Chapter 1

Introduction

Proteins are crucial for virtually all biochemical processes. Evolution has thus created an amazingly broad spectrum of protein functions ranging from enzyme catalysis and signal transduction, coordinated motion and maintenance of cellular shape to storage and transport of small molecules and ions. The importance of proteins in life has already been realised at the beginning of the 19th century, when the Swedish chemist Jöns J. Berzelius coined the term "protein" from the greek *πρωτεῖον* (first rank). Although proteins are synthesised linearly from amino-acid monomers, they fold in a to date poorly understood way into in general compact three-dimensional structures. Since only correctly folded proteins are functional, the knowledge of a protein's three-dimensional structure is crucial to fully understand its function.

The amino-acid sequence of a protein is encoded in the DNA sequence of its gene. As a consequense, on an evolutionary level new protein functions can arise in only two ways: creation of new genes and/or diversification of the use of already existing genes. The latter can result from mutation and differential expression of genes, while the former can in principle originate from three processes: first, from import of genes from other organisms (lateral gene transfer); second, from gene duplication, shuffling and/or fusion; and third, though highly unlikely, from spontaneous creation of new genes from random, non-coding DNA sequences by accumulation of mutations. With the availability of more and more genome sequences it has become apparent that novel protein functions have their origin to the largest extent in the duplication and shuffling of genes and gene fragments (Andrade *et al.*, 2001; Patthy, 1999).

One of the perhaps most astonishing outcomes of the immense efforts put into genome sequencing projects is that biological complexity seems to arise only to a minor degree from a larger numbers of genes in the genome. Instead, the amount of non-coding DNA in the genome has increased during evolution much faster than the number of genes leading to less and less compact genomes. While archea and eubacteria encompass approximately 1,000 genes per Mbp (Mega base-pair) genomic DNA, yeast has about 450, fruit-flies around 70 and for humans estimates of this number are as low as 10 (Patthy, 1999). Moreover, whereas prokaryotic genes are continuous, the genes of eukaryotes are interrupted by non-coding DNA (introns). Hence, the coding sequences (exons) in eukaryotic DNA are scattered between vast stretches of not only inter-, but also intragenic, "silent" DNA.

The existence of non-coding DNA presents many challenges to the eukaryotic gene expression machinery, since gene and exon boundaries have to be recognised correctly in a plethora of non-coding DNA. Once the gene is transcribed into a messenger RNA precursor (pre-mRNA), introns have to be removed from the primary transcript and coding regions have to be fused in an elaborate process known as splicing (Witkowski, 1988). Although an enormous amount of energy is required to replicate, maintain and express non-coding DNA sequences, their presence in the genome must obviously offer higher developed organisms evolutionary advantages (Gilbert, 1978; Herbert & Rich, 1999). Whether non-coding sequences within genes are an ancient feature of genome architecture and have been lost by non-eukaryotic organisms during evolution to stream-line their genomes (introns-early theory) (Darnell, 1978; Doolittle, 1978) or whether discontinuous genes arose late in evolution (introns-late theory) (Crick, 1978), is still a matter of intense debate.

Whatever their origins, introns have proliferated rapidly during eukaryote radiation, as they allow for two key advantageous mechanisms of protein innovation: First, splicing permits to produce different mature mRNAs (and thereby more than one protein) from the same primary gene transcript by differential joining of 5' and 3' splice sites (alternative splicing). Exons can be extended, shortened or skipped and even whole introns retained in the mature mRNA giving rise to several protein isoforms (Maniatis & Tasic, 2002). Alternative splicing thus increases considerably the functional repertoire of the eukaryotic genome without changing the total number of genes. In fact, alternative splicing is considered to be the most important source of protein diversity in vertebrates (Black, 2000; Graveley, 2001). Secondly, exon boundaries often coincide with the boundaries of independently folding and functioning protein domains[‡]. Exon shuffling and duplication provide thus a powerful means to fuse otherwise unrelated domains rapidly (in evolutionary terms) in a combinatorial fashion to so-called modular or mosaic proteins (Patthy, 1999). Furthermore, a modular protein structure confers multiple binding sites enabling their cooperation and the formation of complex regulatory networks.

The fact that a human exon comprises on average about 150 nucleotides compared to 3,500 nucleotides for an average intron (Deutsch & Long, 1999) and that about 50% of the human genes are transcribed to alternatively spliced forms (Sorek & Amitai, 2001; Modrek & Lee, 2002) underlines the importance and advantages of discontinuous eukaryotic gene structure in the innovation of protein function.

[‡] strictly, a protein domain is defined as an independently folding and functioning tertiary structure element. A protein module is a protein domain encoded by an exon that appears in the genomic DNA of two or more proteins that are otherwise unrelated to each other (Kyte, 1995).

1.1 Pre-mRNA splicing

Although introns can vary considerably in size and sequence, they contain at least three conserved motifs: the dinucleotides GU and AG at their 5' and 3' ends (5' and 3' splice-site), respectively, and an adenosine at the branch-point (BP) (Fig. 1.1(a)). During splicing these motifs are recognised specifically and play a crucial role in the two occuring transesterification reactions. The first splicing reaction results in the cleavage of the 5' splice-site (ss) generating a free 5' exon and a lariat intermediate, in which the branch-point adenosine is connected to the 5' end of the intron. In the second step, the free 3' hydroxyl group of the 5' exon attacks the phosphodiester bond at the 3' ss such that the 5' and 3' exon are fused and the intron is released (Burge *et al.*, 1999).

The splicing reaction is catalysed by the spliceosome, a sophisticated ≈4.8 MDa ribonucleoprotein complex consisting of U snRNPs (uridine-rich small nuclear ribonucleoprotein particles) and about 100 accessory proteins termed splicing factors (Burge *et al.*, 1999; Yu *et al.*, 1999). Different sets of U snRNPs and splicing factors associate in a temporally specific fashion with the mRNA substrate as it emerges from the transcription machinery. At least seven spliceosome complexes can be distinguished in yeast: the commitment complexes CC1 and CC2, the pre-spliceosome B and the mature spliceosomes A2-1, A1, A2-2, and A2-3 (Fig. 1.1(b)). As constituents of a large, dynamic ribonucleoprotein complex, spliceosomal proteins often display a modular architecture: they are built from multiple domains, some of which are catalytic (for instance kinase and phosphatase domains), but the majority of which mediates protein-RNA interactions (*e.g.* RNA-recognition motifs (RRM) and KH (hnRNP K homology) domains) or protein-protein interactions. In the spliceosome, protein-protein interactions are often mediated by arginine-serine (RS) domains, tetratricopeptide repeats (TPR), WW and FF domains.

1.1.1 The splicing factor Prp40

The core of U snRNPs is formed by the given U snRNA and a set of Sm proteins common to all snRNPs with the exception of the U6 snRNP, which contains Lsm (like-Sm) proteins. The Sm proteins are thought to assemble in a heptameric, "doughnut"-like structure around a short, conserved uridine-rich region in the U snRNA (Kambach *et al.*, 1999). In addition to the canonical Sm proteins, U snRNPs encompass tightly associated splicing factors specific to the given U snRNP. The Prp40 protein (pre-mRNA processing), which has been studied in this Thesis, is a splicing factor in the budding yeast *Saccharomyces cerevisiae* (*Sc.*) and tightly associated with the U1 snRNP (Kao & Siliciano, 1996). The U1 snRNP recognises the 5' ss and thereby commits the pre-mRNA to splicing (Fig. 1.1(b)). Prp40 is a modular protein, which comprises an N-terminal WW domain pair (Bork & Sudol, 1994) and six consecutive FF domains (Bedford & Leder, 1999) (Fig. 1.2). The region spanning both WW domains has been implicated in cross-intron bridging (Abovich & Rosbach, 1997). During

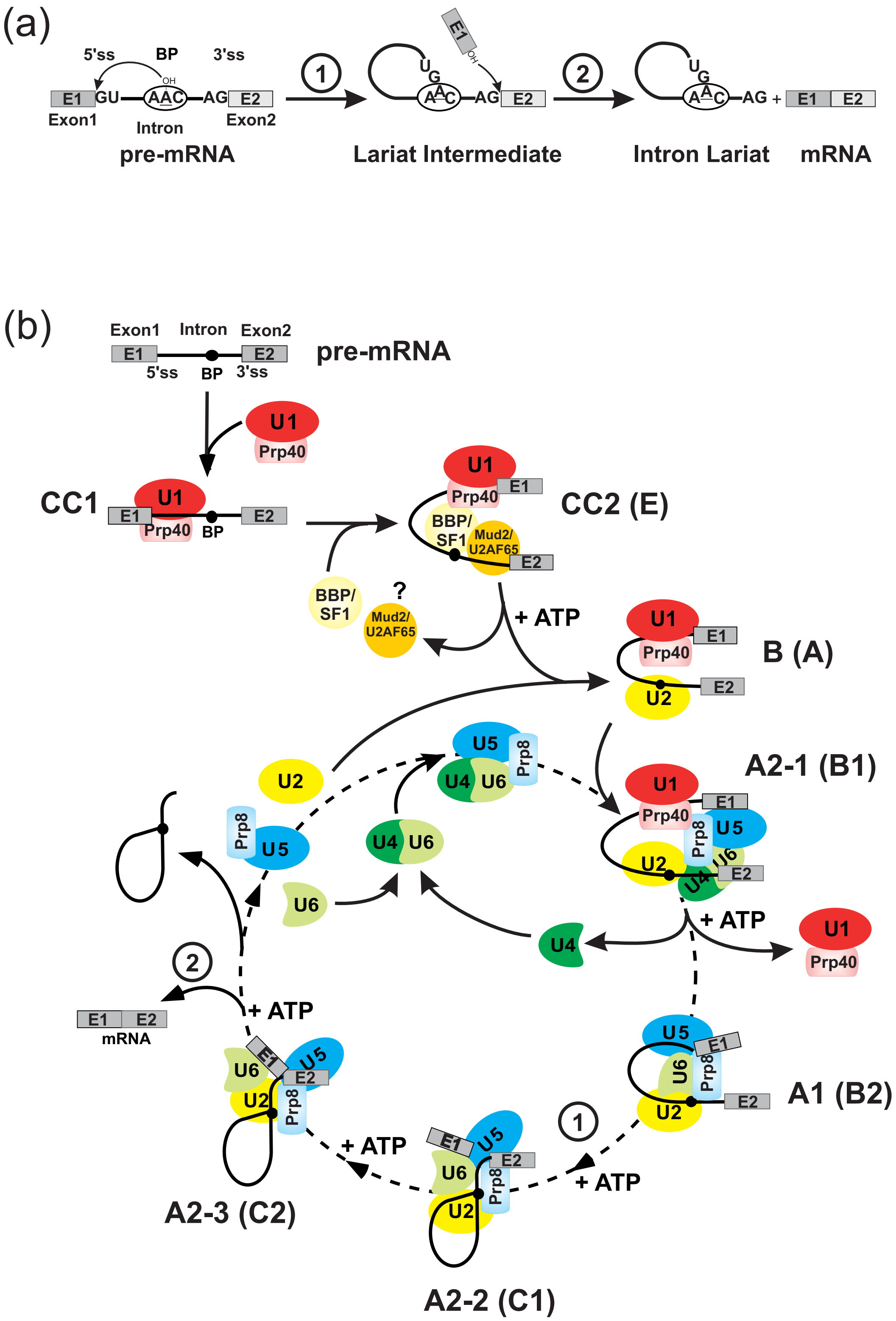

Figure 1.1: Mechanism of the splicing reaction. Schematic representation (a) of the two transesterification reactions and (b) of the spliceosome assembly and the functions of the splicing factors Prp40, BBP and Prp8. The complexes are denoted as for the yeast spliceosome, while the mammalian complexes are given in parentheses. Ellipses represent snRNPs, rounded boxes the snRNP associated splicing factors Prp40 (light red) and Prp8 (light blue), while spheres represent the auxiliary splicing factors BBP (yellow) and Mud2 (orange) in spatial proximity to Prp40. It is not known when Mud2 leaves the spliceosome, which is indicated by the question mark.

the transition from the CC1 to the CC2 splicing complex, the branch-point is specifically recognised by the branch-point binding protein (BBP/ySF1) (Berglund *et al.*, 1997). It is believed that the interaction between BBP and the WW domains of Prp40 brings the 5' ss and the branch-point in spatial proximity (Abovich & Rosbach, 1997).

Another interaction partner of Prp40 in the splicing complex is Prp8. Prp8 is the largest component of the U5 snRNP and has emerged as a tantalising candidate for a protein with a catalytic function in the spliceosome (Collins & Guthrie, 1999). Prp8 exhibits an unusually high phylogenetic conservation with two-thirds sequence identity between yeast and humans throughout the entire sequence of around 2400 amino-acids. Interestingly, the sequence of Prp8 provides to date no clues to its domain organisation. Nevertheless, Prp8 represents the only spliceosomal factor known to directly interact with all pre-mRNA sequence elements important for splicing: the 5' ss, the branch-point and the 3' ss (MacMillan *et al.*, 1994; Reyes *et al.*, 1996; Teigelkamp *et al.*, 1995; Umen & Guthrie, 1996). Prp8 possesses a proline-rich region at the N-terminus, which is believed to interact with the Prp40 WW domains (Abovich & Rosbach, 1997).

Despite the unique reactions that mRNA processing complexes catalyse, all of them have recently been found to be coupled to one another and to the transcription machinery. Capping, splicing and poly-adenylation are since regarded as co-transcriptional rather than post-transcriptional processes (Steinmetz, 1997; Proudfoot *et al.*, 2002; Maniatis & Reed, 2002). Only RNA polymerase II (Pol II) transcripts are processed to mature mRNA in the nucleus and the flexible C-terminal domain (CTD) of Pol II serves as the central interaction platform for various mRNA processing factors (Proudfoot *et al.*, 2002). The CTD is composed of tandem repeats (52 in vertebrates and 26 in yeast) of the consensus sequence YSPTSPS. The reversible phosphorylation of the CTD repeats at serine positions 2 and 5 (Dahmus, 1996) is thought to regulate the exchange of binding partners as transcription proceeds. Both the WW domains and the FF domains of Prp40 have been shown to interact with the phosphorylated CTD tail and this interaction has been suggested to link splicing and transcription in yeast (Morris & Greenleaf, 2000).

To gain insight into this complex interaction network involving Prp40, the three-dimensional structures of the N-terminal WW domain pair and the consecutive FF domain were determined in this Thesis by NMR spectroscopy (Chapter 5 and Chapter 7). Moreover, the binding sites of the WW domains and the N-terminal FF domain of Prp40 have been mapped by chemical shift titration experiments with peptides derived from the potential binding partners (Chapter 6 and 7).

1.2 Nature's LEGO bricks: Protein domains

The shuffling and fusion of independently folding and functioning protein domains to "mosaic" proteins has allowed the latter to fine-tune their biological functions and to confer multiple binding sites (Pawson, 1995; Andrade *et al.*, 2001). In general, protein domains can be classified in different families based on either similarity of sequence, three-dimensional structure or biological function. Now that the complete sequences of various organisms' genomes are available, the future task will be to ascribe a biological function to the newly identified genes.

Protein-protein interactions are believed to be mediated by a relatively small number of protein-interaction modules. The most prominent examples are Src homology 2 (SH2) and SH3 domains, PDZ domains (postsynaptic density/disc-large/ZO1), phosphotyrosine binding (PTB) domains, PH (pleckstrin homology) domains, 14-3-3 proteins, WW domains and Eps15 homology (EH) domains. To all of these domains well-defined ligand binding motifs were ascribed at first. Examples are the "PPxY" motif (where x is any residue) suggested for WW domains (Chen *et al.*, 1997) and the "PxxP" motif suggested for SH3 domains (Mayer & Eck, 1995). More recently, however, several exceptions have challenged the generality of these proposed "consensus" interaction motifs. For instance, different proline-rich ligands were discovered for WW domains including phosphoSer/Thr-Pro motifs (Kay *et al.*, 2000), while for SH3 domains interaction partners that are devoid of the consensus "PxxP" motif were found (Kang *et al.*, 2000; Barnett *et al.*, 2000). These findings suggest that the binding potential of the WW and SH3 domain family (and probably many other domain families) is larger than originally thought. The ability of protein interaction domains to mediate multiple types of binding is even more evident for domains made up of repeated units. For example, TPR repeats, HEAT motifs or WD 40 domains assemble into larger domains and thereby create extensive binding surfaces with diverse specificities (Groves & Barford, 1999; Andrade *et al.*, 2001). In general, protein domain types may thus rather share a common fold than a common specific binding motif and form subclasses with quite different targets and functions. To understand the molecular mechanisms underlying the apparent binding promiscuity of protein domain families, three-dimensional structures will have to be combined with mutagenesis studies and biophysical methods. An attempt to classify WW domains based on their three-dimensional structures and binding specificities was made in Chapter 4.

1.2.1 WW and FF domains

WW domains are the smallest, naturally occurring protein modules composed of approximately 40 amino acids. The name refers to two signature tryptophan (W) residues that are spaced 20–22 amino acids apart and are conserved in most WW domains known to date (Bork & Sudol, 1994). WW domains recognise ligands rich in prolines and fold as a stable, triple stranded anti-parallel β-sheet in absence of ligands or disulfide bridges (Chen & Sudol, 1995; Macias *et al.*, 1996). They are found in many different proteins, often localised in the cytoplasm as well as in the cell nucleus. Shortly after their characterisation, WW domains attracted attention because the signaling complexes they mediate have been implicated directly or indirectly in several human diseases including Liddle's syndrome of hypertension, muscular dystrophy, Alzheimer's and Huntington's diseases and, more recently, cancer (Sudol, 1996; Faber *et al.*, 1998).

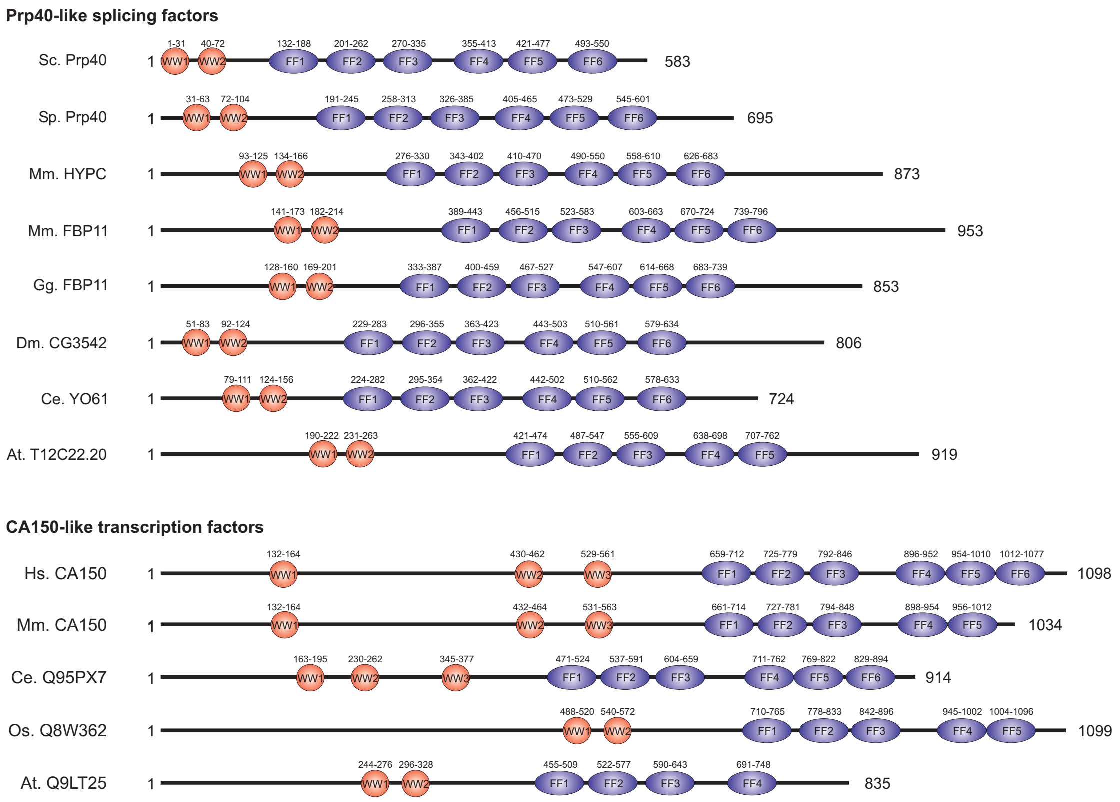

Figure 1.2: Domain architecture of Prp40-like splicing factors and CA150-like transcription factors. Red spheres represent WW domains, while blue ellipses indicate FF domains. Proteins of yet unknown function are noted by their gene names. Abbreviations used are: Hs., *Homo sapiens*; Mm., *Mus musculus*; Gg., *Gallus gallus*; Dm., *Drosophila melanogaster*; Ce., *Caenorhabditis elegans*; At., *Arabidopsis thaliana*; Os., *Oryza sativa*; Sc., *Saccharomyces cerevisiae*; Sp., *Schizosaccharomyces pombe.*

Interestingly, all Prp40 orthologues contain an N-terminal WW domain pair followed by six copies of a recently identified protein-protein interaction motif termed FF domain (Bedford & Leder, 1999) (Fig. 1.2, top). In addition, a group of transcription factors similar to the negative transcription regulator CA150 exhibits a modular architecture, which is almost identical with that of Prp40 comprising in general three N-terminal WW domains accompanied by six FF domains (Fig. 1.2, bottom). Intriguingly, CA150 and Prp40-like proteins have at least three common interaction partners: the C-terminal domain (CTD) of the largest subunit of RNA polymerase II (Suñé *et al.*, 1997; Morris & Greenleaf, 2000), the pre-mRNA splicing factor BBP/SF1 (Goldstrohm *et al.*, 2001; Abovich & Rosbach, 1997) and the Huntington's disease gene product huntingtin (Holbert *et al.*, 2001; Faber *et al.*, 1998). While the interaction with huntingtin and BBP/SF1 is mediated by the WW domains, the CTD binds to the FF domains of CA150 and Prp40. However, for Prp40 also the WW domains were shown to interact with phosphorylated CTD sequences. Moreover, the Prp40 FF domains interact with the splicing factor Clf1 (crooked neck like factor 1), which is believed to play an important role in the transition of the commitment complex to the first fully assembled splicing complex (Fig. 1.1) (Chung *et al.*, 1999). The interaction is mediated by the tetratrico peptide repeat (TPR) motifs of Clf1, which also bind to Mud2/U2AF65 at the 3'ss. However, how the FF domains interact with the CTD and the TPR repeats is unknown.

FF domains harbour two conserved phenylalanine residues and are about 60 residues in length. Secondary structure predictions and the pattern of residue conservation suggest that FF domains are α-helical. Apart from splicing and transcription factors, repeated FF domains are also present in the p190 family of GTPases (Settleman *et al.*, 1992; Burbelo, 1995), which play an important role in signal transduction pathways (Hall, 1994) to regulate cytoskeletal organisation (Lim, 1996). Since FF domains are often preceeded by WW domains, the close relation of the FF domain with WW domains and the fact that no structural information was available pointed us to determine the solution structure of the Prp40 FF domain adjacent to the WW pair, which is presented in Chapter 7.

1.3 Aim of this Thesis

At the time when this Thesis was started no structural information was available about the yeast splicing factor Prp40. However, biochemical and genetic data suggested an important function for Prp40 in bridging the 5' and the 3' splice sites by interacting with two other splicing factors, namely BBP and Prp8. Moreover, residual dipolar couplings had recently been demonstarted to be applicable in biomolecular NMR opening the way to determine three-dimensional structures of proteins with intrinsically low NOE densities, such as elongated and multi-domain proteins.

To explore how the WW domain pair of Prp40 can function in cross-intron bridging and whether their relative domain orientation allows for interactions with multiple binding partners their solution structure should be solved. Furthermore, the interactions of the Prp40 WW domains with other spliceosome components should be investigated by chemical shift titration experiments. The close relation of a novel protein-interaction domain, the FF domain, with WW domains in splicing and transcription factors pointed us to solve the three-dimensional structure of the FF domain adjacent to the WW domain pair in Prp40. Since the binding site of FF domains in general and that of the Prp40 FF1 domain in particular were unknown, chemical shift perturbation experiments should be performed with a suitable binding partner.

1.4 References

Abovich, N. & Rosbach, M. (1997). Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. *Cell,* **89**, 403–412.

Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. (2001). Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* **134** (2–3), 117–31.

Barnett, P., Bottger, G., Klein, A. T., Tabak, H. F. & Distel, B. (2000). The peroxisomal membrane protein Pex13p shows a novel mode of SH3 interaction. *EMBO J.* **19**, 6382–91.

Bedford, M. T. & Leder, P. (1999). The FF domain: a novel motif that often accompanies WW domains. *Trends Biochem. Sci.* **24** (7), 264–5.

Berglund, J. A., Chua, N., Abovich, N., Reed, R. & Rosbach, M. (1997). The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell,* **89**, 781–7.

Black, D. L. (2000). Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell,* **103**, 367–70.

Bork, P. & Sudol, M. (1994). The WW domain: a signalling site in dystrophin? *Trends Biochem. Sci.* **19** (12), 531–3.

Burbelo, P. D. (1995). p190-B, a new member of the Rho GAP family, and Rho are induced to cluster after integrin cross-linking. *J. Cell. Biol.* **270**, 30919–26.

Burge, C. B., Tuschl, T. & Sharp, P. A. (1999). Splicing of precursors to mRNAs by the spliceosome. In: *The RNA World.* pp. 525–560. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA.

Chen, H. I., Einbond, A., Kwak, S. J., H., L., Koepf, E., Peterson, S., Kelly, J. W. & Sudol, M. (1997). Characterization of the WW domain of human yes-associated protein and its polyproline-containing ligands. *J. Biol. Chem.* **272**, 17070–7.

Chen, H. I. & Sudol, M. (1995). The WW domains of Yes-associated protein binds a proline-rich ligand that differs from the consensus established for Src homology 3-binding modules. *Proc. Natl. Acad. Sci. USA,* **92**, 7819–7823.

Chung, S., McLean, M. R. & Rymond, B. C. (1999). Yeast ortholog of the Drosophila crooked neck protein promotes spliceosome assembly through stable U4/U6.U5 snRNP addition. *RNA,* **5** (8), 1042–54.

Collins, C. A. & Guthrie, C. (1999). Allele-specific genetic interactions between Prp8 and RNA active site residues suggest a function for Prp8 at the catalytic core of the spliceosome. *Genes* & *Dev.* **13**, 1970–82.

Crick, F. (1978). Split genes and RNA splicing. *Science,* **204**, 264–71.

Dahmus, M. (1996). Reversible phosphorylation of the C-terminal domain of RNA polymerase II. *J. Bio. Chem.* **271**, 19009–12.

Darnell, J. E. (1978). Implications of RNA.RNA splicing in evolution of eukaryotic cells. *Science,* **202**, 1257–60.

Deutsch, M. & Long, M. (1999). Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.* **27**, 3219–28.

Doolittle, R. F. (1978). Genes in pieces: were they ever together? *Nature,* **272**, 581–82.

Faber, P. W., Barnes, G. T., Srinidhi, J., Chen, J., Gusella, J. F. & MacDonald, M. E. (1998). Huntingtin interacts with a family of WW domain proteins. *Hum. Mol. Genet.* **7**, 1463–74.

Gilbert, W. (1978). Why genes in pieces? *Nature,* **271**, 501.

Goldstrohm, A. C., Albrecht, T. R., Suñé, C., Bedford, M. T. & Garcia-Blanco, M. A. (2001). The transcription elongation factor CA150 interacts with RNA polymerase II and the pre-mRNA splicing factor SF1. *Mol. Cell. Biol.* **21**, 7617–28.

Graveley, B. R. (2001). Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* **17**, 100–7.

Groves, M. R. & Barford, D. (1999). Topological characteristics of helical repeat proteins. *Curr. Op. Struct. Biol.* **9**, 383–9.

Hall, A. (1994). Small GTP-binding proteins and the regulation of the actin cytoskeleton. *Annu. Rev. Cell. Biol.* **10**, 31–54.

Herbert, A. & Rich, A. (1999). RNA processing in evolution: the logic of soft-wired genomes. *Annals New York Acad. Sci.* **870**, 119–32.

Holbert, S., Denghien, I., Kiechle, T., Rosenblatt, A., Wellington, C., Hayden, M. R., Margolis, R. L., Ross, C. A., Dausset, J., Ferrante, R. J. & Neri, C. (2001). The Gln-Ala repeat transcriptional activator CA150 interacts with huntingtin: neuropathologic and genetic evidence for a role in Huntington's disease pathogenesis. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 1811–6.

Kambach, C., Walke, S. & Nagai, K. (1999). Structure and assembly of the spliceosomal small nuclear ribonucleoprotein particles. *Curr. Op. Struct. Biol.* **9**, 222–230.

Kang, H., Freund, C., Duke-Cohan, J. S., Musacchio, A., Wagner, G. & Rudd, C. E. (2000). SH3 domain recognition of a proline-independent tyrosine-based RKxxYxxY motif in immune cell adaptor SKAP55. *EMBO J.* **19**, 2889–99.

Kao, H. Y. & Siliciano, P. G. (1996). Identification of Prp40, a novel essential yeast splicing factor associated with the U1 small nuclear ribonucleoprotein particle. *Mol. Cell. Biol.* **16** (3), 960–7.

Kay, B. K., Williamson, M. P. & Sudol, M. (2000). The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. *FASEB J.* **14**, 231–241.

Kyte, J. (1995). Structure in protein chemistry. pp. 264–272. Garland Publishing, Inc., New York, USA.

Lim, L. (1996). Regulation of phosphorylation pathways by p21 GTPases. The p21 Ras-related Rho subfamily and its role in phosphorylation signalling pathways. *Eur. J. Biochem.* **242**, 171–84.

Macias, M. J., Hyvönen, M., Baraldi, E., Schultz, J., Sudol, M., Saraste, M. & Oschkinat, H. (1996). Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide. *Nature,* **382**, 646–649.

MacMillan, A. M., Query, C. C., Allerson, S., Chen, G. L., Verdine, G. L. & Sharp, P. A. (1994). Dynamic association of proteins with the pre-mRNA branch region. *Genes & Dev.* **8**, 3008–20.

Maniatis, T. & Reed, R. (2002). An extensive network of coupling among gene expression machines. *Nature,* **416**, 499–506.

Maniatis, T. & Tasic, B. (2002). Alternative pre-mRNA splicing and the proteome expansion in metazoans. *Nature,* **418**, 236–43.

Mayer, B. J. & Eck, M. J. (1995). SH3 domains. Minding your p's and q's. *Curr. Biol.* **5**, 364–7.

Modrek, B. & Lee, C. A. (2002). A genomic view of alternative splicing. *Nat. Genet.* **30**, 13–9.

Morris, D. P. & Greenleaf, A. L. (2000). The splicing factor, Prp40, binds the phosphorylated carboxyl-terminal domain of RNA polymerase II. *J. Biol. Chem.* **275** (51), 39935–43.

Patthy, L. (1999). Genome evolution and the evolution of exon-shuffling – a review. *Gene,* **238**, 103–14.

Pawson, T. (1995). Protein modules and signalling networks. *Nature,* **373**, 573–80.

Proudfoot, N. J., Furger, A. & Dye, M. J. (2002). Integrating mRNA processing with transcription. *Cell,* **108**, 501–512.

Reyes, J. L., Kois, P., Konforti, B. B. & Konarska, M. M. (1996). The canonical GU dinucleotide at the 5' splice site is recognized by p220 of the U5 sn RNP within the spliceosome. *RNA,* **2**, 213–25.

Settleman, J., Albright, C., Foster, L. & Weinberg, R. (1992). Association between GTPase activators for Rho and Ras families. *Nature,* **359**, 153–4.

Sorek, R. & Amitai, M. (2001). Piecing together the significance of splicing. *Nat. Biotech.* **19**, 196.

Steinmetz, E. J. (1997). Pre-mRNA processing and the CTD of RNA polymerase II: The tail that wags the dog? *Cell,* **89**, 491–4.

Sudol, M. (1996). Structure and function of the WW domain. *Prog. Biophys. Molec. Biol.* **65**, 113–32.

Suñé, C., Hayashi, T., Liu, Y., Lane, W., Young, R. & Garcia-Blanco, M. (1997). CA150, a nuclear protein associated with the RNA polymerase II holoenzyme, is involved in Tat-activated human immunodeficiency virus type 1 transcription. *Mol. Cell. Biol.* **17**, 6029–39.

Teigelkamp, S., Newman, A. J. & Beggs, J. D. (1995). Extensive interactions between the PRP8 protein with the 5' and 3' splice sites during splicing suggests a role in stabilization of exon alignment by U5 snRNA. *EMBO J.* **14**, 2602–12.

Umen, J. G. & Guthrie, C. (1996). Mutagenesis of the yeast gene PRP8 reveals domains governing the specificity and fidelity of 3' splice site selection. *Genetics,* **143**, 723–39.

Witkowski, J. A. (1988). The discovery of "split" genes: a scientific revolution. *Trends Biochem. Sci.* **10**, 110–3.

Yu, Y.-T., Scharl, E. C., Smith, C. M. & Steitz, J. A. (1999). The growing world of small nuclear ribonucleoproteins. In: *The RNA World.* pp. 487–524. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, USA.