A *syn–anti* conformational difference allows SRSF2 to recognize guanines and cytosines equally well

Gerrit M Daubner¹, Antoine Cléry¹,
Sandrine Jayne¹, James Stevenin²
and Frédéric H-T Allain^{1,*}

¹Institute of Molecular Biology and Biophysics, ETH Zürich, Zürich, Switzerland and ²Department of Functional Genomics, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), University of Strasbourg, Strasbourg, France

SRSF2 (SC35) is a key player in the regulation of alternative splicing events and binds degenerated RNA sequences with similar affinity in nanomolar range. We have determined the solution structure of the SRSF2 RRM bound to the 5'-UCCAGU-3' and 5'-UGGAGU-3' RNA, both identified as SRSF2 binding sites in the HIV-1 tat exon 2. RNA recognition is achieved through a novel sandwich-like structure with both termini forming a positively charged cavity to accommodate the four central nucleotides. To bind both RNA sequences equally well, SRSF2 forms a nearly identical network of intermolecular interactions by simply flipping the bases of the two consecutive C or G nucleotides into either *anti* or *syn* conformation. We validate this unusual mode of RNA recognition functionally by *in-vitro* and *in-vivo* splicing assays and propose a 5'-SSNG-3' (S = C/G) high-affinity binding consensus sequence for SRSF2. In conclusion, in addition to describe for the first time the RNA recognition mode of SRSF2, we provide the precise consensus sequence to identify new putative binding sites for this splicing factor.

The EMBO Journal (2012) 31, 162–174. doi:10.1038/emboj.2011.367; Published online 14 October 2011

Subject Categories: RNA; structural biology

Keywords: alternative splicing; HIV tat exon; NMR; protein–RNA complex; SR protein

Introduction

Alternative splicing is a fundamental biological process in eukaryotic cells, which is important for proteome variety. It is estimated that 95% of the human multi-exon genes undergo alternative splicing (Black, 2003; Wahl *et al.*, 2009). Splicing cannot be seen as an independent cellular process. Factors involved in the regulation of alternative splicing are intimately coupled with transcription (de la Mata and Kornblihtt, 2006; Das *et al.*, 2007; Oesterreich *et al.*, 2011)

*Corresponding author. Institute for Molecular Biology and Biophysics, ETH Zürich, HPG G18, Zürich CH 8093, Switzerland.
Tel.: +41 44 633 3940; Fax: +41 44 633 1294;
E-mail: allain@mol.biol.ethz.ch

Received: 15 June 2011; accepted: 12 September 2011; published online: 14 October 2011

and involved in many downstream events reaching from mRNA export (Huang and Steitz, 2001) to translation (Sanford *et al.*, 2004). It is, therefore, not surprising to find many links between this central process and widespread diseases like cancer or AIDS (Wang and Cooper, 2007; Grosso *et al.*, 2008; Saliou *et al.*, 2009). Consequently, current efforts aim to assemble a splicing code including all factors involved in the regulation of alternative splicing and allowing the prediction of splicing patterns of any primary transcript from its sequence (Wang and Burge, 2008; Barash *et al.*, 2010).

In general, regulation of alternative splicing is achieved through *trans*-acting protein factors binding to *cis*-regulatory RNA elements. Members of the heterogeneous nuclear ribonucleoprotein (hnRNP; Dreyfuss *et al.*, 1993; Venables *et al.*, 2008) and the serine/arginine-rich (SR) protein family are among the most abundant splicing regulators (Bourgeois *et al.*, 2004; Long and Cáceres, 2009). Both families usually act antagonistically, with hnRNPs as silencers and SR proteins as enhancers of splicing events. SRSF2 (also referred previously as SC35) is a member of the SR protein family and shares its prototypical modular structure with an N-terminal RNA recognition domain (RRM) and a C-terminal RS domain. While the RRM domain is mainly involved in specific RNA recognition, the RS domain modulates SRSF2 activity through phosphorylation (Colwill *et al.*, 1996; Rossi *et al.*, 1996; Wang *et al.*, 1998), directly contacts RNA (Shen *et al.*, 2004; Hertel and Graveley, 2005) and promotes protein–protein interactions with the spliceosome (Wu and Maniatis, 1993; Kohtz *et al.*, 1994). SRSF2 is the only family member which is exclusively localized in the nucleus, and is therefore restricted to nuclear processes (Cazalla *et al.*, 2002). Besides its crucial function as a regulator of alternative splicing, SRSF2 was shown to be actively involved in transcription elongation by directly or indirectly mediating the recruitment of elongation factors to the Polymerase II C-terminal domain (Lin *et al.*, 2008).

To regulate alternative splicing events, the RRM of SRSF2 specifically binds *cis*-regulatory elements on the pre-mRNA. Although there are reports on SRSF2 acting on intronic splicing enhancers (Gabut *et al.*, 2005), most studies reported about SRSF2 binding to exonic splicing enhancers (ESEs). A well-studied example is the competition between SRSF2 and hnRNP A1 for binding to the tat exons 2 or 3 of the HIV-1 tat pre-mRNA. Cooperative binding of hnRNP A1 to an exonic splicing silencer inhibits splicing by blocking the binding of SRSF2 to an overlapping ESE. This competition seems to depend directly on the concentration of each antagonist and on their affinity for the targeted RNA sequence (Zhu *et al.*, 2001; Zahler *et al.*, 2004; Hallay *et al.*, 2006; Okunola and Krainer, 2009). One single mutation introduced into a *cis*-acting element bound by SRSF2 can strongly affect binding affinity, alter the splicing pattern and ultimately lead to severe diseases like human growth hormone deficiency (Solis

et al, 2008a,b). Therefore, many efforts have been invested into the search of a consensus motif for *trans*-acting factors to predict critical RNA binding sequences (Cartegni *et al*, 2003; Smith *et al*, 2006). But while *in-vitro* SELEX (systematic evolution of ligands by exponential enrichment) could identify closely related high-affinity consensus motifs for SR proteins SRSF1, SRSF3 and SRSF7, very degenerated consensus sequences were selected in the presence of SRSF2 (Tacke and Manley, 1995; Cavaloc *et al*, 1999). In addition, functional SELEX (Liu *et al*, 2000) and its two-step variant (Schaal and Maniatis, 1999b) resulted also in the not overlapping consensus sequences 5'-GRYYcSYR-3' (R = purine, Y = pyrimidine, c = tendency to C, S = G/C) and 5'-UGCNGYY-3' (Bourgeois *et al*, 2004). This hints either to a huge variability in the SRSF2 binding specificity or to an unspecific binding mode for RNA recognition.

In this study, we reveal the molecular basis of RNA recognition by SRSF2. We solved the solution structure of SRSF2 in the free state and in complex with a pyrimidine- and a more purine-rich RNA sequence, both found in the HIV-1 tat exon 2 (Zahler *et al*, 2004; Hallay *et al*, 2006). Based on these results and in combination with exhaustive isothermal titration calorimetry (ITC) measurements, we are able to present a consensus motif for high-affinity binding by SR protein SRSF2. We further show that this consensus motif is valid in a functional context and that we can affect splicing activity *in vivo* with specific point mutation in the RNA-binding interface of SRSF2. Altogether, our data will enhance the prediction and identification of new ESE sequences for SRSF2 and will provide a further piece of the Rosetta stone to decipher the splicing code.

Results

Solution structure of the SRSF2 RRM domain

SRSF2 among SR proteins is unusual for its ability to bind with similar affinity in nanomolar range two very different types of RNA sequences as identified by SELEX. For example, a pyrimidine-rich sequence like 5'-UGUCCAGUU-3' or a purine-rich sequence like 5'-AGGAGAU-3' (Cavaloc *et al*, 1999). This unusual sequence specificity motivated us to characterize its mode of RNA recognition by NMR spectroscopy. We focused our study on the first 101 amino acids that embed an RNA recognition motif (RRM; Figure 1A), known to be responsible for the sequence-specific RNA recognition of SRSF2 (Cavaloc *et al*, 1999). The recombinant protein was purified by affinity chromatography using an N-terminal HIS tag. This protein fragment is most stable and soluble in a buffer containing 50 mM arginine, 50 mM glutamate and 20 mM NaH₂PO₄ at pH 5.5 (Hautbergue and Golovanov, 2008). After collecting 2585 NOE-derived distance restraints, we obtained a precise ensemble of 20 structures for the SRSF2 RRM (r.m.s. deviation of 1.3 Å for all heavy atoms; Figure 1B; Table I).

The segment spanning residues 16–90 adopts a canonical RRM fold ($\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$). Although the RRM exposes a lysine (Lys17) instead of the classical aromatic residue within the RNP2 motif on β -strand 1, this is somehow compensated by the presence of an exposed aromatic residue on strand β_2 (Tyr44) (Figure 1C). In total, three aromatics (Tyr44, Phe57 and Phe59) are exposed on the β -sheet surface. A characteristic feature of SRSF2 is the presence of a long loop 3 between

β -strands 2 and 3, which spans 13 amino acids (Bourgeois *et al*, 2004). More unusual, the region immediately before β -strand 1 (aa 10–15) is structured and well defined. Several main-chain hydrogen bonds and a small hydrophobic patch involving Val10, Met13 and Met89 stabilize this structural feature (Figure 1D). The remaining parts of the SRSF2 N- and C-termini (aa 1–8 and 94–101, respectively) are flexible as confirmed by relaxation data (¹H-¹⁵N NOE, Supplementary Figure S1A).

Overview of SRSF2 RRM in complex with 5'-UCCAGU-3' RNA

We next aimed at elucidating the RNA binding mode of SRSF2 by solving its structure bound to RNA. We decided to use the sequence 5'-UCCAGU-3' for our NMR studies, since this sequence is found both in the ESE of HIV-1 tat exon 2 (Zahler *et al*, 2004) and in the SELEX-derived pyrimidine-rich sequence 5'-UGUCCAGUU-3'. NMR titration with increasing amounts of this single-stranded hexanucleotide induced large chemical shift perturbations upon binding to SRSF2 RRM and saturation was reached at a ratio of 1:1 (Figure 2A). The largest chemical shift perturbations are observed on β -strands 1–3, which is the canonical RRM binding site, and for residues of loop 3 and of the regions just N- and C-terminal to the RRM (Figure 2B). A dissociation constant (K_d) of 0.27 μ M was measured by ITC in our NMR conditions (Figure 2C). Since the protein contains an N-terminal HIS tag, we verified that the tag did not perturb RNA interaction (Supplementary Figure S1B). Considering the good quality of the spectra, we solved the solution structure of SRSF2 RRM in complex with 5'-UCCAGU-3' using 2113 NOE-based constraints (including 107 intermolecular ones; Supplementary Table S1). The final ensemble of 20 conformers presents a precise structure with an r.m.s. deviation of 1.11 Å for all heavy atoms (Figure 2D; Table I).

The six nucleotides of the RNA are stretched over the β -sheet surface, with all the sugar puckers adopting a C2'-endo conformation and the bases adopting an *anti* conformation (Figure 2E). Only three of the six nucleotides, namely C2, C3 and G5 are sequence specifically recognized by the RRM. The two flanking uracils and A4 do not show any sequence-specific intra-RNA or intermolecular contacts. A particularity of this complex is the involvement of residues of the loop 3 and both termini in RNA binding. Parts of the N- (aa 3–8) and C-termini (aa 94–96) are more ordered in the RNA bound state when compared with the free form. This is confirmed by relaxation studies, although unambiguous evidence is hindered by five prolines in this region (Supplementary Figure S2). Pro6 and especially Pro95 form extensive contacts with the RNA through stacking as experimentally shown by intermolecular NOEs (Supplementary Table S1). In addition, the region N-terminal to the RRM shows large chemical shift perturbations between the free and the RNA-bound form due to the major change of conformation in this region (Figures 1B and 2D). The intramolecular interactions formed in the free form are broken upon RNA binding resulting in large chemical shift perturbations for the amides of Thr14 and Ser15 and the carbonyl of Tyr92 (Supplementary Figure S3). While the amide of Thr14 forms a new hydrogen bond with the His63 carbonyl at the end of β -strand 3 and the Tyr92 carbonyl interacts with the amino group of C3, the amide of Ser15 is not hydrogen bonded

Figure 1 Solution structure of the SRSF2 RRM (aa 1–101) in the free state. **(A)** Schematic representation of the full-length SRSF2. The protein is shown with the amino-acid sequence of the RRM used in these studies. The β -strands are coloured in orange, α -helices in red and both conserved RNP motifs are underlined. Numbering is according to the PDB sequence. **(B)** Backbone traces (N, C α and C') of the 20 lowest energy structures of the free-state SRSF2 RRM (aa 1–101) superimposed on the backbone of the structured part (aa 10–93). Protein backbone of the RRM (aa 15–90) in grey, N-terminus (aa 1–14) in blue and C-terminus (aa 91–101) in dark green. **(C)** Ribbon structure of the free-state SRSF2 RRM. Characteristic residues are in stick representation, coloured green for carbon, red for oxygen and blue for nitrogen. **(D)** Close-up view on the hinge between β -strand 1 and the N-terminus with the same colour code than for **(C)** plus yellow for sulphur. Hydrogen bonds are presented as violet dashed lines. Figures **(B–D)** were generated with MOLMOL (Koradi *et al*, 1996).

anymore. Overall, this leads to a significant conformational relocation of both termini to form a hydrophobic and positively charged cavity to accommodate the single-stranded RNA (Supplementary Figure S4).

Sequence-specific recognition of a 5'-CCNG-3' sequence

Three of the six nucleotides are sequence specifically recognized by SRSF2 RRM (Figure 3A) and we evaluated the conservation of all hydrogen bonds within the 20 structures with the program HBAT (Supplementary Table S2; Supplementary Figure S5; Tiwari and Panigrahi, 2007). The first base to be specifically recognized is C2 through two hydrogen bonds to the protein main chain at the C-terminal end of β -strand 4 (between its N4 amino group and the main-chain carbonyl of Met89 and between N3 and the main-chain amide of Arg91; Figure 3A). In support of these contacts, large chemical shift perturbations of those two backbone groups are observed upon RNA binding (Figure 2B; Supplementary Figure S3; Dominguez *et al*, 2011). It is a conserved mode of cytosine recognition within RRM, although in the case of SRSF2 no aromatic is present on β -strand 1. Instead, the hydrophobic part of Lys17 resides beneath the base of C2 and its amino group contacts the C2 phosphate oxygen. In addition, the base of C2 interacts also

with the side chains of Arg91 and Tyr92 from the region C-terminal to the RRM.

C3 is located in the middle of the RRM on β -strand 3 and stacks with its base on Phe59 of the RNP1 motif and further with the side chain of Arg94 from the region C-terminal to the RRM (Figure 3A). The side chain of Arg61 provides binding specificity by forming hydrogen bonds to N3 and O2. The N4 amino group of C3 is hydrogen bonded to the main-chain carbonyl of Tyr92, the same that was in the free protein hydrogen bonded to Thr14 and Ser15 amides (Figure 2B).

The base of A4 is sandwiched between residue Ser98 from the C-terminus and Pro46, Arg49 and Tyr50 from loop 3. Our structure shows many contacts to the sugar and phosphate backbone, but no sequence-specific contacts to the A4 base.

Among the four central nucleotides, G5 is most obvious recognized sequence specifically (Figure 3B). G5 forms with its Watson–Crick edge two hydrogen bonds to the carboxyl groups of the Asp42 side chain and one to the main-chain carbonyl of Gly4 from the region N-terminal to the RRM. Furthermore, the base moiety of G5 is sandwiched between Tyr44 from β -strand 2 and Pro6 from the N-terminal region.

To evaluate the importance of the protein–RNA interactions identified in the structure, we substituted most interacting residues by alanine and measured the effect of each

Table I Structural statistics of the SRSF2 RRM in complex with 5'-UCCAGU-3' and 5'-UGGAGU-3' RNA

	SRSF2	SRSF2 + 5'-UCCAGU-3'		SRSF2 + 5'-UGGAGU-3'	
	Protein	Protein	RNA	Protein	RNA
<i>NMR distance and dihedral constraints</i>					
Distance constraints					
Total NOE	2585	1933	73	2098	83
Intra-residue	458	411	65	420	66
Inter-residue					
Sequential ($ i-j = 1$)	736	560	8	609	17
Non-sequential ($ i-j > 1$)	1391	962	0	1069	0
Hydrogen bonds	29	29	3	29	3
Protein–RNA intermolecular		107		83	
Total dihedral angle restraints		6		6	
<i>Structure statistics</i>					
Violations (mean and s.d.)					
Distance constraints (>0.3 Å)	0.33 ± 0.02	0.32 ± 0.04		0.3 ± 0.11	
Dihedral angle constraints (deg)	0	0		0	
Max. dihedral angle violation (deg)	0	0		0	
Max. distance constraint violation (Å)	0.35 ± 0.01	0.32 ± 0.04		0.3 ± 0.11	
<i>Deviations from idealized geometry</i>					
Bond lengths (Å)	0.0041	0.0041		0.0042	
Bond angles (deg)	1.7	1.67		1.77	
<i>Average pairwise r.m.s. deviation^a (Å)</i>					
Protein (10–93)					
Heavy	1.3 ± 0.19	1.08 ± 0.19		1.46 ± 0.18	
Backbone	0.54 ± 0.16	0.63 ± 0.17		0.78 ± 0.17	
RNA (103–106)					
All RNA heavy		1.11 ± 0.45		1.3 ± 0.37	
Complex (10–93/103–106)					
Protein and RNA heavy		1.11 ± 0.21		1.47 ± 0.18	
<i>Ramachandran statistics</i>					
Most favoured regions	63.8%	63.8%		63.3%	
Additional allowed regions	31.3%	33.2%		33.0%	
Generously allowed regions	3.1%	2.3%		3.3%	
Disallowed regions	1.7%	0.7%		0.4%	

^aCalculated for the ensemble of the 20 lowest energy structures.

mutation on binding affinity by ITC measurements (Table II; Supplementary Figure S6). We also recorded ¹H-¹⁵N-HSQC of each SRSF2 mutant to confirm that the RRM fold was not affected (Supplementary Figure S7). Except for the Phe59 mutant, all the proteins were correctly folded. In perfect agreement with our structure, all the side chains involved in C2, C3 and G5 binding or sequence-specific recognition showed a high decrease in affinity when mutated to an alanine. Deletion of the N- or C-terminus led to an affinity decrease smaller than expected considering the number of protein–RNA contacts involving both termini. Yet, the large conformational loss of entropy of the termini upon RNA binding could explain the overall weak affinity gain provided by these termini. Altogether, the structure of SRSF2 in complex with 5'-UCCAGU-3' suggests a clear sequence-specific binding of the 5'-CCNG-3' sequence, which appears in contradiction with the degenerated consensus motifs identified by SELEX (Cavaloc *et al*, 1999).

Affinity measurements establish a 5'-SSNG-3' consensus binding motif for SFSR2

To further confirm or invalidate the 5'-CCNG-3' binding consensus sequence derived by the complex structure, we decided to replace each nucleotide of the central motif with one of the other three nucleotides and to measure the difference

in binding affinity by ITC. In agreement with our structure, the importance of G5 is emphasized by this study, since a change of G5 into any other base led to a drastic decrease in binding affinity (Table II). Similarly, the non sequence-specific recognition of A4 observed in the structure is confirmed since its replacement by a different nucleotide did not result in any change in affinity. More surprisingly, although our structure indicated a high specificity for cytosines at positions 2 and 3, ITC measurements revealed that SRSF2 can equally well accommodate a guanine at each position without any loss in binding affinity but that affinity is reduced when changed to uracil or adenine (Figure 3C; Table II). In fact, replacement of both cytosines by guanines (5'-UGGAGU-3') resulted in an even higher affinity (K_d of 0.22 μM) compared with cytosine containing RNA (Figure 4A).

Overall, our ITC data confirm that SRSF2 binds sequence specifically 5'-CCNG-3'; however, the binding consensus sequence is more degenerated because guanine is equally well tolerated in the first two positions. The high-affinity SRSF2 binding consensus sequence is therefore 5'-SSNG-3' (S = G/C) with a K_d ranging from 0.2 to 0.4 μM. Yet, the next question for us to address was how SRSF2 can accommodate two guanines instead of two cytosines? Considering that the Watson–Crick edge of cytosine and guanine is perfectly complementary, it was clear that the guanines cannot just

Figure 2 Overview of the SRSF2 RRM binding to the 5'-UCCAGU-3' RNA. (A) Overlay of ^1H - ^{15}N HSQC spectra representing an NMR titration of ^{15}N -labelled SRSF2 RRM with unlabelled 5'-UCCAGU-3' RNA. The peaks corresponding to the free and RNA-bound protein states (RNA:protein ratio 1:1) are in blue and red. Negative peaks are coloured green. Arrows indicate chemical shift perturbations higher than 0.5 p.p.m. (B) Mapping of the combined chemical shift perturbations ($\Delta\delta = [(\delta\text{HN})^2 + (\delta\text{N}/6.41)^2]^{1/2}$) upon RNA binding over the amino-acid residue number. The position of the secondary structure elements is shown above the graph. The largest chemical shift perturbations (>0.5 p.p.m.) are indicated. The largest shift for Arg91 (*) can only be seen in a construct without HIS tag (Supplementary Figure S1). (C) Binding affinity of SRSF2 RRM for the 5'-UCCAGU-3' RNA measured by ITC. Raw data and corresponding binding curve are depicted. Mean value for the dissociation constant (K_d) with standard deviation is based on three independent measurements. (D) Backbone traces (N, C α and C') of the 20 lowest energy structures of the SRSF2 RRM (1–101) in complex with the 5'-UCCAGU-3' RNA superimposed on the backbone of the structured part. The protein backbone is coloured as in Figure 1B. The central motif of the RNA is shown in stick representation, with the carbon atoms in yellow, nitrogen in blue, phosphate in orange and oxygen in red; the two non-defined flanking uracils were omitted for better overview. Amino acids 1–6 and 94–101 are omitted in the 45° rotated view. (E) Surface representation in stereo view of the most representative structure, with the protein backbone in ribbon and the 5'-UCCAGU-3' RNA in sticks. Important protein side chains involved in RNA interaction are represented as sticks. Colour code as in Figure 1B and D. Figures (D, E) were generated with MOLMOL (Koradi *et al*, 1996).

Figure 3 Specificity of the interaction between SRSF2 and the 5'-UCCAGU-3' RNA. (A) Close-up view of the four central nucleotides C2, C3, A4 and G5 as described in Figure 2D. Hydrogen bonds are presented as violet dashed lines. Figure was generated with MOLMOL (Koradi *et al*, 1996). (B) Affinity matrix obtained by ITC measurements. Nucleotides in big bold letters have no effect on binding affinity and small blue letters decrease binding affinity more than five-fold when compared with the original sequence 5'-UCCAGU-3'.

replace the cytosines; therefore, either the protein must adapt to the RNA or the RNA to the protein.

Two guanines adopting a *syn* conformations explains how SFSR2 recognizes 5'-SSNG-3'

To understand, how SRSF2 recognizes equally well a cytosine and guanine at positions 1 and 2 of the 5'-SSNG-3' consensus

sequence, we determined the solution structure of the SRSF2 RRM bound to 5'-UGGAGU-3'. Note that 5'-GGAG-3' is found in the second divergent sequence selected by *in-vitro* SELEX (5'-AGGAGAU-3') and in the loop II (5'-GAGGAG-3') of the HIV-1 tat exon 2 (Cavaloc *et al*, 1999; Zahler *et al*, 2004; Hallay *et al*, 2006). The NMR data recorded in the presence of the 5'-UGGAGU-3' RNA indicated that the same regions of the

Table II Binding affinities of wild-type and variants of the SRSF2 RRM with 5'-UCCAGU-3' RNA and 5'-UGGAGU-3' RNA

	K_d (μ M)	Affinity decrease
SRSF2 + 5'-UCCAGU-3'		
Wild type (aa 1–101)	0.27 ± 0.02	
K17A	2.38 ± 0.38	9
D42A	4.09 ± 0.58	15
Y44A	> 5	> 20
D48A	0.83	3
S54A	0.37	1
F59A	Unfolded	—
R61A	> 5	> 20
Q88A	0.35	1
Δ N-Ter (aa 11–101)	0.6 ± 0.05	2
Δ C-Ter (aa 1–91)	1.1 ± 0.02	4
SRSF2 + 5'-UGGAGU-3'		
Wild type (aa 1–101)	0.22 ± 0.03	
K17A	2.8 ± 0.1	13
D42A	3.7 ± 0.8	17
Y44A	> 5	> 20
D48A	0.6	3
F59A	Unfolded	—
R61A	> 5	> 20
R86A	0.46	2
Q88A	0.23	1
Δ N-Ter (aa 11–101)	0.86 ± 0.04	4
Δ C-Ter (aa 1–91)	4.4 ± 0.0	20
RNA (5' $\rightarrow$ 3')		
cCCAGU	0.47 ± 0.02	2
UgCAGU	0.41 ± 0.03	2
UaCAGU	1.92 ± 0.06	7
UuCAGU	2.7 ± 0.1	10
UCaAGU	5.2 ± 0.38	19
UCgAGU	0.36 ± 0.02	1
UCuAGU	6.4 ± 0.2	> 20
UCCuGU	0.36 ± 0.02	1
UCCgGU	0.34 ± 0.04	1
UCCcGU	0.39 ± 0.03	1
UCCaAu	7.6 ± 1.4	> 20
UCCAcU	17.2 ± 3	> 20
UCCAuU	11.4 ± 1	> 20

SRSF2 RRM were affected by this interaction, although the directionality and the magnitude of the chemical shift perturbations were quite different due to different ring current effect of purine compared with pyrimidine (Supplementary Figure S8). Using 2264 NOE-based distance constraints (including 83 intermolecular ones), we could determine a precise structure of the complex (r.m.s. deviation of 1.47 Å for all heavy atoms; Table I).

The overall conformation of the complex shows the same sandwich-like structure, with the RNA located between the β -sheet surface on one side and the N- and C-termini on the other side (Figure 4B). In this second complex, the RNA bases are located in the same binding pockets than in the complex with 5'-UCCAGU-3' (Figure 4C; Supplementary Table S3); however, G2 and G3 adopt now a *syn* conformation whereas all the nucleotides of the first complex adopted an *anti* conformation (Figure 4D). NMR evidence for the *syn* conformations for G2 and G3 is the observation of a strong H8-H1' intranucleotide NOE (Supplementary Figure S9). In adopting a *syn* conformation, guanines expose their Hoogsteen edge that resembles the Watson–Crick edge of a cytosine. Indeed, G2 N7 forms a hydrogen bond with the main-chain amide of Arg91 resulting in a similar shifted peak in the ^1H - ^{15}N -HSQC (Supplementary Figure S7C). Furthermore, G2 stacks on the hydrophobic part of Lys17,

but instead of contacting the phosphate backbone, the amino group of Lys17 is hydrogen bonded to G2 N3. Thus, the unusual lysine in the RNP2 motif could be an important side chain to allow the recognition of both cytosine and guanine at this position in SRSF2.

Like C3 in the first complex, G3 stacks on Phe59 and forms a very similar network of hydrogen bonds with both the Arg61 side chain and the Tyr92 main chain (Figure 4D). In addition, the *syn* conformation for G2 and G3 results in an intra-RNA hydrogen bond between G3 N2 and G2 O2'. This additional contact might explain the slightly higher binding affinity observed in the presence of two guanines.

In order to determine whether the *syn* guanine is also observed when only one cytosine is substituted, we formed another protein–RNA complex between the SRSF2 RRM and 5'-UGCAGU-3' and assigned the resonances of this complex. Here, C3 adopts an *anti* conformation and G2 again a *syn* conformation as evidenced by the strong correlation of the ribose H1' and the H8 of the base (Supplementary Figure S9). The binding mode is similar to what was observed with the other two RNA sequences, since the pattern of intermolecular NOE is nearly identical to the other complexes.

In conclusion, with these two structures of SRSF2 RRM bound to RNA, we can now understand how the same RRM can recognize the different sequences 5'-CCAG-3', 5'-GGAG-3', 5'-GCAG-3' and 5'-CGAG-3' equally well and sequence specifically. The RNA simply adapts to the binding surface in the first two nucleotide binding pockets, adopting an *anti* conformation when a cytosine is present and a *syn* conformation when it is a guanine. This explains how SRSF2 can achieve a highly degenerated sequence specificity.

The 5'-SSNG-3' consensus motif is also required for splicing activation of SRSF2 *in vitro*

In order to investigate, whether the identified consensus motif 5'-SSNG-3' is functionally important for the role of SRSF2 as splicing activator, we conducted *in-vitro* splicing assays with the full-length SRSF2 produced in insect cells to allow phosphorylation of the RS domain (Qian *et al*, 2011). We used the heterologous Sp1 'inverted exon 2' derived from an adenovirus pre-mRNA as a splicing reporter (Figure 5A; Dreumont *et al*, 2010). Only the insertion of a valid enhancer sequence into exon 2 can lead to splicing activation *in vitro*. Therefore, we inserted a 20 nucleotide SELEX-derived sequence (previously named S94) that contains the sequence 5'-UCCAGU-3' that we used to solve the first complex structure. This sequence gave rise to a splicing activation of 26% upon addition of SRSF2, while a mutant (5'-aaaaa-3') within the same 20 nucleotide sequence showed under the same condition only 6% residual splicing activation (4.3-fold decrease) (Figure 5B). Consequently, this demonstrates that the hexanucleotide sequence acts as an ESE recognized by SRSF2 in this *in-vitro* system.

Next, we investigated the effect of specific point mutations in this ESE on splicing activation. In good agreement with our structural data, each sequence containing a cytosine or a guanine at positions 1 and 2 and a guanine at position 4 of the consensus sequence allows splicing activation in a range from 26 to 13% (Figure 5C). This corresponds to a two- to four-fold increase in splicing activity when compared with the control. A substitution by another nucleotide like adenine at position 2 or uracil at position 1 severely disrupted splicing

Figure 4 Overview of SRSF2 RRM binding to the 5'-UGGAGU-3' RNA. (A) Binding affinity of SRSF2 RRM to the 5'-UGGAGU-3' RNA measured by ITC, as described in Figure 2C. (B) Backbone traces of the 20 lowest energy structures of the SRSF2 RRM in complex with the 5'-UGGAGU-3' RNA superimposed on the backbone of the structured part, as described in Figure 2D. (C) Surface representation (stereo view) of the most representative structure, as described in Figure 2E. (D) Close-up view of the semi-specifically recognized G2 and G3 as described in Figure 2D. Figures (B–D) were generated with MOLMOL (Koradi *et al*, 1996).

(only 5–6% activation, equivalent to the residual activation). Further mutations revealed that the highly specifically bound G5 shows a similar strong decrease in splicing activity upon replacement by either an adenine or a cytosine (8 and 9% activation) and is so only slightly above the residual splicing activity. Together, these results show that the consensus binding sequence 5'-SSNG-3' found with our structural approach parallels the consensus sequence needed for splicing activation of the full-length protein *in vitro*.

Mutants of the SRSF2 RRM affecting RNA binding also affect splicing *in vivo*

To test the functional importance of the RNA binding contacts we found in the structures, we conducted *in-vivo* splicing assays with the full-length protein. SRSF2 autoregulates its own expression by splicing of its own pre-mRNA leading to non-sense-mediated decay (Dreumont *et al*, 2010). The splicing activating element 5'-UGCAGU-3' within the 3'-terminal exon of the SRSF2 gene is essential for SRSF2-mediated splicing and similar to the RNA sequences we used for our NMR investigations. This 3'-terminal region of the SRSF2 gene was inserted into the β -globin intron of a CMV promoter-driven plasmid resulting in a minigene referred to as pSC35- β Glo (Figure 6A; Dreumont *et al*, 2010). The construct contains two mutually exclusive 3' splice sites and overexpression of SRSF2 in HeLa cells results in a two-fold increase of splicing of the most proximal, SRSF2-specific 3' splice site (Figure 6B). We then tested the effect of SRSF2 single amino acid mutants on RNA

binding in this *in-vivo* context. The expression of each mutant was monitored by western blot to ensure that a decrease in splicing activity is not associated with poor protein expression (Supplementary Figure S10). Asp42 and Tyr44, which are both involved in the specific recognition of G5, showed a decrease in splicing activity upon mutation to alanine (Figure 6C). Furthermore, an alanine substitution of Arg61, involved in C3 recognition, also severely disrupted splicing activation. The Phe59 mutant on the other hand could not be detected by western blot, possibly as it is unfolded similarly to our *in-vitro* observations. We used mutants of Ser54 and Arg86 as negative control, since they did not show a strong decrease of binding affinity in the ITC measurements and correspondingly did not affect splicing activity *in vivo*. In summary, the key SRSF2 side chains shown to be important for the RNA binding interface proved also to be functionally important for splicing activity of the full-length protein in a cellular context. This demonstrates that splicing activation by SRSF2 *in vitro* and *in vivo* strongly depends on the specific recognition of a 5'-SSNG-3' consensus sequence within an ESE.

Discussion

The unusual flexibility of RNA recognition by SRSF2

In this study, we explain how SRSF2 can bind an apparent high diversity of sequences with similar affinity in nanomolar range and clarify its specificity of RNA recognition. Until now,

Figure 5 Effect of nucleotide substitutions in the 5'-UCCAGU-3' motif on SRSF2 splicing activity *in vitro*. **(A)** Scheme of the Sp1 'inverted exon 2' reporter, adapted from Dreumont *et al* (2010). A test sequence containing the SELEX-derived S94 RNA was inserted into exon 2 and tested for splicing activation. **(B)** *In-vitro* splicing assays using cytoplasmic S100 and nuclear extract (ratio 4:1) with the Sp1 'inverted exon 2' splicing substrate. Each transcript embodied the SELEX-derived S94 sequence in exon 2 with various point mutations in the 5'-UCCAGU-3' SRSF2 binding site. After normalization, the ratio for splicing activation of the respective negative control was subtracted for each sample. Mean value and standard deviation of three independent experiments are shown below. Colours are as described in Figure 3B, with low-affinity mutants in small blue and high-affinity mutants in big bold letters. **(C)** Graphical depiction of mean value and standard deviation for each transcript. The dashed blue line illustrates the residual splicing activity of the control.

only three RRM structures of RS-containing proteins have been solved in complex with RNA, namely SRSF3 (Hargous *et al*, 2006), Tra2- β 1 (Clery *et al*, 2011; Tsuda *et al*, 2011) and in this paper, SRSF2. Whereas the consensus motif and the mode of interaction are quite different, the mode of binding has common features between these proteins. Similarly to SRSF3 (Hargous *et al*, 2006), the RNA adopts the same conformation on the surface of the RRM with nucleotide one, two and four located on the β -sheet surface and nucleotide three interacting with loop 3 (Supplementary Figure S11A). The mode of recognition of the 5' cytosine is identical between the two proteins and is observed in a total of eight different structures of RRM-protein complexes (Supplementary Table S4; Auweter *et al*, 2006), but only SRSF2 can equally well recognize a 5' guanine at this position. Yet, sequence specificity is different for the RRM of SRSF2 that recognizes a 5'-SSNG-3' sequence when SRSF3 recognizes a 5'-CNNC-3' sequence. Moreover, the N- and C-termini play an important role for RNA binding in SRSF2 and not for SRSF3. In this respect, RNA binding by SRSF2

resembles Tra2- β 1 where both termini are involved in RNA recognition. However, while both termini are parallel in SRSF2, they cross each other upon RNA binding in Tra2- β 1 (Clery *et al*, 2011; Tsuda *et al*, 2011). Unique to SRSF2 is the fact that the N-terminal region is relocated in the RNA-bound state compared with the free protein (Figures 1 and 2). This protein conformational change is somehow reminiscent of what was observed earlier in the N-terminal RRM of U1A where a C-terminal helix is also relocated upon RNA binding (Allain *et al*, 1996).

Similarities can also be observed between SRSF2 bound to the 5'-GGNG-3' motif and hnRNP A1 in complex with a telomeric DNA repeat (Supplementary Figure S11B). In both cases, the guanine that stacks on the conserved phenylalanine of the hnRNP1 A1 motif adopts a *syn* conformation and forms hydrogen bonds between its Hoogsteen face and a lysine side chain. Furthermore, the imino group forms a hydrogen bond with the main chain of the C-terminal region. Other common features include the guanine besides β -strand 2 that forms identical hydrogen bonds with an aspartate side

Figure 6 Effect of point mutations in the SRSF2 RNA binding interface on its splicing activity *in vivo*. (A) Scheme of the pSC35-βGlo minigene, adapted from Dreumont *et al* (2010). The 3'-terminal intronic and exonic region of the SRSF2 gene was cloned into the rabbit β-globin intron 2. (B) RNA analysis of *in-vivo* splicing assays after RT-PCR. The minigene was co-transfected into HeLa cells, overexpressing various SRSF2 mutants. Splicing activation by SRSF2 led to an increased use of the 3' splice site of the SRSF2 terminal exon compared with the β-globin 3' splice site. Mean value and standard deviation of three independent experiments are shown below the gel. (C) Graphical depiction of mean value and standard deviation for each SRSF2 mutant.

chain and a β-turn right before β-strand 1. Interestingly, these two proteins are antagonists in splicing regulation and compete for overlapping binding sites (Zhu *et al*, 2001; Okunola and Krainer, 2009). As a result, they not only recognize similar RNA sequences but also seem to share a very similar RNA binding mode for these two guanines.

The most unusual feature of SRSF2 RNA recognition is the unexpected manner used by the RRM to allow the accommodation of either G or C nucleotides at the two most 5' positions of the binding sequence. It was surprising that guanine and cytosine could be equally well recognized in two binding pockets since their Watson-Crick faces are perfectly complementary to each other. There are three possibilities for two so different bases to be equally well recognized: either the protein pocket adapts to the RNA or the RNA to the protein or both. An example of the first possibility is found when comparing the structure of HuD and Sex lethal. In both proteins, the same binding pocket can accommodate either an adenine or a uracil through a conserved glutamine that can rotate its side chain to contact one or the other base without changing the RNA conformation (Supplementary Figure S12; Handa *et al*, 1999; Wang and Tanaka Hall, 2001). SRSF2 uses the second possibility in the binding

pockets for the first two nucleotides with the cytosines adopting the expected *anti* conformation while the guanines adopt the more unusual *syn* conformation. This is possible without changing the protein conformation because the Hoogsteen edge of a guanine resembles the Watson-Crick edge of a cytosine permitting similar hydrogen bonds to be formed. This equivalence between a *syn* G and an *anti* C has been observed in several RNA structures like the MicroROSE element (Chowdhury *et al*, 2006) or in the ribozyme core of the hepatitis delta virus RNA (Been and Perrotta, 1995), where base pairing between the Hoogsteen edge of a *syn* G and the Watson-Crick edge of an *anti* G occurs. It was even previously reported that the most efficient interaction between a G-G base pair requires one guanine to be flipped into a *syn* conformation in order to form hydrogen bonds with its Hoogsteen face, thereby mimicking the cytosine (Burkard and Turner, 2000). The two structures of SRSF2 bound to the 5'-UCCAGU-3' and 5'-UGGAGU-3' RNA reveal that exactly the same mechanism is used by the RRM for protein-RNA recognition (Supplementary Figure S13).

Interestingly, we could find another very similar mode of RNA adaptation in another RRM but involving adenine and uracil. The structure of HuD was previously solved in com-

plex with two different AU-rich RNA sequences (Wang and Tanaka Hall, 2001). In one binding pocket, a uracil or an adenine can be accommodated both with the uracil adopting an *anti* conformation and the adenine a *syn* conformation (Supplementary Figure S14). Although no affinity measurement was performed to verify that the RRM can equally well accommodate both bases, the resemblance with how SRSF2 RRM accommodates the first nucleotide but with C *anti* and G *syn* is striking and explains why HuD has been associated with AU-rich binding preference (Wang and Tanaka Hall, 2001). To know if this simple mode of RNA adaptation seen for two RRM structures is more generally found in other RRMs with degenerated sequence recognition, more examples will need to be investigated.

The 5'-SSNG-3' consensus sequence facilitates SRSF2-specific ESE prediction

To date, the specificity of interaction between SRSF2 and its RNA targets has been elusive. Due to the lack of a clear consensus binding sequence, several SRSF2-specific ESEs were identified with no certainty that the ESE is really responsible for SRSF2 binding. The two structures of SRSF2 RRM bound to the two RNA sequences described here (5'-GGAG-3' and 5'-CCAG-3') combined with our *in-vitro* and *in-vivo* splicing data clearly lift the mystery associated with SRSF2 sequence specificity. Indeed, our structure revealed a 5'-SSNG-3' high-affinity binding consensus sequence for SRSF2 that could now be found in all SELEX consensus sequences and in all identified SRSF2-specific ESEs (Table III; Supplementary Table S5).

In addition to be found in all ESEs regulated by SRSF2, biochemical data obtained from others support direct interactions between natural 5'-SSNG-3' type of sequences and the full-length protein SRSF2 (Hallay *et al*, 2006). Footprinting experiments performed *in vitro* with the SLS3 HIV-1 RNA region in presence of full-length recombinant SRSF2 protein revealed two parts protected on this RNA: the 5'-strand of loop II, containing a 5'-GGAG-3' sequence, and the ESS2 with the terminal loop of the B motif containing 5'-AGAG-3' and 5'-GAAG-3' sequences. The strongest protections were observed at 5'-GGAG-3' in good agreement with our expected

higher affinity for SRSF2 for this sequence compared with the two other binding sites. The protection observed in the other sequence is most likely due to the close proximity of two binding sites of weaker affinity. Using the same method, it was also shown that the 14-nt apical loop located in the terminal exon of SRSF2 is directly bound by SRSF2 which can now easily be explained by the presence of two overlapping sequences 5'-GGUG-3' and 5'-GGCG-3' that perfectly fit the 5'-SSNG-3' consensus motif (Dreumont *et al*, 2010).

The consensus binding sequence that we could now define precisely will be very useful to identify more accurately SRSF2 binding sites and to predict new putative RNA targets of this splicing factor. For example, the current ESEfinder software searches for C-rich consensus sequences based on the functional selection of SRSF2-dependent ESEs and does not consider the possibility that SRSF2 can bind equally well a 5'-GGNG-3' motif (Cartegni *et al*, 2003). In fact, in good agreement with our results showing that the 5'-GGNG-3' motif has the higher affinity for SRSF2 compared with the 5'-CCNG-3' motif (Table II), most of the ESEs contain G-rich motifs rather than C-rich motifs (Table III). This will need to be considered for a more precise prediction of possible SRSF2-dependent ESEs and in addition, our final consensus sequence will be important to refine the proposed splicing code (Barash *et al*, 2010).

Identification of the 5'-SSNG-3' consensus sequence has also strong implications when it is required to detect new mutations in SRSF2 binding sites that could be at the origin of a genetic disease. Indeed, a single nucleotide substitution can be sufficient to prevent the recruitment of SRSF2 and then profoundly change the splicing pattern of a pre-mRNA. For example, it was shown that the two substitutions 5'-CCAGTA-3' to 5'-CCAATC-3' in an ESE bound by SRSF2 in the exon 2 of HIV-1 tat pre-mRNA strongly decreased tat intron 1 splicing (Zahler *et al*, 2004). Moreover, a single 5'-CCAG-3' to 5'-CCAA-3' substitution in the tau exon 10 strongly affects the recruitment of SRSF2 and as a result decreases the level of inclusion of this exon (Qian *et al*, 2011). These two examples can now be easily explained by the expected strong decrease of SRSF2 binding when the G of the 5'-CCNG-3' sequence is changed into another base (Figure 3B). In the example of tau

Table III RNA target sequences identified as binding sites *in vitro* and ESE *in vivo* for SRSF2

	SSNG ^a	Description	
A	GSAG	AGUA	SELEX
GUU	CGAG	UA	
UGUU	CSAG	WU	SELEX
GWUU	CCUG	CUA	
GUGU	GCUG		
GA	CGAG	UAGKS	
A	GGAG	AU	
U	CCNG	YY	Functional SELEX
GRYY	CSYR		Functional SELEX (Supplementary Table S5)
	CCAG	UAGAUC	HIV-1 tat exon 2 (ESE2)
GA	GGAG		HIV-1 tat exon 2 (5' loop II)
GUAGCC	CGAG	A	HIV-1 exon 6D (ESE)
U	CGAG	AUAAUU	Tau exon 10 (ESE)
U	CCUG	UU	β-globin exon 2 (ESE)
AU	CCCG	U	GluR2 flop exon (ESE)
GUU	CGAG	U	SRSF2 gene (ESE)
UGAU	CGAG	A	Apolipoprotein A-II exon 3 (ESE)
			Arrisi-Mercado <i>et al</i> (2004)

^aS = G/C, W = A/U, K = G/U, Y = C/U, R = G/A.

exon 10, our findings could be useful to explain mutations responsible for the neurodegenerative disease tauopathies (Ballatore *et al*, 2007).

A single nucleotide substitution can not only disrupt an existing ESE site, but also create a new ESE. In an HIV-1 strain identified for its abnormal splicing pattern associated with a dramatic decrease in viral replication, it was shown that a naturally occurring U to C mutation in the sequence 5'-AGUAG-3' of exon 6D creates an ESE for SRSF2 (Wentz *et al*, 1997). While a mutation to 5'-AGGAG-3' also increased splicing, a mutation to 5'-AGAAG-3' had only a weak effect (Caputi and Zahler, 2002). This example again confirms our 5'-SSNG-3' consensus sequence for SRSF2 and shows that it is very suitable to predict SRSF2 targets. In conclusion, our structure-based consensus sequence for RNA binding by SRSF2 sharpens our knowledge about the regulation of splicing events. It enhances the detection of disrupted or created SRSF2-dependent ESEs in the genome of patients and could therefore be useful to find new strategies to target specifically the cause of a disease and to develop appropriate treatments.

Materials and methods

Protein and RNA preparation

We cloned the ORF of SRSF2 RRM (1–101) into the pet28a plasmid using the restriction sites *Bam*H1/*Xho*1. The protein was over-expressed at 37°C for 3 h in *Escherichia coli* BL21 (DE3) codon plus cells in minimal M9 medium (1 g l⁻¹ ¹⁵N-NH₄Cl, 2 g l⁻¹ ¹³C-glucose) using 0.1 mM IPTG. The protein was then purified by two successive nickel affinity chromatography steps and dialysed against the NMR buffer (50 mM L-Glu, 50 mM L-Arg and 20 mM Na₂HPO₄/NaH₂PO₄ at pH 5.5). A last purification step by size exclusion chromatography with a Superdex75 column (GE Healthcare) was necessary to remove residual RNases in the solution. The protein could be concentrated to over 2 mM with a 10-kDa molecular mass cutoff membrane.

RNA was purchased from Dharmacon, deprotected according to the manufacturer's protocol, purified by butanol extraction, lyophilized and resuspended in NMR buffer.

RNA-protein complexes used as NMR samples were prepared at an RNA:protein ratio of 1:1 in a final volume of 250 µl and a concentration of 0.75 mM.

NMR measurements and resonance assignments

All the NMR measurements were recorded in the NMR buffer (50 mM L-Glu, 50 mM L-Arg and 20 mM Na₂HPO₄/NaH₂PO₄ at pH 5.5) and at 310 K using Bruker AVIII-500 MHz, AVIII-600 MHz, AVIII-700 MHz and Avance-900 MHz spectrometers, all equipped with cryoprobes.

Data processing was performed with Topspin 2.1 (Bruker) and analysis with sparky (<http://www.cgl.ucsf.edu/home/sparky/>).

We used for the backbone, aliphatic and aromatic side-chain assignments: 2D (¹⁵N-¹H) HSQC, 2D (¹³C-¹H) HSQC, 3D HNCA, 3D HNCOC, 3D HNCOC, 3D HNCAC, 3D CBCACONH, 3D HNCACB, 3D HcccNH TOCSY, 3D hCcccNH TOCSY, 3D HCcH TOCSY, 3D NOESY (¹⁵N-¹H) HSQC and 3D NOESY (¹³C-¹H) HSQC, all recorded in H₂O (Sattler *et al*, 1999).

To assign the resonances of the unlabelled RNA, we recorded: 2D (¹H-¹H) TOCSY, 2D (¹H-¹H) NOESY, 2D 1F-filtered 2F-filtered (¹H-¹H) NOESY (Peterson *et al*, 2004) and natural abundance 2D (¹³C-¹H) HSQC, all in D₂O.

Intermolecular NOEs were obtained by using a 2D (¹H-¹H) NOESY, 2D F2 filtered (¹H-¹H) NOESY and 3D F1 filtered F2 edited (¹H-¹H) NOESY (Lee *et al*, 1994), all recorded in D₂O. In addition, intermolecular NOEs between imino proton of G3 and the SRSF2 RRM protons were obtained using 2D (¹H-¹H) NOESY in H₂O at 288 K.

We used a mixing time of 150 ms for NOESY spectra, 23 ms for 3D TOCSY spectra and 50 ms for 2D TOCSY spectra.

Structure calculation and refinement

AtnosCandid software (Herrmann *et al*, 2002a,b) was used to automatically generate peak lists from 2D (¹H-¹H) NOESY and 3D NOESY (¹⁵N- and ¹³C-edited) HSQC spectra. After manually refinement of each list, intermolecular NOE distance constraints were automatically assigned through seven cycles using CYANA with the macro *noeassign* (Herrmann *et al*, 2002a). We included additional hydrogen bond constraints derived from hydrogen-deuterium exchange experiments on the amide protons. Intramolecular RNA and intermolecular distance restraints were manually assigned and calibrated based on fixed interatomic distances. Calculation was then conducted with CYANA 3.0. Starting from random structures, 250 preliminary structures were calculated and the 50 structures having the lowest target function were selected for further refinement. This was done by a restrained simulated annealing run in implicit water with the SANDER module of AMBER 9 (Case *et al*, 2005) using the ff99 force field (Wang *et al*, 2000). The final best 20 structures were selected based on lowest energy and NOE violations and analysed with PROCHECK (Laskowski *et al*, 1996).

Isothermal titration calorimetry

Measurements were conducted on a VP-ITC instrument (MicroCal), which was calibrated according to the manufacturer's protocol. Concentrations of RNA and protein were calculated based on their optical density absorbance at 260 or 280 nm, respectively. The sample cell was loaded with 1.4 ml of 20 µM RNA and the syringe with 0.4 mM of protein. Measurements were done at 37°C in the NMR buffer using at least 35 consecutive injections of protein (5 µl). Data were integrated and normalized using the Origin 7.0 software according to a 1:1 RNA:protein ratio binding model. Standard deviation is based on two (mutants) to three (wild type) independent measurements.

In-vitro splicing assay

For each splicing reaction, 200 ng of SRSF2 protein expressed in baculovirus was used with a ratio of S100 cytoplasmic fraction to nuclear extract of 4:1. The splicing reaction with ³²P-labelled RNA was conducted at 31°C for 90 min in the buffer conditions described before (Disset *et al*, 2006). After protein denaturation with proteinase K and phenol/chloroform extraction, RNA was precipitated and loaded onto a 6% denaturing acrylamide gel. The radioactivity corresponding to the unspliced and spliced RNA was detected using a Phosphor-Imager and radioactive bands quantified using the ImageQuant software (GE Healthcare). For each RNA mutant, the percentage of splicing activation corresponding to the negative control (in the absence of SRSF2 addition) was subtracted from the percentage determined in the presence of SRSF2. The mean value with standard deviation is based on three independent *in-vitro* splicing experiments.

Transfection and RT-PCR analysis

HeLa cells were maintained in Dulbecco's modified Eagle's medium (DMEM; GibcoBRL) supplemented with 10% FBS (GibcoBRL). The pSC35-βGlo reporter was described before (Dreumont *et al*, 2010) and is schematically represented in Figure 6A. The ORF of full-length SRSF2 was cloned into the mammalian expression vector pcDNA 3.1 with an N-terminal HA tag. HeLa cells were grown in 6-well plates and after 1 day transfected with the calcium-phosphate method using 1 µg of each plasmid. Cells were grown for 1 day, then harvested and total RNA isolated. RNA was reverse transcribed using the M-MuLV reverse transcriptase RNaseH⁻ (Promega) according to the manufacturer's protocol. RT-PCR was performed for 30 cycles (94°C/30 s—55°C/20 s—72°C/60 s) with forward (5'-ACG GTG CAT TGG AAC GGA CCC-3') and reverse (5'-GTA ACC ATT ATA AGC TGC AAT-3') primers. The labelled cDNAs were then detected and quantified as described above. The percentage of exon inclusion (SC35-specific isoform) from four independent *in-vivo* splicing assays was calculated with standard deviation.

Accession codes

Atomic coordinates and NMR restraints for the structures of SRSF2, SRSF2 + 5'-UCCAGU-3' and SRSF2 + 5'-UGGAGU-3' have been deposited in the Protein Data Bank under accession codes 21ea, 21eb and 21ec, respectively.

Supplementary data

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

Acknowledgements

We would like to thank M Blatter for his help with structure calculations; F Damberger and M Schubert for their invaluable knowledge of NMR spectroscopy; J Boudet for his advice and S Gerhardt for his assistance in the ITC measurements. Research of FH-TA is supported by the Swiss National Foundation, National Center for Competence in Research Structural Biology and

EURASNET. AC was supported by the European Molecular Biology Organization for a postdoctoral fellowship.

Author contributions: FH-TA, AC and JS designed the project; GMD prepared protein and RNA samples for structural studies; GMD, AC and FH-TA analysed NMR data; GMD and AC set up structure calculations; GMD and AC did ITC measurements; GMD and SJ conducted *in-vitro* splicing assays; GMD did *in-vivo* splicing assays; GMD, AC and FH-TA wrote the manuscript; all authors discussed the results and approved the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Allain FH, Gubser CC, Howe PW, Nagai K, Neuhaus D, Varani G (1996) Specificity of ribonucleoprotein interaction determined by RNA folding during complex formulation. *Nature* **380**: 646–650
- Arrisi-Mercado P, Romano M, Muro AF, Baralle FE (2004) An exonic splicing enhancer offsets the atypical GU-rich 3' splice site of human apolipoprotein A-II exon 3. *J Biol Chem* **279**: 39331–39339
- Auweter SD, Oberstrass FC, Allain FH (2006) Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res* **34**: 4943–4959
- Ballatore C, Lee VM, Trojanowski JQ (2007) Tau-mediated neurodegeneration in Alzheimer's disease and related disorders. *Nat Rev Neurosci* **8**: 663–672
- Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ (2010) Deciphering the splicing code. *Nature* **465**: 53–59
- Been MD, Perrotta AT (1995) Optimal self-cleavage activity of the hepatitis delta virus RNA is dependent on a homopurine base pair in the ribozyme core. *RNA* **1**: 1061–1070
- Black DL (2003) Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**: 291–336
- 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
- Burkard ME, Turner DH (2000) NMR structures of r(GCAGGCGUGC)2 and determinants of stability for single guanosine-guanosine base pairs. *Biochemistry* **39**: 11748–11762
- Caputi M, Zahler AM (2002) SR proteins and hnRNP H regulate the splicing of the HIV-1 *tev*-specific exon 6D. *EMBO J* **21**: 845–855
- Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR (2003) ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res* **31**: 3568–3571
- Case DA, Cheatham III TE, Darden T, Gohlke H, Luo R, Merz Jr KM, Onufriev A, Simmerling C, Wang B, Woods RJ (2005) The Amber biomolecular simulation programs. *J Comput Chem* **26**: 1668–1688
- 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
- Cazalla D, Zhu J, Manche L, Huber E, Krainer AR, Caceres JF (2002) Nuclear export and retention signals in the RS domain of SR proteins. *Mol Cell Biol* **22**: 6871–6882
- Chowdhury S, Maris C, Allain FH, Narberhaus F (2006) Molecular basis for temperature sensing by an RNA thermometer. *EMBO J* **25**: 2487–2497
- Clery A, Jayne S, Benderska N, Dominguez C, Stamm S, Allain FH (2011) Molecular basis of purine-rich RNA recognition by the human SR-like protein Tra2-beta1. *Nat Struct Mol Biol* **18**: 443–450
- Colwill K, Pawson T, Andrews B, Prasad J, Manley JL, Bell JC, Duncan PI (1996) The Clk/Sty protein kinase phosphorylates SR splicing factors and regulates their intranuclear distribution. *EMBO J* **15**: 265–275
- Crovato TE, Egebjerg J (2005) ASF/SF2 and SC35 regulate the glutamate receptor subunit 2 alternative flip/flop splicing. *FEBS Lett* **579**: 4138–4144
- Das R, Yu J, Zhang Z, Gygi MP, Krainer AR, Gygi SP, Reed R (2007) SR proteins function in coupling RNAP II transcription to pre-mRNA splicing. *Mol Cell* **26**: 867–881
- de la Mata M, Kornblihtt AR (2006) RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20. *Nat Struct Mol Biol* **13**: 973–980
- Disset A, Bourgeois CF, Benmalek N, Claustres M, Stevenin J, Tuffery-Giraud S (2006) An exon skipping-associated nonsense mutation in the dystrophin gene uncovers a complex interplay between multiple antagonistic splicing elements. *Hum Mol Genet* **15**: 999–1013
- Dominguez C, Schubert M, Duss O, Ravindranathan S, Allain FH (2011) Structure determination and dynamics of protein-RNA complexes by NMR spectroscopy. *Prog Nucl Magn Reson Spectrosc* **58**: 1–61
- Dreumont N, Hardy S, Behm-Ansmant I, Kister L, Branlant C, Stevenin J, Bourgeois CF (2010) Antagonistic factors control the unproductive splicing of SC35 terminal intron. *Nucleic Acids Res* **38**: 1353–1366
- Dreyfuss G, Matunis MJ, Pinol-Roma S, Burd CG (1993) hnRNP proteins and the biogenesis of mRNA. *Annu Rev Biochem* **62**: 289–321
- Gabut M, Mine M, Marsac C, Brivet M, Tazi J, Soret J (2005) The SR protein SC35 is responsible for aberrant splicing of the E1alpha pyruvate dehydrogenase mRNA in a case of mental retardation with lactic acidosis. *Mol Cell Biol* **25**: 3286–3294
- Grosso AR, Martins S, Carmo-Fonseca M (2008) The emerging role of splicing factors in cancer. *EMBO Rep* **9**: 1087–1093
- Hallay H, Locker N, Ayadi L, Ropers D, Guittet E, Branlant C (2006) Biochemical and NMR study on the competition between proteins SC35, SRp40, and heterogeneous nuclear ribonucleoprotein A1 at the HIV-1 Tat exon 2 splicing site. *J Biol Chem* **281**: 37159–37174
- 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
- Hargous Y, Hautbergue GM, Tintaru AM, Skrisovska L, Golovanov AP, Stevenin J, Lian LY, Wilson SA, Allain FH (2006) Molecular basis of RNA recognition and TAP binding by the SR proteins SRp20 and 9G8. *EMBO J* **25**: 5126–5137
- Hautbergue GM, Golovanov AP (2008) Increasing the sensitivity of cryoprobe protein NMR experiments by using the sole low-conductivity arginine glutamate salt. *J Magn Reson* **191**: 335–339
- Herrmann T, Guntert P, Wuthrich K (2002a) Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J Biomol NMR* **24**: 171–189
- Herrmann T, Guntert P, Wuthrich K (2002b) 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
- Hertel KJ, Graveley BR (2005) RS domains contact the pre-mRNA throughout spliceosome assembly. *Trends Biochem Sci* **30**: 115–118
- Huang Y, Steitz JA (2001) Splicing factors SRp20 and 9G8 promote the nucleocytoplasmic export of mRNA. *Mol Cell* **7**: 899–905
- Kohtz JD, Jamison SF, Will CL, Zuo P, Luhrmann R, Garcia-Blanco MA, Manley JL (1994) Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors. *Nature* **368**: 119–124

- Koradi R, Billeter M, Wuthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* **14**: 51–55, 29–32
- Laskowski RA, Rullmann JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* **8**: 477–486
- Lee W, Revington MJ, Arrowsmith C, Kay LE (1994) A pulsed field gradient isotope-filtered 3D ¹³C HMQC-NOESY experiment for extracting intermolecular NOE contacts in molecular complexes. *FEBS Lett* **350**: 87–90
- Lin S, Coutinho-Mansfield G, Wang D, Pandit S, Fu XD (2008) The splicing factor SC35 has an active role in transcriptional elongation. *Nat Struct Mol Biol* **15**: 819–826
- Liu HX, Chew SL, Cartegni L, Zhang MQ, Krainer AR (2000) Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol Cell Biol* **20**: 1063–1071
- Long JC, Caceres JF (2009) The SR protein family of splicing factors: master regulators of gene expression. *Biochem J* **417**: 15–27
- Oesterreich FC, Bieberstein N, Neugebauer KM (2011) Pause locally, splice globally. *Trends Cell Biol* **21**: 328–335
- Okunola HL, Krainer AR (2009) Cooperative-binding and splicing-repressive properties of hnRNP A1. *Mol Cell Biol* **29**: 5620–5631
- 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
- Qian W, Liang H, Shi J, Jin N, Grundke-Iqbal I, Iqbal K, Gong CX, Liu F (2011) Regulation of the alternative splicing of tau exon 10 by SC35 and Dyrk1A. *Nucleic Acids Res* **39**: 6161–6171
- Rossi F, Labourier E, Forne T, Divita G, Derancourt J, Riou JF, Antoine E, Cathala G, Brunel C, Tazi J (1996) Specific phosphorylation of SR proteins by mammalian DNA topoisomerase I. *Nature* **381**: 80–82
- Saliou JM, Bourgeois CF, Ayadi-Ben Mena L, Ropers D, Jacquenet S, Marchand V, Stevenin J, Branlant C (2009) Role of RNA structure and protein factors in the control of HIV-1 splicing. *Front Biosci* **14**: 2714–2729
- Sanford JR, Gray NK, Beckmann K, Caceres JF (2004) A novel role for shuttling SR proteins in mRNA translation. *Genes Dev* **18**: 755–768
- Sattler M, Schleucher J, Griesinger C (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog NMR Spectrosc* **34**: 93–158
- Schaal TD, Maniatis T (1999a) Multiple distinct splicing enhancers in the protein-coding sequences of a constitutively spliced pre-mRNA. *Mol Cell Biol* **19**: 261–273
- Schaal TD, Maniatis T (1999b) Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. *Mol Cell Biol* **19**: 1705–1719
- Shen H, Kan JL, Green MR (2004) Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. *Mol Cell* **13**: 367–376
- Smith PJ, Zhang C, Wang J, Chew SL, Zhang MQ, Krainer AR (2006) An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. *Hum Mol Genet* **15**: 2490–2508
- Solis AS, Peng R, Crawford JB, Phillips III JA, Patton JG (2008a) Growth hormone deficiency and splicing fidelity: two serine/arginine-rich proteins, ASF/SF2 and SC35, act antagonistically. *J Biol Chem* **283**: 23619–23626
- Solis AS, Shariat N, Patton JG (2008b) Splicing fidelity, enhancers, and disease. *Front Biosci* **13**: 1926–1942
- Tacke R, Manley JL (1995) The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. *EMBO J* **14**: 3540–3551
- Tiwari A, Panigrahi SK (2007) HBAT: a complete package for analysing strong and weak hydrogen bonds in macromolecular crystal structures. *In Silico Biol* **7**: 651–661
- Tsuda K, Someya T, Kuwasako K, Takahashi M, He F, Unzai S, Inoue M, Harada T, Watanabe S, Terada T, Kobayashi N, Shirouzu M, Kigawa T, Tanaka A, Sugano S, Guntert P, Yokoyama S, Muto Y (2011) Structural basis for the dual RNA-recognition modes of human Tra2-beta RRM. *Nucleic Acids Res* **39**: 1538–1553
- Venables JP, Koh CS, Froehlich U, Lapointe E, Couture S, Inkel L, Bramard A, Paquet ER, Watier V, Durand M, Lucier JF, Gervais-Bird J, Tremblay K, Prinos P, Klinck R, Elela SA, Chabot B (2008) Multiple and specific mRNA processing targets for the major human hnRNP proteins. *Mol Cell Biol* **28**: 6033–6043
- Wahl MC, Will CL, Luhrmann R (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**: 701–718
- Wang GS, Cooper TA (2007) Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8**: 749–761
- Wang HY, Lin W, Dyck JA, Yeakley JM, Songyang Z, Cantley LC, Fu XD (1998) SRPK2: a differentially expressed SR protein-specific kinase involved in mediating the interaction and localization of pre-mRNA splicing factors in mammalian cells. *J Cell Biol* **140**: 737–750
- Wang JM, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem* **21**: 1049–1074
- Wang X, Tanaka Hall TM (2001) Structural basis for recognition of AU-rich element RNA by the HuD protein. *Nat Struct Biol* **8**: 141–145
- Wang Z, Burge CB (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**: 802–813
- Wentz MP, Moore BE, Cloyd MW, Berget SM, Donehower LA (1997) A naturally arising mutation of a potential silencer of exon splicing in human immunodeficiency virus type 1 induces dominant aberrant splicing and arrests virus production. *J Virol* **71**: 8542–8551
- Wu JY, Maniatis T (1993) Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* **75**: 1061–1070
- Zahler AM, Damgaard CK, Kjems J, Caputi M (2004) SC35 and heterogeneous nuclear ribonucleoprotein A/B proteins bind to a juxtaposed exonic splicing enhancer/exonic splicing silencer element to regulate HIV-1 tat exon 2 splicing. *J Biol Chem* **279**: 10077–10084
- Zhu J, Mayeda A, Krainer AR (2001) Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. *Mol Cell* **8**: 1351–1361