Molecular basis of RNA recognition and TAP binding by the SR proteins SRp20 and 9G8

Yann Hargous^{1,8}, Guillaume M Hautbergue^{2,8}, Aura M Tintaru^{2,3,8}, Lenka Skrisovska¹, Alexander P Golovanov³, James Stevenin^{4,5,6,7}, Lu-Yun Lian^{3,9}, Stuart A Wilson^{2,*} and Frédéric H-T Allain^{1,*}

¹Institute of Molecular Biology and Biophysics, ETH Zurich, Zurich, Switzerland, ²Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield, UK, ³Faculty of Life Science, University of Manchester, Manchester, UK, ⁴IGBMC, Department of Transcription, Illkirch, France, ⁵Insertion U596, Illkirch, France, ⁶CNRS UMR7104, Illkirch, France and ⁷University of Strasbourg, Strasbourg, France

The sequence-specific RNA-binding proteins SRp20 and 9G8 are the smallest members of the serine- and arginine-rich (SR) protein family, well known for their role in splicing. They also play a role in mRNA export, in particular of histone mRNAs. We present the solution structures of the free 9G8 and SRp20 RNA recognition motifs (RRMs) and of SRp20 RRM in complex with the RNA sequence 5'CAUC3'. The SRp20-RNA structure reveals that although all 4 nt are contacted by the RRM, only the 5' cytosine is primarily recognized in a specific way. This might explain the numerous consensus sequences found by SELEX (systematic evolution of ligands by exponential enrichment) for the RRM of 9G8 and SRp20. Furthermore, we identify a short arginine-rich peptide adjacent to the SRp20 and 9G8 RRM, which does not contact RNA but is necessary and sufficient for interaction with the export factor Tip-associated protein (TAP). Together, these results provide a molecular description for mRNA and TAP recognition by SRp20 and 9G8.

The EMBO Journal (2006) 25, 5126–5137. doi:10.1038/sj.emboj.7601385; Published online 12 October 2006

Subject Categories: RNA; structural biology

Keywords: mRNA export; NMR; RRM; SR protein; TAP

Introduction

The family of serine- and arginine-rich (SR) proteins has long been identified as key regulators of alternative splicing (Fu, 1995; Bourgeois *et al*, 2004). They also play a role in other

stages in the gene expression pathway, including mRNA stability (Zhang and Krainer, 2004) and translation (Sanford *et al*, 2004), and recently three SR proteins, namely SRp20, 9G8 and ASF/SF2, were shown to function as mRNA export factors (Huang and Steitz, 2001; Huang *et al*, 2003). SRp20 and 9G8 are the smallest members of the SR protein family (Zahler *et al*, 1992; Cavaloc *et al*, 1994), with a single N-terminal RNA recognition motif (RRM), which is 80% identical between the two proteins. In SRp20, the SR-rich domain immediately follows the RRM, whereas in 9G8, a zinc-knuckle is present between the RRM and the SR domain. The zinc-knuckle modifies the RNA-binding specificity of 9G8 compared to SRp20, as 9G8 binds a purine-rich sequence AGAC(G/U)ACGA(C/U), whereas SRp20 or 9G8 lacking the zinc-knuckle binds a pyrimidine-rich sequence with a consensus (A/U)C(A/U)(A/U)C (Heinrichs and Baker, 1995; Cavaloc *et al*, 1999). These consensus sequences identified *in vitro* closely match several naturally occurring pre-mRNA sequences bound *in vivo* by 9G8 and SRp20. These include the *Drosophila* doublesex gene (*dsx*) (Heinrichs and Baker, 1995; Lynch and Maniatis, 1996), CD44 alternative exon v9 (Galiana-Arnoux *et al*, 2003), SRp20 exon 4 (Jumaa and Nielsen, 2000), calcitonin/caltonin gene-related peptide (Lou *et al*, 1998) and probably many other exonic sequences where these consensus sequences serve as exonic splicing enhancers (Schaal and Maniatis, 1999; Fairbrother *et al*, 2002). In addition, the consensus sequences for both proteins are found within a 22 nt RNA element of the histone H2A mRNA. In this context, both 9G8 and SRp20 bind histone H2A mRNA and may individually or together promote its nucleocytoplasmic export through their ability to interact with the Tip-associated protein (TAP) mRNA export factor (Huang *et al*, 2003).

Despite being studied intensely since their discovery 14 years ago, there are no published structures for canonical SR proteins. This may in part be explained not only by the poor solubility of these proteins in their free state, but also by the degenerate RNA-binding sequences recognized by SR proteins, which may have prevented their study in the bound form. We have surmounted the solubility problem for the RRM of SRp20 and 9G8 by studying the proteins either fused with the immunoglobulin G-binding domain 1 of Streptococcal protein G (GB1) solubility tag (Zhou *et al*, 2001) or in a solution containing charged amino acids (aa) (Golovanov *et al*, 2004). This allowed us to determine the structures of the free 9G8 and SRp20 RRM and of SRp20 RRM in complex with the RNA sequence CAUC using NMR spectroscopy. These represent the first structures for canonical SR proteins, which play a vital role in eukaryotic gene expression. The β -sheet of their RRM is very hydrophobic with two phenylalanines, a tyrosine, an alanine and a tryptophan exposed on its surface. The RRM β -sheet surface of SRp20 is used to bind all 4 nt of the CAUC sequence in a semi-sequence-specific manner with only the 5' cytosine recognized specifically by the RRM. We have also identified a short

*Corresponding authors. FH-T Allain, Institute of Molecular Biology and Biophysics, ETH Zurich, Swiss Federal Institute of Technology (ETH) Zurich, 8093 Zurich, Switzerland. Tel.: +41 1 633 39 40; Fax: +41 1 633 12 94; E-mail: allain@mol.biol.ethz.ch or SA Wilson, Department of Molecular Biology and Biotechnology, University of Sheffield, Firth Court, Sheffield S10 2TN, UK. Tel.: +44 114 222 2849; Fax: +44 114 222 2800; E-mail: stuart.wilson@sheffield.ac.uk

⁸These authors contributed equally to this work

⁹Present address: School of Biological Sciences, Biosciences Building, University of Liverpool, Liverpool L69 7ZB, UK

Received: 30 May 2006; accepted: 18 September 2006; published online: 12 October 2006

arginine-rich peptide C-terminal of the RRM in 9G8 and SRp20, which interacts with TAP but not with the RNA. Altogether, these data provide the first molecular details of mRNA and TAP recognition by canonical SR proteins.

Results

Solution structures of 9G8 and SRp20 RRM

A fragment consisting of aa 12–98 of 9G8 (Figure 1A) was used for structural studies, which includes the RRM domain and a C-terminal arginine-rich region situated just before the zinc-knuckle (Cavaloc *et al*, 1999). This fragment has good expression and solubility properties in a buffer containing L-Arg and L-Glu (Golovanov *et al*, 2004); however, without these additives, its solubility was 10 times less. Moreover, this fragment retains RNA- and TAP-binding properties, and thus is functionally competent. The solution structure of the 9G8 fragment was calculated based on 951 non-trivial NOE-derived interproton distance constraints, 68 hydrogen bond and 119 dihedral angle restraints. The RRM domain structure was determined with high precision (pairwise RMSD for the 20 conformers of the final ensemble are 0.58 and 1.51 Å for the backbone and heavy atoms, respectively; Table I and Figure 1B). The RRM of 9G8 adopts the classical $\beta\alpha\beta\alpha\beta$ topology with two α -helices packed against one side of the four-stranded β -sheet (Maris *et al*, 2005). The RRM presents an unusually large hydrophobic core on the β -sheet surface, which is composed of Tyr 14 (β 1), Phe 51 and Phe 49 (β 3), and Trp 41 and Ala 43 (β 2) (Figure 1C and D). The conformation of the α 2– β 4 loop is relatively well defined, as the tip of the loop forms a number of hydrophobic contacts (involving side chains of Ile 72, Cys 73, Ser 75) with the adjacent β -sheet, thus protruding into the cleft between the two α -helices. The conformation of the β 1– α 1 and β 2– β 3 loops is less well defined, which is not uncommon for solution structures of RRM. The narrow signal linewidth and small chemical shift dispersion observed in the ^1H - ^{15}N -correlated (HSQC) NMR spectrum reveal that the C-terminal arginine-rich region, required for TAP binding (Figure 1A and below), is unstructured and highly flexible.

We have also determined the structure of the SRp20 RRM. To overcome solubility problems, we fused the RRM domain (aa 1–86) of human SRp20 to the 58-aa GB1 solubility tag (Zhou *et al*, 2001). The structure of GB1-SRp20 RRM was calculated following the assignment of 1832 NOE cross-peaks including 690 constraints for the RRM only. This resulted in an RRM structure with a pairwise RMSD for the 19 conformers of the ensemble of 1.24 and 2.12 Å for the backbone and heavy atoms, respectively (Table II and Figure 1E). The RRM of SRp20 adopts the expected $\beta\alpha\beta\alpha\beta$ topology (Maris *et al*, 2005) and a structure very similar to free 9G8 (Figure 1). The only structural difference resides in the loops between α -helix 2 and β 4, where several amino-acid changes are found between the two proteins (Figure 1A).

Solution structure of SRp20 RRM in complex with CAUC

To understand the molecular basis for RNA recognition by the SRp20 RRM, we studied its interaction with a large number of RNA oligonucleotides using NMR spectroscopy. The RNAs were 3–8 nt long and corresponded to several natural binding sequences for SRp20 (Lou *et al*, 1998; Huang and Steitz, 2001) and consensus sequences identified

by *in vitro* (Cavaloc *et al*, 1999) or *in vivo* systematic evolution of ligands by exponential enrichment (SELEX) (Schaal and Maniatis, 1999; Supplementary Table 1). For all RNAs containing a CUUC or a CAUC sequence within an RNA length of 5 nt or above, we obtained remarkably similar ^1H - ^{15}N HSQC spectra with several large shifts compared to the free protein spectrum and a good linewidth (Figure 2A and Supplementary Figure 1). These shifts indicate that the RNA binds the β -sheet surface of the RRM (Figure 2B). However, in the NMR spectra of these complexes, the resonance lines of the RNA aromatic protons were either very broad or absent (i.e. broadened beyond detection). Such line broadening is frequent in protein–nucleic acid complexes and originates from exchange between different conformations at the interface. This broadening makes structure determination of the RNA in complex impossible under these conditions. With an RNA of only 4 nt (CUUC or CAUC), binding is still observed but the affinity is decreased, as indicated by a smaller chemical shift difference between the free and bound form of the protein at a 1:1 ratio compared to larger RNA (Supplementary Figure 2). Nevertheless, in these complexes, the RNA resonances now become observable, and changes in chemical shifts upon binding indicate that all 4 nt appear to bind the protein (Figure 2C). As the complex is in fast exchange, the RNA resonance linewidths are averaged between the sharp linewidths of the free state and the broader ones of the bound state. It was only in such conditions that structure determination was possible as both the protein and RNA resonances presented good linewidths. Similar results were obtained for the RRM of 9G8 (aa 1–104) in complex with RNA, confirming that the highly related RRM binds RNA in the same manner (compare Figure 2A and D). We therefore pursued the structure determination of SRp20 RRM in complex with CAUC.

The structure of the SRp20 RRM:CAUC complex was calculated following the assignment of 2173 NOE cross-peaks, including 40 intermolecular ones. This resulted in a structure with a pairwise RMSD for the 29 conformers of the ensemble of 1.21 and 2.06 Å for the backbone and heavy atoms, respectively (Table II and Figure 3A). Within the RRM:CAUC complex, the RRM adopts the same structure as the free protein with the same differences compared to 9G8 (Figure 1). In the complex, the RNA completely covers the unusually large hydrophobic surface of the β -sheet of the RRM with each of the four bases contacting one of the four aromatic residues. This extensive hydrophobic interaction surface explains most of the affinity of the RNA for the protein ($\approx 18 \mu\text{M}$; Supplementary Figure 3). Here, C1 and A2 stack on the conserved RNP2 (Y13 in β 1) and RNP1 (F50 in β 3) aromatic side chains, respectively, and the RNP1 residue F48 (β 3) inserts between the sugar rings of C1 and A2 (Figure 3B and C). This binding mode is common to almost all RRM–RNA complexes (Maris *et al*, 2005), although here A2 adopts an unusual *syn* conformation that was previously observed only for guanine in this position (Maris *et al*, 2005; Auweter *et al*, 2006). The conformation of U3 and C4 is more unusual, as U3 is bulged out whereas C4 stacks partially over A2. U3 interacts with Phe 48 (β 3), Trp 40, Ala 42 (both β 2) and with the β 2– β 3 loop of the RRM, whereas C4 is maintained in this position by a hydrogen bond between C4 amino and A2 2' oxygen (Figure 3C).

Figure 1 Amino-acid sequences of 9G8, SRp20 and RBP1 and overviews of the solution structures of the RRM for 9G8 and SRp20. (A) Amino-acid sequences of human 9G8, SRp20 and *Drosophila* RBP1. Only the first 100 aa are shown. The amino-acid number and the secondary structure are indicated in blue for 9G8 and black for SRp20. Amino acids in red are involved in the SRp20–CAUC complex. The arginines in SRp20 and 9G8, whose mutation into glutamate abolishes TAP binding, are highlighted in green. (B) Overlay of the final 20 structures of 9G8 RRM superposed on the backbone atoms of the structured parts of the protein. The protein backbone is red. Only the ordered region of the protein (aa 10–82) is shown. (C) The lowest energy structure in ribbon (protein backbone) and stick representation. Important protein side chains interacting with RNA in the complex are represented as green sticks. (D) Surface representation of the 9G8 RRM, indicating residues interacting with RNA by homology with the SRp20:RNA complex. (E) Overlay of the final 19 structures of SRp20 RRM superposed on the backbone atoms of the structured parts of the protein. The protein backbone is gray. Only the ordered region of the protein (aa 9–81) is shown. (F) A representative conformer of SRp20 in ribbon (protein backbone) and stick representation. Important protein side chains interacting with RNA in the complex are represented as green sticks. (G) Surface representation of the SRp20 RRM, indicating residues interacting with RNA by homology with the SRp20:RNA complex.

Table 1 NMR and refinement statistics of free 9G8 RRM

	9G8 RRM
<i>NMR distance and dihedral constraints</i>	
Distance restraints	
Total NOE	951
Intra-residue	209
Inter-residue	742
Sequential ($ i-j = 1$)	330
Non-sequential ($ i-j > 1$)	412
Hydrogen bonds	68
Dihedral angle restraints	119
<i>Structure statistics^a</i>	
Violations (mean and s.d.)	
Number of distance restraint violations $> 0.3 \text{ \AA}$	0.95 ± 1.0
Maximum distance constraint violation ($\text{\AA}$)	0.33 ± 0.07
Deviations from idealized geometry	
Bond lengths ($\text{\AA}$)	0.011 ± 0.0001
Bond angles (deg)	2.27 ± 0.03
Average pairwise RMSD ($\text{\AA}$) ^b	
9G8 RRM (10–81)	
Heavy	1.51 ± 0.17
Backbone	0.58 ± 0.13
<i>Ramachandran statistics^a</i>	
Residues in most favored regions	77.9%
Residues in additionally allowed regions	19.9%
Residues in generously allowed regions	2.0%
Residues in disallowed regions	0.1%

^aCalculated for an ensemble of the 20 lowest energy structures.**SRp20 RRM recognizes CAUC sequence specifically but only partially**

Considering the different SELEX consensus sequences obtained for SRp20 RRM, C-A/U-A/U-C emerges as the core binding sequence (Cavaloc *et al*, 1999; Schaal and Maniatis, 1999). Our structure confirms that 4 nt can be accommodated on the β -sheet surface of the RRM, but the recognition appears only partially sequence specific. C1 is clearly recognized specifically by the RRM, whereas the selective recognition of C4 is less evident from the structure. Indeed, four intermolecular hydrogen bonds mediate the recognition of C1 (Figure 3C and D). The C1 amino protons are hydrogen-bonded with Leu 80 backbone carbonyl oxygen and Glu 79 side-chain carboxyl oxygen, C1 N3 is hydrogen-bonded with Asn 82 amide and finally C1 O2 is hydrogen-bonded with Ser 81 hydroxyl (Figure 3C and D). For C4, Lys 11 ($\beta 1$) is found close to the base in some conformers, which could explain a preference for cytosine in this position via contacts to C4 O2 and N3 (Figure 3C), but the severe line broadenings of the resonances for this side chain prevent its accurate positioning. The sequence-specific recognition of C1 was independently confirmed by binding measurements of SRp20 with GAUC, which has approximately 10-fold lower affinity than CAUC (Supplementary Figure 3).

On the basis of the consensus recognition sequence for SRp20, one would not expect perfect sequence-specific recognition for A2 and U3 and the structure is consistent with this. Two hydrogen bonds stabilize A2 (one between its amino proton and Ser 81 hydroxyl oxygen and another between its H2 and C1 2' hydroxyl oxygen; Figure 3C) and these contacts partially explain why an adenine is preferred at this second position over a guanine, but we found no interacting side chains in the protein that could favor the adenine

over a uracil or a cytosine. For U3, although one hydrogen bond stabilizes U3 (between its O2 and Asn 44 amino; Figure 3D), any nucleotide could be accommodated in this binding pocket with the same contact (involving O2 for a cytosine and N3 for a guanine or an adenine). Thus, the structure supports earlier findings indicating that SRp20-RNA recognition is only semi-sequence-specific (Cavaloc *et al*, 1999; Schaal and Maniatis, 1999).

A short peptide C-terminal of the RRM provides a binding site for TAP in SRp20 and 9G8

Having defined the regions of SRp20 responsible for binding RNA, we next sought to identify how SRp20 binds the TAP mRNA export factor. Previous work has shown that aa 1–105 for 9G8, including the RRM and flanking sequences, were sufficient to support the interaction. However, the regions of SRp20 responsible for interaction with TAP were not defined (Huang *et al*, 2003). We used a glutathione S-transferase (GST)-TAP/p15 fusion protein immobilized on glutathione Sepharose beads together with radiolabeled SRp20 truncations in a pull-down binding assay to define the regions of SRp20 required for interaction with TAP. We found that full-length SRp20 bound GST-TAP/p15 beads efficiently in this assay, whereas the RRM did not, suggesting that the TAP- and RNA-binding sites on SRp20 were distinct (Figure 4A and B). The presence of a short arginine-rich peptide adjacent to the RRM of SRp20 restored the interaction with GST-TAP/p15 beads to the same levels seen with full-length SRp20 (Figure 4B, lanes 1 and 3), indicating that this motif was important for TAP interactions. Within this motif, the arginines stood out, as the TAP-binding domain in the REF2-I mRNA export adaptor is arginine rich (Rodrigues *et al*, 2001). A single mutation to glutamate of each of these arginines in SRp20 completely prevented TAP binding, demonstrating their importance for this interaction (Figure 4B). To confirm that the potential TAP-binding motif functioned in isolation, we fused it to GST and found that it bound TAP-p15 efficiently in a pull-down assay (Figure 4C).

A similar analysis was undertaken with 9G8 and it was shown that a short arginine-rich peptide (aa 81–98), which lies between the RRM and the zinc-knuckle, was sufficient for interaction with TAP-p15 (Figure 4D–F). Within this sequence, the arginine dipeptides at positions 87, 88 and 97, 98 were required for the interaction, whereas Arg 90, 93 were not. These data indicate that 9G8 and SRp20 harbor a TAP-binding motif, which consists of an arginine-rich peptide, which from the NMR analysis, appears flexible and is tightly juxtaposed with their RRMs.

To confirm the functional importance of the TAP-binding motif, firstly we tested the ability of full-length SRp20 and a mutant form (R88E) to interact with TAP by co-immunoprecipitation (Figure 5A), using extracts from transiently transfected human 293T cells treated with ribonuclease A and alkaline phosphatase. Whereas wild-type SRp20 bound TAP efficiently, the R88E point mutation in the TAP-binding motif completely abrogated binding. These data indicate that the extensive arginine-serine-rich domain C-terminal of the TAP-binding motif cannot substitute for this motif in supporting TAP interactions. Secondly, we used a tethered mRNA export assay to test whether the TAP-binding motif in SRp20 was a transferable signal. This assay is based on the nuclear export of an unspliced reporter RNA artificially tethered to an export

Table II NMR and refinement statistics for the free GB1-SRp20 RRM and bound to CAUC

	Free GB1-SRp20	SRp20 RRM-CAUC RNA complex	
		GB1-SRp20 RRM	CAUC
<i>NMR distance and dihedral constraints^a</i>			
Distance restraints			
Total NOE	1832 (690)	2101 (692)	32
Intra-residue	318 (115)	398 (131)	
Inter-residue	1514 (575)	1703 (543)	
Sequential ($ i-j = 1$)	523 (225)	537 (177)	
Non-sequential ($ i-j > 1$)	991 (350)	1166 (376)	
Hydrogen bonds	47	47	
Protein-RNA intermolecular		40	
Dihedral angle restraints RNA (δ)			4
<i>Structure statistics^b</i>			
Violations (mean and s.d.)			
Number of distance restraint violations $> 0.3 \text{ \AA}$	8.5 ± 1.7	8.6 ± 1.8	
Maximum distance constraint violation ($\text{\AA}$)	0.74 ± 0.10	0.52 ± 0.08	
Deviations from idealized geometry			
Bond lengths ($\text{\AA}$)	0.011 ± 0.0002	0.011 ± 0.0001	
Bond angles (deg)	2.58 ± 0.035	2.6 ± 0.032	
Average pairwise RMSD ($\text{\AA}$) ^b			
Gb1 (1-56)			
Heavy	1.29 ± 0.22	0.99 ± 0.14	
Backbone	0.70 ± 0.17	0.44 ± 0.11	
SRp20 RRM (75-146)			
Heavy	2.12 ± 0.35	2.06 ± 0.33	
Backbone	1.24 ± 0.29	1.17 ± 0.26	
RNA			
Heavy		1.03 ± 0.20	
Complex SRp20 RRM + CAUC			
Heavy		2.00 ± 0.26	
Backbone		1.21 ± 0.25	
<i>Ramachandran statistics^a</i>			
Residues in most favored regions	64.4%	72.4%	
Residues in additional allowed regions	28.6%	22.1%	
Residues in generously allowed regions	4.3%	3.6%	
Residues in disallowed regions	2.7%	1.9%	

^aIn parentheses are the statistics for the RRM alone.

^bCalculated for an ensemble of the 19 and 29 lowest energy structures for free SRp20 and in complex, respectively.

factor by the bacteriophage MS2 RNA-binding coat protein (MS2) and operator RNA sequences (Figure 5B). The reporter RNA contains luciferase within an inefficiently spliced intron derived from HIV-1. The spliced RNA, which is exported normally, loses the luciferase gene and does not lead to luciferase expression. The unspliced RNA containing the luciferase gene is normally actively retained in the nucleus. However, the direct tethering of an mRNA export factor via the MS2 coat protein/operator can override nuclear retention, leading to export of the unspliced RNA, which in turn is translated giving rise to luciferase activity, thus providing an assay for mRNA export (Wiegand *et al*, 2003; Williams *et al*, 2005). The expression levels for all MS2 fusions were tested by Western blotting and found to be equivalent except the REF2-I RRM and SRp20-REF2-I-RRM chimera, which consistently gave low expression levels (Figure 5C). Nuclear staining was also observed for all MS2 fusions by indirect immunofluorescence (not shown). When full-length REF2-I is tethered to an unspliced mRNA, it promotes export of that RNA, presumably by recruiting TAP, whose direct tethering leads to significantly greater export (Wiegand *et al*, 2003) (Figure 5D). However, the REF2-I RRM, lacking the N- and C-terminal TAP-binding sites (Rodrigues *et al*, 2001), does not function in this assay. MS2-SRp20 (aa 1–90) is also

nonfunctional in this assay (Figure 5D; Huang *et al*, 2004). Normally, SRp20 is phosphorylated on serines in the cytoplasm and binds to TAP alone once it has been dephosphorylated during nuclear splicing (Huang *et al*, 2004; Lai and Tarn, 2004). As MS2-SRp20 (aa 1–90) retains a serine near its C-terminus, it may be phosphorylated *in vivo*. As an MS2 fusion protein, it will directly bind the pre-mRNA, whose export is being monitored, thus bypassing splicing-dependent dephosphorylation, which may render it incapable of binding TAP. This may explain why it is unable to promote the export of the unspliced reporter RNA. In contrast, a chimera between the TAP-binding motif from SRp20 and the REF RRM was functional, despite poor expression of the chimeric protein (Figure 5C). This indicates that even low steady-state levels of a functional export adaptor can promote export in this assay; thus, the REF2-I RRM, whose expression was also poor, most likely does not function as it lacks TAP-binding sites. The ability of the REF2-I-SRp20 chimera to promote export suggests that it is not recognized as a substrate by SR protein kinases and evades serine phosphorylation. Together, these data indicate that the TAP-binding motif of SRp20 is transferable and can stimulate mRNA export *in vivo* in the context of a chimera with a defective REF2-I.

Figure 2 RNA containing CAUC binds to the RRM of SRp20 and 9G8 and affects residues in the β -sheet. (A) ^{15}N HSQC spectra of ~ 1 mM solutions of free GB1-SRp20 RRM (red) and that bound in the presence of one equivalent of 5'UCAUC3' (blue) at 315 K. (B) The changes in chemical shifts of the backbone amide nitrogens and protons between free and bound RRM are plotted versus the amino-acid number. Large chemical shift changes occur in the β -strands as well as in the region immediately after $\beta 4$ (X represents the position of proline residues). (C) Sections of 2D TOCSY spectra showing the H5–H6 correlations of C1, U2 and C4 for ~ 1 mM solutions of free 5'CAUC3' (red) and of 5'CAUC3' in the presence of one-fifth (black) and one equivalent of protein (blue). (D) ^{15}N HSQC spectra for ~ 1 mM solutions of free GB1-9G8 RRM (red) and that bound in the presence of one equivalent of 5'CUCUUCAC3' (blue) at 315 K.

Discussion

Structure of SR proteins: a technical challenge

The SR proteins have been an elusive structural target for many years, probably owing in part to their intrinsic insolubility. By overcoming the solubility problems, either by fusing the RRM to the solubility tag GB1 (Zhou *et al*, 2001) or by studying it in the presence of charged amino acids (Golovanov *et al*, 2004) (see Materials and methods), we provide the first molecular structures for this important class of proteins. The structures of 9G8 and SRp20 RRMs revealed an unusually large exposed hydrophobic surface (Figure 1B), which can partly explain the low solubility of these RRMs. Studying SRp20 RRM in complex with RNA presented additional technical challenges. Although it was well established that SRp20 is a sequence-specific RNA-binding protein, like

most SR proteins, it binds a degenerate RNA consensus sequence, making target sequence choices difficult (Tacke and Manley, 1999; Bourgeois *et al*, 2004). In this context, 13 different sequences, ranging in size from three to 8 nt, covering the different consensus sequences established for SRp20 (Supplementary Table S1), were tested. With RNA sequences of 5 nt or above, the binding affinity is higher (Supplementary Figure 2), but the RNAs experience severe line broadening originating from conformational exchange at the interface and, for long RNA sequences, from exchange between multiple binding registers when more than one cytosine is present in the sequence. On decreasing the RNA length to 4 nt and using the CAUC sequence, the affinity decreases, but the RNA resonances become observable, making it possible to determine the structure of the complex. This approach is similar to the one we used to study the four RRMs of polypyrimidine-

Figure 3 Overview of the solution structure of the SRp20 RRM in complex with CAUC. (A) Overlay of the final 29 structures superposed on the heavy atoms of the structured parts of the protein and RNA. The protein backbone is shown in cyan and RNA heavy atoms are shown in yellow (C), red (O) and blue (N and P). Only the ordered region of the RRM (residues 9–83) is shown. (B) Surface representation for the RRM (residues 9–83) and stick (RNA heavy atoms) representation for the RNA of the most representative structure of the complex. The protein surface is painted according to surface potential, with red indicating negative and blue indicating positive charges. The RNA is colored as in (A). (C) The most representative structure of the complex in ribbon (protein backbone) and stick (RNA) representation. The color scheme is the same as in (A), and important protein side chains involved in interactions with the RNA are represented as green sticks. Putative hydrogen bonds are shown by dotted violet lines. Schematic representations of the intermolecular hydrogen-bond interactions stabilizing C1 and A2 (D), and U3 (E). The RNA is shown in black and the protein in blue.

Figure 4 Identification of a TAP-binding motif in SRp20 and 9G8. **(A)** Schematic of full-length SRp20, deletion and point mutations used together with the amino-acid sequence for the linker (L) domain. **(B)** Pull-down assays. GST (control, lane –) and GST-TAP-p15 (lanes 1–6) expressed in *E. coli* were first immobilized on glutathione-coated beads. Various ³⁵S-radiolabeled SRp20 proteins synthesized in rabbit reticulocytes were added to the binding reactions in the presence of ribonuclease A. Eluted proteins were analyzed following SDS–PAGE by Phosphorimaging (left, middle panels) and Coomassie blue (right panel). **(C)** Pull-down assays. Lanes 1 and 2 show purified proteins. Recombinant GST (control, lane 3) and GST fusions of SRp20 aa 1–90 (lane 4) or of SRp20 aa 84–90 (lane 5) were immobilized on glutathione beads and purified TAP-p15 expressed in *E. coli* was added to the reactions. Eluted proteins were analyzed by SDS–PAGE stained with Coomassie blue. **(D)** Schematic of full-length 9G8, deletion and point arginine mutations used together with the amino-acid sequence for the linker (L) domain. **(E)** Pull-down assays. GST (control, lane –) and GST-TAP-p15 (lanes 1–8) expressed in *E. coli* were first immobilized on glutathione-coated beads. Various ³⁵S-radiolabeled 9G8 proteins synthesized in rabbit reticulocytes were added to the binding reactions in the presence of ribonuclease. Eluted proteins were analyzed following SDS–PAGE by Phosphorimaging (left, middle panels) and Coomassie blue staining (right panel). **(F)** Pull-down assays. Lanes 1 and 2 show purified proteins. Recombinant GST (control, lane 3) and GST fusions of 9G8 aa 12–98 (lane 4) or of 9G8 aa 81–98 (lane 5) were immobilized on glutathione beads and purified TAP-p15 expressed in *E. coli* was added to the reactions. Eluted proteins were analyzed by SDS–PAGE stained with Coomassie blue. Amino acids encoded by constructs used in all pull-down assays are shown in brackets after the protein name.

tract-binding protein (PTB) in complex with RNA, where decreasing the size of the RNA at the expense of affinity was necessary to solve the structures (Oberstrass *et al*, 2005).

An unusual RNA-binding mode

CAUC binding to SRp20 has similarities to other RRM-RNA complexes and also striking differences. Indeed, C1 is recog-

nized in an almost identical manner in the four RRM of PTB in complex with CUCUCU, except that the hydrogen bond between C1 amino and Glu 79 in SRp20 is replaced by an RNA intramolecular hydrogen bond with the preceding uracil in PTB (Oberstrass *et al.*, 2005). Binding of the 3'AUC is more unusual, as the interaction is not sequence specific and the topology of the bound RNA is unusual. For example, A2

Figure 5 Functional analysis of the TAP-binding motif. **(A)** Co-immunoprecipitation assays. Total extracts from 293T cells (Mock) co-transfected with a Myc-tagged TAP plasmid and either a control (Flag) or various Flag-Myc-tagged SRp20 constructs were treated with ribonuclease A and alkaline phosphatase before immunoprecipitation with α -Flag antibodies. Total extracts (left panel) and purified complexes (right panel) were analyzed by Western immunoblotting with α -Myc antibodies. Asterisks indicate heavy and light IgG chains. **(B)** Schematic representation of the tethered mRNA export reporter assay. SD: splice donor; SA: splice acceptor. **(C)** α -Myc Western blot of total 293T extracts used for the tethered export assay. MS2 fusions of REF2-I-RRM (aa 71–155) and of the SRp20 (aa 84–90)-REF2-I-RRM chimera (chimera) reproducibly led to poor expression (shown with arrows), which can be detected only by overexposure of the blot. **(D)** Luciferase activity generated by the MS2 fusions in the tethered export assay. Error bars represent standard deviations from four independent sets of assays, each carried out in triplicate.

adopts a *syn* conformation, which is unprecedented in RRM–RNA complexes (Maris *et al*, 2005). Similarly, the positioning of U3 and C4 is unusual compared with other RRM–RNA complexes. In the ones solved to date, the residue equivalent to U3 would either stack over the preceding one (A2) or interact with the β 2 strand (Maris *et al*, 2005). Here, it is C4 and not U3 that stacks partially over A2, whereas U3 is flipped out (Figure 3C). Furthermore, neither A2 nor U3 is recognized specifically, whereas C4 may be recognized by Lys 11. Such semi-sequence-specific binding is consistent with the degenerate consensus RNA-binding sequence found previously by *in vitro* and *in vivo* methods for SRp20 (Cavaloc *et al*, 1999; Schaal and Maniatis, 1999). This mode of RNA binding is likely to be conserved in the 9G8 RRM, as binding of the RNA leads to chemical shift changes similar to those seen for the SRp20 RRM (Figure 2). The same should be true for the *Drosophila* RBP1, as all the amino acids whose side chain interacts with the RNA are conserved compared to SRp20 (Figure 1A). However, for 9G8, the specificity of full-length protein is modulated by the zinc-knuckle in SELEX experiments; therefore, 9G8 probably associates with different mRNA sequences *in vivo* (Cavaloc *et al*, 1999).

This semi-specific mode of RNA binding is very unusual compared to other alternative-splicing factors (e.g. Fox-1 (Auweter *et al*, 2006), hnRNPA1 (Ding *et al*, 1999), Sex-lethal (Handa *et al*, 1999) or Nova-1 (Lewis *et al*, 2000)), which

recognize their RNA targets with high sequence specificity. The mode of binding for SRp20 most closely resembles that for PTB, where each of the four RRMs recognizes short pyrimidine tracts (3–5 nt) with weak affinity and in a semi-sequence-specific manner (Oberstrass *et al*, 2005).

Implication for RNA splicing and export of histone mRNAs

The structure of the SRp20 RRM:CAUC complex helps rationalize the basis of the partial sequence selectivity of SRp20 and SR proteins in general (Bourgeois *et al*, 2004). The structure and affinity measurements (Supplementary Figure 3) revealed that C1 in particular is sequence-specifically recognized. This recognition is functionally essential, as a C to G mutation in the first cytosine of the CAUC sequence within the histone mRNA can impair RNA export (Huang and Steitz, 2001). Yet, having degenerate specificity offers multiple functional advantages. More RNA sequences can be targeted and there is less evolutionary pressure on the bound RNA, which is ideal for exonic sequences. Considering the weak binding affinity of SRp20, the protein can associate and dissociate more easily from the RNA, which is an important physical property in the context of the highly dynamic RNA metabolism processes that involve SRp20 (Bourgeois *et al*, 2004). Furthermore, RNA-binding affinity can be modulated by protein–protein interactions that are themselves dependent on the level of phosphoryla-

tion of the interacting proteins. This regulated RNA-binding mode can then be used for fine-tuning post-transcriptional gene expression (Singh and Valcarcel, 2005).

As for the role of SRp20 in mRNA export, our structure of the complex and the identification of the regions in SRp20 and 9G8 interacting with TAP provide the first molecular insight into how these proteins mediate mRNA export. From the SRp20:RNA structure, we see that the region C-terminal of the RRM (after G84 in 9G8 and G83 in SRp20) is not involved in RNA binding and is free to interact with TAP and promote the export of histone and other mRNAs containing SRp20- and 9G8-binding sites. The short TAP-binding motif present in SRp20 is conserved in *Drosophila* RBP1 but shares little in common with 9G8 except that they have multiple arginine residues and they are adjacent to the RRM (Figure 1A). The spacing of the essential arginines does not appear to be vital, as it differs in both proteins, but given that the peptides are flexible, they are likely to be able to fit into a common binding site on TAP.

Earlier studies showed that a fragment from the 9G8 RRM (aa 13–63), lacking the TAP-binding motif, could compete for interaction between 9G8 (aa 1–123) and TAP, implying TAP bound the RRM; however, these studies required a 100 molar excess of peptide for competition *in vitro*, implying weak binding to the RRM (Huang *et al*, 2003). Moreover, the control peptide and proteins used had a mutation in Phe 54, which, as revealed by our structure, would disrupt the hydrophobic core of the RRM and almost certainly disrupt folding, no doubt preventing the weak interactions observed with the wild-type peptide. At physiological salt concentrations, we fail to see any interaction between the 9G8 RRM and TAP (Figure 4B) and see only a very weak interaction in less chaotropic buffers (potassium acetate versus sodium chloride; not shown). However, both SRp20 and 9G8 RRM with the intact C-terminal TAP-binding motif do interact slightly more efficiently with TAP than the isolated C-terminal motif (Figure 4C and F), consistent with the RRM providing a minor contribution to the TAP interaction. Nevertheless, the failure of SRp20 R88E to co-immunoprecipitate with TAP, together with the ability of the isolated C-terminal arginine-rich peptides to bind TAP well, in contrast to the isolated RRM, indicates that these short C-terminal peptides provide the major contribution to the interaction.

It is not clear at present why the many arginines present in the C-terminal RS domains of 9G8 and SRp20 could not also recruit TAP. In this context, weak interactions between TAP and the SR protein RRM might assist in selection of the arginine-rich TAP-binding motif juxtaposed with the RRM. As hyperphosphorylated SR proteins do not bind TAP (Huang *et al*, 2004; Lai and Tarn, 2004), it is tempting to speculate that negatively charged phosphoserines associate with the arginines in the TAP-binding motif to block their interaction with TAP. Altogether, our work provides a detailed role for two domains identified in the adaptor proteins SRp20 and 9G8 in the export of specific RNAs such as histone mRNAs. The role of the RRM is to target specific RNAs and the role of the small peptide C-terminal to the RRM is to help recruit TAP. The two events appear independent, as the RRM can bind RNA on its own and the TAP-binding motif binds TAP independently of RNA. Yet, the small spatial proximity of the two domains is intriguing and could suggest a possible coupling between the two events.

Materials and methods

Cloning, expression and purification of 9G8 and SRp20 RRM domains

The DNA encoding aa 12–98 from human 9G8 was amplified by PCR from cDNA clone MGC22746 and subcloned into a modified pET24b containing a protease cleavage site. The DNA encoding the RRM domain (aa 1–86) of human SRp20 was subcloned into pET30a+ containing the 58-aa GB1 solubility tag (Zhou *et al*, 2001) followed by a 6xHis tag. 9G8 RRM (aa 1–104) was also subcloned fused to GB1 into a pET30a+ vector. 9G8 RRM was expressed in *Escherichia coli* BL21(DE3)RP, GB1-SRp20 RRM was expressed in *E. coli* BL21(DE3) C + RIL and GB1-9G8 RRM was expressed in *E. coli* BL21(DE3) pLysS. For labeled samples, expression was carried out in M9 media containing ^{15}N ammonium chloride and ^{13}C glucose. 9G8 was purified using cobalt affinity chromatography, then dialyzed against 9G8 NMR buffer (20 mM sodium acetate pH 5.5, 50 mM NaCl, 25 mM L-arginine, 50 mM L-glutamic acid, 5 mM EDTA and 10 mM DTT) and subsequently concentrated to 0.6 mM, as determined by Bradford assay. The Arg/Glu addition (Golovanov *et al*, 2004) was essential for concentration of 9G8 to 0.6 mM. GB1-SRp20 RRM was purified by three successive nickel affinity chromatography (Qiagen) steps in the presence of protease inhibitor cocktail (Sigma). The protein was dialyzed against 50 mM sodium dihydrogenophosphate pH 6.4 and 1 mM DTT and concentrated to 1 mM as measured by UV spectroscopy at 274 nm in 6 M guanidium hydrochloride. The absence of RNases was confirmed by the RNase Alert Lab Kit (Ambion). The GB1-9G8 RRM was purified by two steps of cation exchange chromatography, dialyzed against 10 mM sodium dihydrogenophosphate pH 7, 175 mM NaCl and 2 mM β -mercaptoethanol and concentrated to 1 mM.

Complex formation with RNA and NMR spectroscopy

The RNAs used were purchased from Dharmacon, deprotected and desalted on a G15 column and then lyophilized and dissolved in 50 mM Na-phosphate at pH 6.4. RNA concentration was measured by UV spectroscopy at 260 nm. To identify the RNA-binding surface of SRp20 RRM and measure the dissociation constant, a titration of the protein was performed. The titration was carried out in at least three steps for complex formation and in 9–11 steps for affinity measurements (Supplementary Figures 2 and 3). All steps were monitored by ^1H , ^{15}N HSQC spectra at 42°C. Dissociation constant (K_d) could be evaluated by using the program CRVFT (R Boyko and BD Sykes, University of Alberta, Edmonton, Canada). For all RNAs tested with 5 nt or more (except UUCAC), saturation was reached at a 1:1 molar ratio of protein:RNA, whereas with short RNAs, saturation was reached only with an excess of RNA. The titrations indicate that the protein is in fast to intermediate exchange relative to the NMR time-scale. The SRp20RRM-CAUC complex for NMR measurement was prepared at a 1:1 stoichiometric ratio of protein:RNA and studied at a concentration of 1 mM in 50 mM sodium dihydrogenophosphate pH 6.4. Studies with the GB1-9G8 construct in complex with RNA were carried out under the same condition as with the GB1-SRp20 construct.

NMR spectroscopy and resonance assignments of free SRp20 and in complex with CAUC

Owing to improvement of spectral linewidths at higher temperature (Supplementary Figure 4), NMR spectra of the free SRp20 RRM and of the CAUC complex were recorded at 42 and 46°C on Bruker DRX 500 (with a cryoprobe), DRX 600 or Avance 900 MHz spectrometers. All the spectra of the free GB1-SRp20 RRM were recorded on a 0.6 mM sample containing a mutation (E65D) that does not affect the structure of the protein.

The backbone resonance assignments of free and bound SRp20 RRM were achieved using several 3D triple resonance experiments (Bax and Grzesiek, 1993)—HNCA, HN(CO)CA, CBCA(CO)NH— together with a 2D ^1H , ^{15}N HSQC and a 3D ^1H , ^{15}N , ^1H NOESY. Resonances of the aliphatic side chains were initially assigned using 3D ^1H , ^{13}C , ^1H NOESY and 3D ^1H , ^{15}N , ^1H NOESY spectra. Each NOESY-based side-chain resonance assignment was subsequently confirmed using 3D HCCH TOCSY. Resonances of aromatic side chains of phenylalanine and histidine residues were assigned using 2D homonuclear TOCSY, 2D homonuclear NOESY, 3D ^1H , ^{13}C , ^1H NOESY and 2D ^1H , ^{13}C HSQC spectra recorded in D_2O .

The bound RNA resonances were initially assigned using a combination of through-space and through-bond NMR experiments

(Varani *et al*, 1996) and later confirmed with ^{13}C -labeled RNA with a labeled sugar that was made by chemical synthesis (Wenter *et al*, 2006). The sugar ring conformation of all four RNA nucleotides is C2' -endo, as identified in the 2D homonuclear TOCSY spectra by the presence of strong cross-peaks between the H1' and H2' resonances and weak cross-peaks between the H1' and H3' resonances. All nucleotides of the bound CAUC have their bases in the *anti*-conformation, except for A2, as indicated by a very strong NOE correlation between A2 H8 and its own H1' and a weaker NOE correlation between A2 H8 and other protons in the 2D homonuclear NOESY spectrum.

Structure determination for the 9G8 RRM

NMR spectra were recorded at 20°C on a Bruker DRX 600 MHz spectrometer equipped with a CryoProbe. Data were processed using NMRPipe and NMRDraw (Delaglio *et al*, 1995) and analyzed using NMRVIEW (Johnson and Blevins, 1994). Sequence-specific backbone and side-chain resonance assignment of 9G8 was achieved using HNCA, HN(CO)CA, HNC(O), HN(CA)CO, CBCA(CO)NH, CBCANH, HBHA(CO)NH and HCCH-TOCSY experiments. Resonances of aromatic side chains were assigned using 2D [^1H , ^{13}C] HSQC and homonuclear TOCSY and NOESY spectra recorded in D_2O .

Secondary structural elements and β -strand topology were determined from a combination of H_α , C_α , $\text{C}\beta$ and C' secondary chemical shift analysis (Wishart and Sykes, 1994) and identification of unambiguous NOE contacts. Structure calculations were performed with the program CYANA 2.1 (Guntert *et al*, 1997), using as input data partially manually assigned list of NOESY cross-peaks and additional restraints for 34 hydrogen bonds, and 119 ϕ and ψ backbone torsion angles derived from TALOS (Table I). The standard protocol was used with seven cycles of combined automated NOE assignment and structure calculation of 100 conformers in each cycle, out of which the 20 with lowest target function values were retained for analysis. For each conformer, the standards simulated annealing schedule with 10^3 torsion angle dynamics steps was applied, starting from initial structures with random values of the torsion angles. The best 20 structures obtained from CYANA calculations were further refined in AMBER 7.0 using a total of 951 unambiguous interproton distance restraints. A final family of 20 9G8 structures was selected on the basis of low restraint and AMBER energies, and was characterized with PROCHECK-NMR (Laskowski *et al*, 1993). Atomic coordinates and NMR restraints of the 9G8 RRM free have been deposited in the Protein Data Bank under the accession code 2HVZ.

Structure determination of the RRM of SRp20 free and in complex with CAUC

For intermolecular NOE assignments, a 2D homonuclear NOESY, a 3D filtered [^1H , ^{13}C , ^1H] NOESY (Zwahlen *et al*, 1997) and a 2D filtered/edited NOESY (Peterson *et al*, 2004) NMR experiment recorded in D_2O and a 2D homonuclear NOESY experiment recorded in H_2O were used. In total, 40 intermolecular NOEs between SRp20 RRM and CAUC could be assigned. Preliminary structure determinations of the free protein and of the protein in complex were performed with the automated NOE assignments module CANDID (Herrmann *et al*, 2002) in the DYANA program (Guntert *et al*, 1997). Before this, NOE cross-peaks in the 2D homonuclear NOESY, 3D [^1H , ^{15}N , ^1H] NOESY and 3D [^1H , ^{13}C , ^1H] NOESY spectra of the protein were automatically picked using the peak picking routine in the Sparky program and subsequently manually inspected. In addition, weak peaks and all NOE peaks in the 2D homonuclear NOESY were picked manually. For the complex, intermolecular protein-RNA NOEs were manually assigned and distance restraints were estimated using the simple isolated spin pair approximation approach (Wuthrich, 1986). Using this approach, the average intensity of all cross-peaks corresponding to H5-H6 NOEs of pyrimidines was set to 2.4 Å.

All distance restraint lists were used in DYANA to calculate preliminary structures of SRp20, both free and in complex with CAUC. The preliminary structure of the complex helped to identify new intermolecular NOEs. In DYANA calculations, 200 random starting conformers and 30000 steps of simulated annealing were used. The 50 best structures were selected based on the lowest target function and used for further refinement in AMBER 7.0 (Case *et al*, 2002). In all AMBER calculations, AMBER force field (Cornell *et al*, 1995) and the generalized Born solvation model (Bashford and Case, 2000), to mimic the shielding effect of the solvent, were

used. A 10 ps simulated annealing protocol was used for the calculation following the same protocol as previously described for complexes from the group (Oberstrass *et al*, 2006). The simulated annealing protocol was followed by a short energy minimization of 500 cycles. The 19 and 29 conformers with the lowest AMBER energy were selected to form the final ensemble of free SRp20 RRM and that bound to CAUC, respectively. Molecular graphics were generated using MOLMOL (Koradi *et al*, 1996). All the intermolecular hydrogen bonds indicated in Figure 3C-E were derived from the structure ensemble. Only one hydrogen bond between C1 N3 and N82 HN was constrained in the last calculation to improve convergence, as it is additionally supported by the large downfield chemical shift of N82 amide observed upon RNA binding (Figure 2B). Atomic coordinates and NMR restraints of the GB1-SRp20 RRM free and bound have been deposited in the Protein Data Bank under the accession codes 2I38 and 2I2Y, respectively.

Analysis of SR protein interactions with TAP

The GST-TAP expression vector was described previously (Williams *et al*, 2005). p15 was subcloned into pET9a (Novagen) expression vector and coexpressed with GST-TAP in *E. coli*. GST pull-downs were carried out as described previously in PBS + 0.1% Tween 20 (Williams *et al*, 2005). SR protein domains were expressed in reticulocyte lysates or in 293T cells using PCR-amplified domains subcloned respectively into pET24b (Novagen) or p3X-Flag-myc-CMV-26 (Sigma) plasmids. Quikchange mutagenesis was used to create site-directed mutations. For co-immunoprecipitation experiments, 293T cells were transfected with either Flag control or Flag fusions of SRp20 domains and mutants together with a 13-Myc-TAP expression vector. Three wells of a 24-well plate were used for each condition and cells were lysed in 300 μl of 50 mM Hepes pH 7.5, 150 mM NaCl, 0.5% Triton, 1 mM EDTA, 1 mM DTT, 10% glycerol containing 2 mM PMSF and complete protease inhibitors (Roche). Total extracts were treated with RNase and alkaline phosphatase (CIAP, Roche), subjected to immunoprecipitation using Flag-agarose beads (Sigma) and purified proteins were eluted in the presence of 0.25 mg/ml Flag peptide (Sigma). The resulting proteins were resolved on 12% SDS-PAGE and analyzed by Western blotting. The MS2 reporter plasmid was constructed by PCR amplification of tandem MS2 sites from pIII/MS2-2 (SenGupta *et al*, 1996) on an *XhoI* fragment and subcloning into the *SalI* site of pLUCSALRRE(Williams *et al*, 2005). A clone (pLUCSALRRE6MS2) that had three copies of the tandem MS2 sites in the same orientation was selected. The MS2 open reading frame was PCR amplified with oligonucleotides, introducing an N-terminal Myc tag and subcloned into the *NheI*-*XbaI* sites of pCINEO (Promega) to generate pCINEOMycMS2. TAP, GFP, REF and truncations were PCR amplified and subcloned into the *XbaI*-*NotI* sites of pCINEO-MycMS2. The SRp20-REFRRM chimera was created by the insertion of annealed oligonucleotides bearing the coding sequence for aa 84-90 of SRp20 with a six glycine hinge into the *XbaI* site of pCINEOMycMS2-REF RRM. A 100 ng portion of pLUCSALRRE6MS2 was cotransfected into 293T cells with 600 ng of MS2 fusion expression vectors and 50 ng of β -galactosidase expression vector. Transfections were carried out in triplicate and luciferase and β -galactosidase assays were measured in duplicate, 36 h post-transfection. The luciferase activities were averaged and normalized to the β -galactosidase levels, which acted as a transfection control.

Supplementary data

Supplementary data are available at *The EMBO Journal* Online (<http://www.embojournal.org>).

Acknowledgements

We thank Vicky Porteous for technical assistance, Professor Stefan Pitsch, Philip Wenter and Luc Reymond (EPFL, Lausanne) for providing us with ^{13}C sugar-labeled CAUC RNA, Richard Steff for help in the structure calculation and Sigrid Auweter for help in the analysis of titration experiments. This investigation was supported by the Roche Research Fund for Biology at the ETH Zurich, by the Swiss National Science Foundation, Structural Biology National Center of Competence in Research, by a grant from the ETH Zürich (TH- Fonds Nr 0-20960-01) to FHTA and by grants from the BBSRC (UK) to SAW, LYL and APG. The Wellcome Trust is acknowledged for funding the Bruker Avance 600 MHz spectrometer in Manchester. FHTA is an EMBO Young Investigator.

References

- Auweter SD, Fasan R, Reymond L, Underwood JG, Black DL, Pitsch S, Allain FH (2006) Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. *EMBO J* **25**: 163–173
- Bashford D, Case DA (2000) Generalized born models of macromolecular solvation effects. *Annu Rev Phys Chem* **51**: 129–152
- Bax A, Grzesiek S (1993) Methodological advances in protein NMR. *Acc Chem Res* **26**: 131–138
- Bourgeois CF, Lejeune F, Stevenin J (2004) Broad specificity of SR (serine/arginine) proteins in the regulation of alternative splicing of pre-messenger RNA. *Prog Nucleic Acid Res Mol Biol* **78**: 37–88
- Case DA, Pearlman DA, Caldwell JW, Cheatham III TE, Wang J, Ross WS, Simmerling CL, Darden TA, Merz KM, Stanton RV, Cheng AL, Vincent JJ, Crowley M, Tsui V, Gohlke H, Radmer RJ, Duan Y, Pitera J, Massova I, Seibel GL, Singh UC, Weiner PK, Kollman PA (2002) AMBER 7. University of California, San Francisco
- Cavaloc Y, Bourgeois CF, Kister L, Stevenin J (1999) The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. *RNA* **5**: 468–483
- Cavaloc Y, Popielarz M, Fuchs JP, Gattoni R, Stevenin J (1994) Characterization and cloning of the human splicing factor 9G8: a novel 35 kDa factor of the serine/arginine protein family. *EMBO J* **13**: 2639–2649
- Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A 2nd generation force-field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* **117**: 5179–5197
- Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* **6**: 277–293
- Ding J, Hayashi MK, Zhang Y, Manche L, Krainer AR, Xu R-M (1999) Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA. *Genes Dev* **13**: 1102–1115
- Fairbrother WG, Yeh RF, Sharp PA, Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. *Science* **297**: 1007–1013
- Fu XD (1995) The superfamily of arginine/serine-rich splicing factors. *RNA* **1**: 663–680
- Galiana-Arnoux D, Lejeune F, Gesnel M-C, Stevenin J, Breathnach R, Del Gatto-Konczak F (2003) The CD44 alternative v9 exon contains a splicing enhancer responsive to the SR proteins 9G8, ASF/SF2, and SRp20. *J Biol Chem* **278**: 32943–32953
- Golovanov AP, Hautbergue GM, Wilson SA, Lian LY (2004) A simple method for improving protein solubility and long-term stability. *J Am Chem Soc* **126**: 8933–8939
- Guntert P, Mumenthaler C, Wuthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J Mol Biol* **273**: 283–298
- Handa N, Nureki O, Kurimoto K, Kim I, Sakamoto H, Shimura Y, Muto Y, Yokoyama S (1999) Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein. *Nature* **398**: 579–585
- Heinrichs V, Baker BS (1995) The *Drosophila* SR protein RBP1 contributes to the regulation of doublesex alternative splicing by recognizing RBP1 RNA target sequences. *EMBO J* **14**: 3987–4000
- Herrmann T, Guntert P, Wuthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J Mol Biol* **319**: 209–227
- Huang Y, Gattoni R, Stevenin J, Steitz JA (2003) SR splicing factors serve as adapter proteins for TAP-dependent mRNA export. *Mol Cell* **11**: 837–843
- Huang Y, Steitz JA (2001) Splicing factors SRp20 and 9G8 promote the nucleocytoplasmic export of mRNA. *Mol Cell* **7**: 899–905
- Huang Y, Yario TA, Steitz JA (2004) A molecular link between SR protein dephosphorylation and mRNA export. *Proc Natl Acad Sci USA* **101**: 9666–9670
- Johnson BA, Blevins RA (1994) NMR view: a computer program for the visualization and analysis of NMR data. *J Biomol NMR* **4**: 603–614
- Jumaa H, Nielsen PJ (2000) Regulation of SRp20 exon 4 splicing. *Biochim Biophys Acta* **1494**: 137–143
- Koradi R, Billeter M, Wuthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graphics* **14**: 51–55, 29–32
- Lai MC, Tarn WY (2004) Hypophosphorylated ASF/SF2 binds TAP and is present in messenger ribonucleoproteins. *J Biol Chem* **279**: 31745–31749
- Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* **26**: 283–291
- Lewis HA, Musunuru K, Jensen KB, Edo C, Chen H, Darnell RB, Burley SK (2000) Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. *Cell* **100**: 323–332
- Lou H, Neugebauer KM, Gagel RF, Berget SM (1998) Regulation of alternative polyadenylation by U1 snRNPs and SRp20. *Mol Cell Biol* **18**: 4977–4985
- Lynch KW, Maniatis T (1996) Assembly of specific SR protein complexes on distinct regulatory elements of the *Drosophila* doublesex splicing enhancer. *Genes Dev* **10**: 2089–2101
- Maris C, Dominguez C, Allain FHT (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J* **272**: 2118–2131
- Oberstrass FC, Auweter SD, Erat M, Hargous Y, Henning A, Wenter P, Reymond L, Amir-Ahmady B, Pitsch S, Black DL, Allain FH (2005) Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science* **309**: 2054–2057
- Oberstrass FC, Lee A, Stefl R, Janis M, Chanfreau G, Allain FH (2006) Shape-specific recognition in the structure of the Vts1p SAM domain with RNA. *Nat Struct Mol Biol* **13**: 160–167
- Peterson RD, Theimer CA, Wu H, Feigon J (2004) New applications of 2D filtered/edited NOESY for assignment and structure elucidation of RNA and RNA-protein complexes. *J Biomol NMR* **28**: 59–67
- Rodrigues JP, Rode M, Gatfield D, Blencowe BJ, Carmo-Fonseca M, Izaurralde E (2001) REF proteins mediate the export of spliced and unspliced mRNAs from the nucleus. *Proc Natl Acad Sci USA* **98**: 1030–1035
- Sanford JR, Gray NK, Beckmann K, Caceres JF (2004) A novel role for shuttling SR proteins in mRNA translation. *Genes Dev* **18**: 755–768
- Schaal TD, Maniatis T (1999) Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. *Mol Cell Biol* **19**: 1705–1719
- SenGupta DJ, Zhang B, Kraemer B, Pochart P, Fields S, Wickens M (1996) A three-hybrid system to detect RNA-protein interactions *in vivo*. *Proc Natl Acad Sci USA* **93**: 8496–8501
- Singh R, Valcarcel J (2005) Building specificity with nonspecific RNA-binding proteins. *Nat Struct Mol Biol* **12**: 645–653
- Tacke R, Manley JL (1999) Determinants of SR protein specificity. *Curr Opin Cell Biol* **11**: 358–362
- Varani G, Aboul-ela F, Allain FHT (1996) NMR investigation of RNA structure. *Prog Nucl Magn Reson Spectrosc* **29**: 51–127
- Wenter P, Reymond L, Auweter S, Allain FH-T, Pitsch S (2006) Short, synthetic and selectively ¹³C-labeled RNA sequences for the NMR structure determination of protein-RNA complexes. *Nucleic Acids Res* **34**: e79
- Wiegand HL, Lu S, Cullen BR (2003) Exon junction complexes mediate the enhancing effect of splicing on mRNA expression. *Proc Natl Acad Sci USA* **100**: 11327–11332
- Williams BJ, Boyne JR, Goodwin DJ, Roaden L, Hautbergue GM, Wilson SA, Whitehouse A (2005) The prototype gamma-2 herpesvirus nucleocytoplasmic shuttling protein, ORF 57, transports viral RNA through the cellular mRNA export pathway. *Biochem J* **387**: 295–308
- Wishart DS, Sykes BD (1994) The ¹³C chemical-shift index: a simple method for the identification of protein secondary structure using ¹³C chemical-shift data. *J Biomol NMR* **4**: 171–180
- Wuthrich K (1986) *NMR of Proteins and Nucleic Acids*. New York: Wiley
- Zahler AM, Lane WS, Stolk JA, Roth MB (1992) SR proteins: a conserved family of pre-mRNA splicing factors. *Genes Dev* **6**: 837–847
- Zhang Z, Krainer AR (2004) Involvement of SR proteins in mRNA surveillance. *Mol Cell* **16**: 597–607
- Zhou P, Lugovskoy AA, Wagner G (2001) A solubility-enhancement tag (SET) for NMR studies of poorly behaving proteins. *J Biomol NMR* **20**: 11–14
- Zwahlen C, Legault P, Vincent SJF, Greenblatt J, Konrat R, Kay LE (1997) Methods for measurement of intermolecular NOEs by multinuclear NMR spectroscopy: application to a bacteriophage lambda N-peptide/boxB RNA complex. *J Am Chem Soc* **119**: 6711–6721