchpoint sequence is well conserved in yeast and the affinity of BBP for this sequence is rather high ($k_d \approx 0.5\mu M$), mammalian introns show little conservation in their branchpoint sequences and the affinity of mBBP even for a consensus branchpoint sequence is low ($k_d \approx 30\mu M$, Berglund *et al*., 1997). In contrast, polypyrimidine tracts are missing from many yeast introns and Mud2p is not essential for yeast viability, while in mammalian introns polypyrimidine tracts are well conserved and $U2AF^{65}$ plays an indispensable role for the assembly of spliceosomes. Therefore it has been speculated that during the course of evolution $U2AF^{65}$ binding to the polypyrimidine tract has compensated in part for the loss of interaction of BBP/SF1 with the branchpoint (Berglund *et al*., 1997). This model has gained an additional facet by the finding that $U2AF^{35}$, the small subunit of U2AF, which is not present in yeast, binds to the 3' splice site AG and thereby strengthens $U2AF^{65}$ binding to the polypyrimidine tract (see 2.3.b; Merendino *et al*., 1999; Wu *et al*., 1999; Zorio and Blumenthal, 1999).

The first ATP dependent step in spliceosome assembly is the formation of the pre-spliceosome with the binding of U2 snRNP to the branchpoint region. It was speculated that this would replace BBP/SF1 which binds to the same region. In humans, a protein that

may correspond to mBBP/SF1 has been shown to be replaced at the branchpoint concomitant with the binding of U2 snRNP (MacMillan *et al*., 1994; Chiara *et al*., 1996).
The formation of pre-spliceosome requires two accessory complexes of the U2 snRNP, SF3a and SF3b (Krämer, 1996 and references therein, Das *et al*., 1999; Caspary *et al*., 1999; Krämer *et al*., 1999). *Prp11*/ySF3a^{66} has been shown to interact genetically with *MUD2* (Abovich *et al*., 1994). hSAP155/SF3b^{155} contacts the pre-mRNA at both sides of the branchpoint and interacts with U2AF^{65} (Gozani *et al*., 1998). In addition, two DEAD/DEAH-box helicases, Prp5 and Sub2/UAP56 have also been implicated in this step (reviewed in Staley and Guthrie, 1998). Prp5 is required for pre-spliceosome assembly and has been shown to mediate ATP dependent changes in the secondary structure of U2 snRNA (Ruby *et al*., 1993; O'Day *et al*., 1996). These findings led to the idea that Prp5 could prepare the U2 snRNA for binding to the branchpoint. UAP56 was found in a screen for interaction partners of U2AF^{65} and also has been shown to be required for pre-spliceosome assembly (Fleckner *et al*., 1997). A likely homologue of this factor was identified in yeast as a supressor of the cold-sensitive *brr1* snRNP biogenesis mutant, but its function has not been elucidated (reviewed in Staley and Guthrie, 1998). Very recently, the splicing factor Cus2, which shows homology to the mammalian transcription factor TAT-SF1 (not related to the splicing factor BBP/SF1), has been shown to alleviate the requirement of ATP for pre-spliceosome formation when mutated (Perriman and Ares, 2000). This finding allowed the hypothesis that Cus2 could be a proofreading factor that ensures correct binding of U2 snRNP to the branchpoint by regulating the ATP dependence of this step.

4.3. Domain structure of BBP/SF1 and Mud2/U2AF^{65}

The SF1 primary sequence can be divided in four structural domains: the N-terminus, the KH domain, the zinc knuckle region and a proline-rich C-terminus (Figure 7).
The N-terminus which shows little conservation through evolution is required for the interaction with Mud2p or U2AF^{65} (Rain *et al*., 1998). This interaction has been mapped to the amino acids 41-144 for yeast and the first 137 amino acids for human SF1. More recently, homologues of SF1 have been identified in *D. melanogaster* and *C. elegans* (Mazroui *et al*., 1999). These proteins contain a longer N-terminus which is rich in RS dipeptides and could be involved in additional functions of the protein.
The central part of SF1 shows high similarity to the STAR (Signal transduction and regulation of RNA) family of proteins. Members of this family include the *C. elegans* protein gld-1, which is responsible for sex-determination and functions as a regulator of

translation, the mouse "quaking" protein, which affects early embryogenesis and myelination and the mammalian Sam68 protein, which plays a role in signal transduction in mitotic cells (reviewed in Vernet and Artzt, 1997). Members of this family share a region of homology of about 200 amino acids that contains a single maxi-KH domain flanked by the QUA1 and QUA2 domains. While the KH domain (hnRNP K homology domain) which is present in a wide variety of RNA binding proteins (reviewed in Nagai, 1996; Lewis *et al*., 1999) has been clearly established as an RNA binding motif (for example Berglund *et al*., 1998b; Rain *et al*., 1998; Lewis *et al*., 2000) the role of the two QUA domains is less clear. SF1 contains only the QUA2 domain which enhances RNA binding *in vitro* (Berglund *et al*., 1998b). Although there is good evidence that STAR family members are involved in RNA metabolism and signal transduction pathways the link between those has not been established yet.

In this respect it is interesting that SF1 in higher eukaryotes also seems to be the target of a signal transduction pathway. A serine residue (Ser20) in the N-terminus of the human protein which is conserved together with neighbouring residues in *C. elegans* and *D. melanogaster*, but not in yeast, is specifically phosporylated by the cGMP dependent protein kinase (PKG-1) (Wang *et al*., 1999). This phosphorylation disrupts the interaction of SF1 with $U2AF^{65}$ and prevents pre-spliceosome formation indicating a role in splicing regulation.

The zinc knuckle motif (two in yeast, *C. elegans* and *D. melanogaster* and one in human) enhances RNA binding (Berglund *et al*., 1998b). However, it can be replaced by a viral nucleocapsid peptide containing a zinc knuckle motif or by a basic peptide containing seven argine-serine repeats without significant loss in binding affinity (Berglund *et al*., 1998b). Therefore it seems very likely that the zinc knuckles provide a rather unspecific interaction with the RNA backbone to help the binding of the KH domain.

In higher eukaryotes several splice variants of SF1 exist that are expressed in a tissue specific fashion (Toda *et al*., 1994;. Arning *et al*., 1996; Caslini *et al*., 1997; Wrehlke *et al*., 1997; Krämer *et al*., 1998; Zhang and Childs, 1998; Wrehlke *et al*., 1999). These isoforms differ only in the length of the proline rich domain and contain distinct C-termini. The C-terminus is not required for viability in yeast and recombinant human SF1 that lacks the C-terminus can restore pre-spliceosome formation in HeLa cell extract fractions (Rain *et al*., 1998). The role of the C-terminus for the function of SF1 and the relevance of the different SF1 isoforms has not yet been elucidated.

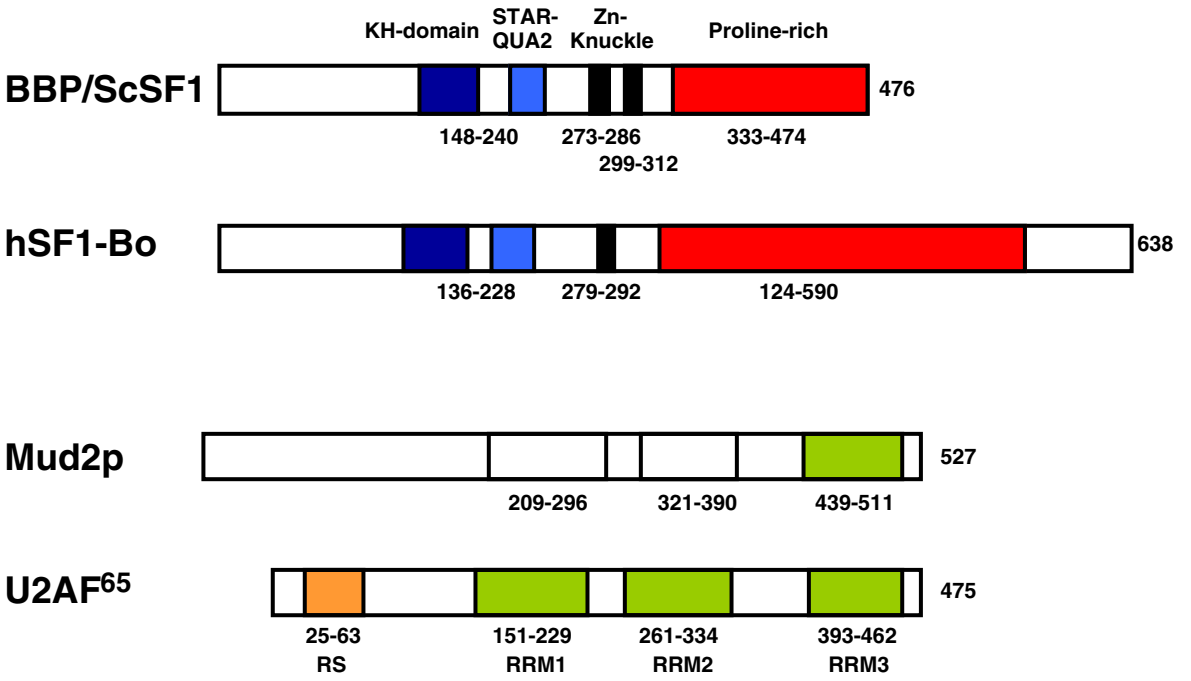

Figure 7. Domain structure of BBP/SF1 and Mud2/U2AF^{65}

Conserved domains are coloured identically and their names are indicated on the bottom and top of the figure. The borders of the domains are given below the primary structure (according to Rain *et al*., 1998).

The Mud2p primary structure shows three domains that can be aligned with the three RRMs of U2AF^{65}. However, sequence similarity between the two proteins is not very high except for the last domain which can clearly be identified as an RNA recognition motif. This domain in the human and in the yeast protein interacts in two-hybrid assays with BBP/SF1 (Rain *et al*., 1998). U2AF^{65} in addition contains an arginine-serine rich motif (RS) that facilitates binding of the U2 snRNA to the branchpoint (Valcárcel *et al*., 1996).

5. Pre-mRNA retention, mRNA export and nonsense-mediated decay

Gene expression is a highly ordered process that has to be controlled temporally and spatially. In eukaryotic cells, the RNA transcript undergoes several processing steps in the nucleus before it is exported to the cytoplasm. Pre-mRNA splicing is required to remove introns from the transcript in the nucleus. Introns often contain stop codons in frame with the upstream protein coding sequence. Thus, escape of the unspliced pre-mRNA from the nucleus would result in the accumulation of aberrant RNAs in the cytoplasm possibly leading to the production of truncated and deleterious proteins. Therefore, tight control to prevent undesired pre-mRNA export is an essential requirement for gene expression. Eukaryotic cells have evolved a proofreading mechanism that can detect and degrade

mRNAs containing pre-mature stop-codons. In addition, it prevents the accumulation of cytoplasmic pre-mRNAs that contain in frame stop codons in the intron and have escaped the nucleus before being spliced (He *et al*., 1993). This process called nonsense-mediated decay (NMD) is present in organisms as divergent as yeast and human (reviewed in Hilleren and Parker, 1999; Czaplinski *et al*., 1999).

How can the cell distinguish between intron containing pre-mRNAs, spliced mRNAs and intronless mRNAs to assure that the first are retained in the nucleus while the latter two are exported to the cytoplasm? In yeast, a pioneering study using a reporter system that allows for the detection of cytoplasmic pre-mRNAs has shown that intact 5' splice site and branchpoint are required for the nuclear retention of pre-mRNAs (Legrain and Rosbash, 1989). Moreover, several factors already known for their involvement in the splicing process affect pre-mRNA retention, namely the proteins Prp6 and Prp9 and the U1 snRNA. This led to a model where assembly of splicing complexes onto splicing signals serves as a retention signal for pre-mRNA. However, in some circumstances cells allow export of partly spliced or unspliced pre-mRNAs to the cytoplasm. A striking example is provided by the HIV-1 Rev protein that overcomes the retention of unspliced viral messages by binding to a Rev Response Element (RRE) thereby promoting export of target RNA independent of its splicing status (reviewed in Stutz and Rosbash, 1998). Other viruses, like the type D retroviruses, use the cellular protein TAP/Mex67p which binds to an element in the viral RNA to export the unspliced RNA (reviewed in Nakielny and Dreyfuss, 1999). In yeast, other splicing factors involved in early steps of spliceosome assembly were subsequently found to also be involved in pre-mRNA retention, like *MUD2*, the homologue of $U2AF^{65}$, and more recently the cap binding complex, CBC (Rain and Legrain, 1997; P.J. Lopez and B. Séraphin, pers. communication). All these factors (except for Prp6, Abovich *et al*., 1990) are involved in early steps of intron recognition that precede complete spliceosome formation (reviewed in Krämer, 1996).

6. Aim of this thesis

Early steps of spliceosome assembly are important for the definition of exons and introns and commit a given pre-mRNA to splicing. Because these steps are also potential targets for the regulation of splicing and are important for pre-mRNA retention in the nucleus we were interested in factors involved in this process.

When this work was started, the function of the splicing factor 1 (SF1) had only been adressed by *in vitro* methods in mammalian nuclear extracts (Krämer, 1988; Krämer and Utans, 1991). A likely homologue in *S. cerevisae* had been identified by sequence similarity

(Arning *et al*., 1996). Given the importance of this factor in the early steps of mammalian spliceosome assembly we set out to analyze the function of its putative yeast homologue by taking advantage of the broad range of *in vitro* and *in vivo* methods avaiable in yeast.

First, we generated genetically modified yeast strains that contained the protein with different affinity tags under its endogenous promoter or under a regulatable promoter. This allowed for the depletion of the essential BBP/ScSF1 protein by genetic and/or biochemical methods and for the detection of the protein in different splicing complexes. We asked if depleted extracts would be deficient in different steps of spliceosome assembly and splicing. In addition, we probed the interaction of BBP/ScSF1 with the Mud2 protein, the closest yeast homologue of $U2AF^{65}$, and determined functional similarities and differences between the two proteins.

In a second approach, we analyzed the *in vivo* function of BBP/ScSF1. Conditional mutants of the protein were generated by error-prone PCR and relevant mutations were mapped. Extracts from the mutant strains were analyzed for the formation of splicing complexes and splicing *in vitro*. Pre-mRNA splicing and retention in the nucleus were investigated *in vivo* with a sensitive reporter system. Synthetic effects with a disruption of the nonsense-mediated decay pathway were analyzed as an indicator for the effects of pre-mRNA leakage on the cell.