MASARYK UNIVERSITY
Faculty of Science
National Centre for Biomolecular Research

Different modes of RNA recognition by proteins involved in RNA processing

Habilitation thesis by Richard Stefl

Brno, 2012

Acknowledgements

I would like to thank all my past supervisors and mentors who contributed to my scientific education. In particular, I thank Profs. Jaroslav Koca, Vladimir Sklenar, and Juli Feigon who supervised me during my undergraduate and graduate studies, for their support, trust, and patience. I wish to thank my postdoc supervisor Prof. Frederic Allain for his support and all the stimulating discussions we had. I would like to thank all my colleagues for their creativity, excellent work, and inspiring discussions.

Contents

1. Preface

This habilitation work is a compilation of scientific publications to which I have contributed as the corresponding author, the first author, and a co-author. These articles were published between 2005 and 2012. A list of these publications is given on page 42. All these publications have a common theme related to the structural insights into protein-RNA interactions and co-transcriptional processing. In particular, they are focused on the structural and mechanistic understanding of various protein-RNA recognition modes involved in RNA editing, transcription termination, processing of non-coding RNA, RNA quality control, and mRNA stability. The accompanying text highlights the author's contribution to the field of protein-RNA interactions and also contains a brief introduction to the topic. A comprehensive information on the studied protein-RNA interactions can be found in the enclosed original publication. The enclosed publications also include three review articles.

2. Introduction

Ribonucleic acids (RNAs) not only carry the genetic information that directs the synthesis of proteins, but they also play a central role in the regulation of gene expression. In the cell, RNAs are always accompanied by RNA-binding proteins (Dreyfuss et al., 2002). The association of RNA-binding proteins with RNA transcripts starts already during transcription. A number of these early-binding RNA-binding proteins remain bound to the RNA transcript until it is degraded, whereas other RNA-binding proteins recognize and transiently bind to the RNA at later stages for specific processes such as splicing, processing, transport and localization (Dreyfuss et al., 2002; Maniatis and Reed, 2002; Moore, 2005; Stefl et al., 2005a). The RNA-binding proteins cover the RNA transcript which facilitates its protection from rapid ribonucleolytic degradation (Vanacova and Stefl, 2007). Furthermore, posttranscriptional control of gene expression by RNA interference and microRNAs involves an array of RNA-binding protein interactions with microRNA and siRNA molecules as well as their cellular targets (Jinek and Doudna, 2009). The recently emerged concept of pervasive and/or hidden transcription revealed a vast number non-coding RNAs that are transcribed in eukaryotes (Jacquier, 2009; Johnson et al., 2005) and these RNAs play a role in various aspects of gene regulation (Esteller, 2011). Importantly, mutations that disrupt the function of RNA-binding proteins are linked to various diseases, including cancer (Cooper et al., 2009; Lukong et al., 2008).

The cell contains large ribonucleoprotein particles (RNPs) that operate as molecular machines and are, for example, involved in protein synthesis (ribosome), RNA transcription (RNA polymerases), or splicing machinery (spliceosome) (Cramer et al., 2008; Klinge et al., 2012; Wahl et al., 2009). In these large RNPs, the majority of proteins and RNAs are involved in tight interactions. In contrast, many multidomain RNA-binding proteins involved in regulatory functions, that they carry out by binding to the RNA substrates during cotranscriptional and posttranscriptional processing, involves flexibility and cooperative interactions with the RNA (Mackereth and Sattler, 2012). Interestingly, there are only a few RNA-binding domains and the variety of functions of RNA-binding proteins originates from the presence of multiple copies and different arrangements of these RNA-binding domains. The RNA-binding domains include the RNA-recognition motif (RRM), K-homology (KH) domain, the double-stranded RNA-binding motif (dsRBM), zinc-finger and zinc-knuckle domains, PAZ and P-element induced wimpy testis (PIWI) domains, sterile alpha motif (SAM) domain, S1 domain, oligonucleotide/oligosaccharide-binding (OB) fold, and cold shock (CSD) domain (Auweter et al., 2006b; Clery et al., 2008; Font and Mackay, 2010; Lunde et al., 2007; Mackereth and Sattler, 2012; Maris et al., 2005; Messias and Sattler, 2004; Stefl et al., 2005a).

The above RNA-binding domains are often tandemly arranged and tethered by flexible or structured linkers. The advantage of such a modular architecture arises from the resulting versatility. The multidomain RNA-binding proteins can bind RNA with higher specificity and affinity when compared to individual RNA-binding domains, which can bind only short RNA stretches with weak affinities. The concept of modularity also enables that RNA-binding proteins can bind RNA with poorly conserved sequence features, which are often observed, for example, in 3'-end processing and splicing sites (Mackereth et al., 2011; Maris et al., 2005; Sickmier et al., 2006).

Some RNA-binding proteins also contain protein-protein interaction modules or catalytic domain in addition to the RNA-binding domains. In the case of proteins with enzymatic domains, the RNA-binding domains define the specificity of the enzymes by recruiting them to their substrates. Adenosine deaminase that act on RNA 2 (ADAR2), an enzyme that recodes genomic information by the site-selective deamination of adenosine, is an example of such an enzyme, with two dsRBMs and a catalytic deaminase domain (Bass, 2002; Macbeth et al., 2005; Stefl et al., 2010). The modules that mediate protein-protein interactions often recruit the RNA-binding proteins to the site of action, for example to the site of transcription for cotranscriptional processing (Jasnovidova and Stefl, 2012). A number of typical RNA-binding domains, such as RRMs, KH, and dsRBMs, were also shown to mediate protein-protein interactions (Lunde et al., 2007).

The enormous diversity of interactions observed in the structures of protein–RNA complexes suggests that a simple recognition code is unlikely to exist in the world of protein–RNA interactions. However, several unifying themes may be inferred from the known complexes and accompanied mechanistic studies. Some RNA-binding proteins bind single-stranded or double-stranded RNA by direct readout of the primary sequence, whereas others recognize primarily the shape of the RNA, or both the sequence and the shape (Clery et al., 2008; Maris et al., 2005; Masliah et al., 2012; Stefl et al., 2005a). The formation of protein-RNA complexes can involve either an induced fit mechanism or rigid-body docking mechanism. In some cases, the protein-RNA recognition occurs via a conformational selection mechanism provided that apo forms of RNA-binding protein or RNA substrate exist in multiple conformations. This implies that conformational dynamics plays a fundamental role in RNA recognition as it enables cooperativity and combinatorial binding by individual RNA-binding proteins (Mackereth et al., 2011; Mackereth and Sattler, 2012; Stefl et al., 2005a).

In our work over the recent years, we have contributed to better understanding of various protein-RNA recognition modes. In particular, we have studied how RRMs, dsRBMs, SAM, and other RNA-binding domains recognize their RNA targets that are involved in RNA editing, transcription termination, processing of non-coding RNA, RNA quality control, and mRNA stability. For example, we demonstrated how RNA-binding domains bind RNA in a cooperative manner, which increased the binding affinity by several orders of magnitude when compared to the binding affinity of isolated domains. We also revealed an unexpected mode of the sequence-specific RNA recognition by the two double-stranded RNA recognition motifs (dsRBMs) present in the N-terminal region of ADAR2 (A-to-I editing enzyme). The solution structure of the two dsRBMs of ADAR2 bound to a specific editing substrate revealed the general principle of the sequence-specific recognition by ADAR2 dsRBMs which explains how ADAR2 selects its editing sites. The following accompanying text highlights the author's contribution in the general context of protein-RNA interactions.

3. NMR as a useful tool for the protein-RNA interaction studies

NMR and X-ray crystallography, both can provide structural snapshots revealing the key aspects of intermolecular interactions and domain organization in the protein-RNA complexes. With the knowledge of the initial structures of unbound components, protein and RNA, it is possible to suggest potential binding pathways and mechanisms. However, these static structures may not address the mechanism by which, for example, the multiple RNA-binding domains functionally cooperate (Mackereth and Sattler, 2012). The dynamical processes of the assembly and disassembly of protein-RNA complexes are dictated by the thermodynamics and the kinetics of their interactions, which in turn, depends on the conformational landscape of interacting molecules and on the entropy of both bound and free states. The inherent mobility of protein-RNA complexes can be studied by NMR but not by X-ray crystallography, which provides only a static picture. NMR *per se* is well suited for characterization of the solution structure and conformational dynamics of smaller systems. In conventional NMR, the analysis of larger systems suffers from spectral overlap and signal-to-noise losses. Only recently, the development of new isotope-labeling strategies and NMR spectroscopy techniques provides tools to overcome these obstacles (Dominguez et al., 2011; Duss et al., 2012; Tugarinov et al., 2004; Tugarinov et al., 2006). In dynamic complexes, the difficulty in getting NOE-derived distance constraints has been addressed by the means of paramagnetic relaxation enhancement (PRE) and residual dipolar coupling (RDC) data. The combination of PRE and RDC provides information about relative domain orientations and enable to determine the structures of large complexes in solution (Mackereth and Sattler, 2012; Simon et al., 2010). In addition, solvent PRE using bulk spin label is a simple and powerful tool to probe binding interfaces (Mackereth and Sattler, 2012; Madl et al., 2011a; Madl et al., 2011b). NMR can also be combined with small angle scattering (SAS), either by X-ray (SAXS) or by neutrons (SANS), yielding a powerful approach in which the high-resolution NMR structures of individual subunits can be accommodated in the SAS-derived molecular envelope of the entire studied system (Grishaev et al., 2005; Madl et al., 2011a; Wang et al., 2010). In the context of the state-of-the-art NMR approaches, we determined the solution structure of the two dsRBMs of ADAR2 bound to a stem-loop pre-mRNA encoding the R/G editing site of GluR-2 (Stefl et al., 2010) (Figure 1). The structure of this more than 50 kDa complex was solved using the divide-and-conquer approach in which atomic details of binding interfaces were obtained with two subcomplexes which were assembled using RDC data to visualize the structure of the entire complex. The structure provides a molecular basis for how dsRBMs recognize the shape, and also more surprisingly, the sequence of the dsRNA. The unexpected direct readout of the RNA primary sequence by dsRBMs is achieved via the minor groove of the dsRNA and this recognition is critical for both editing and binding affinity at the R/G site of GluR-2.

More generally, our findings suggest a solution to the sequence-specific paradox faced by many dsRBM-containing proteins that are involved in post-transcriptional regulation of gene expression (Stefl et al., 2010).

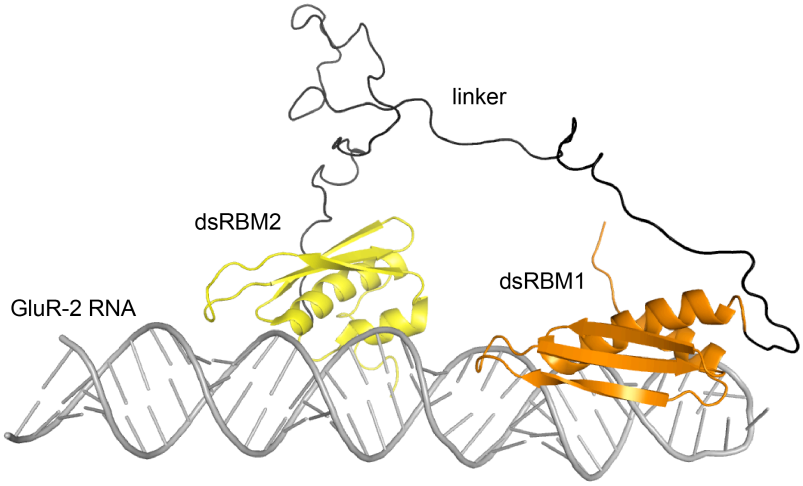


Figure 1. Structure of ADAR2 dsRBM12 Bound to GluR-2 R/G (Stefl et al., 2010). RDC-reconstructed structure of the ADAR2 dsRBM12 bound to GluR-2 R/G. The RNA is represented as a stick model (colored in gray) and the protein is shown as a ribbon model (dsRBM1 in orange; dsRBM2 in yellow; linker in black). Figure was generated with pymol (The PyMOL Molecular Graphics System, Schrödinger, LLC.).

NMR is very useful technique for the structure determination of weak protein-RNA complexes ($K_D \geq 10^{-4}$ M at a physiological salt concentration) (Dominguez et al., 2011). The complexes in this range of affinities are normally undetectable by most biochemical methods. Similarly, X-ray crystallography is very ineffective when tackling weak protein-RNA complexes because they are difficult to crystallize. We were successful in solving of one of these weak protein-RNA complexes that is involved in the poly(A)-independent termination pathway that requires the Nrd1 complex (Hobor et al., 2011). The Nrd1 complex includes two RNA-binding proteins, the nuclear polyadenylated RNA-binding (Nab) 3 and the nuclear pre-mRNA down-regulation (Nrd) 1 that bind their specific termination elements. We determined the solution structure of the RNA-recognition motif (RRM) of Nab3 in complex with a UCUU oligonucleotide, representing the Nab3 termination element. Albeit the binding affinity is very weak (K_D ~ 800 μM), the structure shows that the first three nucleotides of UCUU are accommodated on the β-sheet surface of Nab3 RRM, and reveals a sequence-specific recognition for the YCU sequence consensus. The specific contacts we identified to be important for binding affinity *in vitro* as well as for yeast viability (Hobor et al., 2011).

The following articles are related to the above topic:

Stefl R, Oberstrass FC, Hood JL, Jourdan M, Zimmermann M, Skrisovska L, Maris C, Peng L, Hofr C, Emeson RB Allain FH. (2010) The solution structure of the ADAR2 dsRBM-RNA complex reveals a sequence-specific read out of the minor groove. *Cell* 143, 225-237. Author contributions (Stefl R): 50%, designed and performed experiments, analyzed the data and wrote the paper. IF=32.406

Hobor F, Pergoli R, Kubicek K, Hrossova D, Bacikova V, Zimmermann M, Pasulka J, Hofr C, Vanacova S, Stefl R. (2011) Recognition of transcription termination signal by the nuclear polyadenylated RNA-binding (Nab)3 protein. *J. Biol. Chem.* 286, 3645-3657. Author contributions (Stefl R): 25%, designed and performed experiments, analyzed the data and wrote the paper. IF=4.773

4. RNA recognition by RNA-binding domains

RNA recognition by RNA-binding proteins is mediated by a special class of domains, called RNA-binding domains. RNA-binding domains are rather small and compact, and the proteins can contain from one to many copies of these domains that are usually tandemly arranged (Lunde et al., 2007). These domains play important role in all aspects of gene expression and regulation. Molecular details and principles of RNA recognition by RNA-binding domains have been revealed from high-resolution structures of individual RNA-binding domains bound to their RNA targets (Lunde et al., 2007; Mackereth and Sattler, 2012; Masliah et al., 2012; Stefl et al., 2005a). The basic RNA-binding domains implicated in RNA recognition are described below.

4.1 RNA recognition by RRMs

The RNA-recognition motif (RRM) is the most frequent RNA-binding domain. It is a small globular domain of about 80 amino acids with a typical **βαββαβ** topology that is formed by a four-stranded **β**-sheet packed against two **α**-helices (Mattaj, 1993). RRMs are often present in multiple copies and the RRM-domain-containing proteins are involved in many cellular functions, particularly messenger RNA and ribosomal RNA processing, splicing and translation regulation, RNA export and RNA stability (Auweter et al., 2006b; Dreyfuss et al., 2002; Maris et al., 2005; Stefl et al., 2005a). The structures of RRM in complex with RNA determined using either NMR spectroscopy or X-ray crystallography reveal the complexity of protein–RNA recognition mediated by the RRM. In many cases, the binding mechanism involves not only protein–RNA interactions but also RNA–RNA and protein–protein interactions (Auweter et al., 2006b; Stefl et al., 2005a). The main protein surface of the RRM involved in the interaction with the RNA is the four-stranded **β**-sheet, which usually contacts two or three nucleotides. The **β**-sheet mediates the canonical mode of RRM–RNA interaction and it is demonstrated here by the structure of Nab3 RRM bound to UCUU (Figure 2) (Hobor et al., 2011). The four-stranded **β**-sheet (in yellow) binds the first three nucleotides out of 5'-$U_1C_2U_3U_4$-3' RNA. The three nucleotides are located on the surface of the **β**-sheet, with the bases oriented parallel to the **β**-sheet plane and the bases of C_2 and U_3 are packed against the conserved phenylalanine side-chains (F333 and F368). Only two nucleotides (C_2U_3) are recognized sequence-specifically by interactions with the protein side- and main-chains of the **β**-sheet. The recognition of U_1 is less evident from the structure. The contact from Glu397 allows also for the interaction with a cytidine in this position. Overall, Nab3 RRM recognizes YCU sequence (where Y stands for pyrimidine).

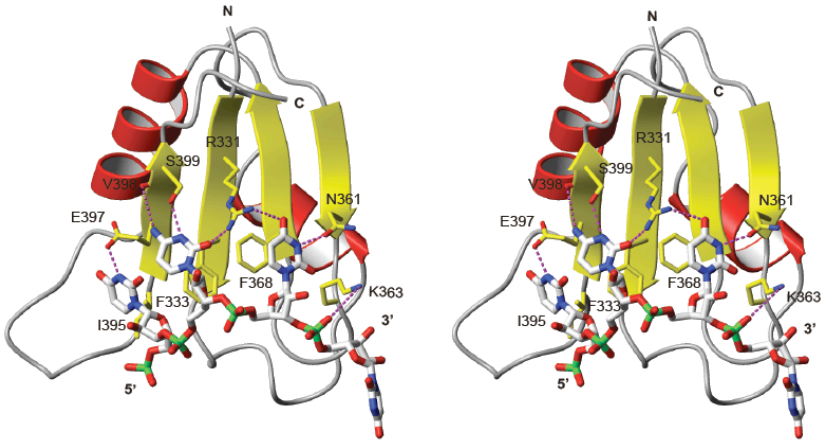


Figure 2. Stereo view of the structure of the Nab3 RRM–UCUU complex. The RNA is represented as a white stick model and the protein is shown as a ribbon model with residues that contact the RNA shown in yellow. Putative hydrogen bonds are shown by dotted magenta lines. Figure was generated with MOLMOL (Koradi et al., 1996).

In many cases, RRM-containing proteins bind more than three nucleotides and recognize longer single-stranded RNA. This is achieved due to the presence of the N- or C-terminal extensions of the domain, to the interdomain linker in case of proteins containing multiple RRMs, or to additional protein cofactors that can also modulate the RNA-binding specificity (Clery et al., 2008; Maris et al., 2005; Muto and Yokoyama, 2012; Stefl et al., 2005a). For example, the C-terminus of hnRNP A1 RRM 1, that is unstructured in the free form, becomes ordered upon nucleic acid, forming a 3_{10} helix. Such induced fit binding mechanism positions specific side-chain residues of the helix, His101 and Arg92, to stack over A203 and G204, respectively (Ding et al., 1999). In some multidomain RBPs, two RRMs and the interdomain linker cooperatively bind RNA providing high affinity and specificity (Allain et al., 2000b; Crowder et al., 1999) (Allain et al., 2000a; Deo et al., 1999; Ding et al., 1999; Handa et al., 1999; Johansson et al., 2004; Price et al., 1998). In the RRM of CBP20, the C- and N-terminal extensions (which are stabilized by the cognate protein CBP80) provide a tight binding pocket for the 5′ capped RNAs (7-methyl-G(5′)ppp(5′)N, where N is any nucleotide) (Mazza et al., 2002).

Not only the β-sheet of RRMs but also the loops connecting β-strands and α-helices are important for RNA recognition. Either one loop (for RMBY (Skrisovska et al., 2007), SF2/ASF (Tintaru et al., 2007), Hrp1 (Perez-Canadillas, 2006)) or more loops (Fox-1 (Auweter et al., 2006a), REF2-I (Golovanov et al., 2006), hnRNP F (Dominguez and Allain, 2006; Dominguez et al., 2010)) were found to form extensive contacts with RNA. In the RRM of human RBMY, the β2-β3 loop is essential for the recognition of the shape of the RNA stem, as all the β2-β3 loop residues contact the sugar-phosphate backbone of the RNA. In the RMBY RRM–RNA complex, the β-sheet surface binds to

the RNA loop in a sequence-specific fashion, combining thus both RNA sequence- and shape-dependent recognition mechanisms (Skrisovska et al., 2007).

In general, the structures of RRM-RNA complexes show that the RRM is a platform with a large capacity for variation in order to achieve high RNA-binding affinity and specificity. For example, it is remarkable that a single domain like nucleolin RRM2 contacts only two nucleotides, whereas U1A RRM1 contacts 12 nucleotides and the RRM of Y14 (Fribourg et al., 2003) does not contact RNA but rather another protein. This fascinating plasticity of the RRM explains why it is so abundant and why it is involved in so many different biological functions; however, this plasticity makes it difficult to predict how the RRM achieves RNA recognition. We have previously summarized and discussed the plasticity of RRM in the following review article (Stefl et al., 2005a).

The following articles are related to the above topic:

Hobor F, Pergoli R, Kubicek K, Hrossova D, Bacikova V, Zimmermann M, Pasulka J, Hofr C, Vanacova S, Stefl R. (2011) Recognition of transcription termination signal by the nuclear polyadenylated RNA-binding (Nab)3 protein. *J. Biol. Chem.* 286, 3645-3657. Author contributions (Stefl R): 25%, designed and performed experiments, analyzed the data and wrote the paper. IF=4.773

Porrua O, Hobor F, Boulay J, Kubicek K, D'Aubenton-Carafa Y, Gudipati RK, Stefl R, Libri D. (2012) In vivo SELEX reveals novel sequence and structural determinants of Nrd1-Nab3-Sen1-dependent transcription termination. *EMBO J.* 31, 3935-3948. Author contributions (Stefl R): 10%, designed experiments and wrote the paper. IF=9.205

Pergoli R, Kubicek K, Hobor F, Pasulka J, Stefl R. (2010) 1H, 13C, and 15N chemical shift assignments for the RNA recognition motif of Nab3. *Biomol. NMR Assign.* 4, 119-121. Author contributions (Stefl R): 15%, designed and performed experiments, and wrote the paper. IF=0.720

Skrisovska L, Bourgeois CF, Stefl R, Grellscheid SN, Kister L, Wenter P, Elliott DJ, Stevenin J, Allain FH. (2007) The testis-specific human protein RBMY recognizes RNA through a novel mode of interaction. *EMBO Rep.* 8, 372-379. Author contributions (Stefl R): 10%, analyzed the data and performed structure calculation. IF=7.355

Stefl R, Skrisovska L, Allain, FH. (2005) RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. *EMBO Rep.* 6:33-38. Author contributions (Stefl R): 33%, analyzed the data and wrote the paper. IF=7.355

4.2 RNA recognition by dsRBMs

The double-stranded RNA-binding motif (dsRBM) is a 70 amino-acid domain with a conserved **αββββα** protein topology in which the two **α**-helices are packed along one face of a three-stranded anti-parallel **β**-sheet (Figure 3). The domain is well conserved and only some minor variations are found, such as length of the loops, additional C-terminal **α**-helix, or the C-terminal extension composed of a short **α**-helical turn followed by a zinc-coordination site (Masliah et al., 2012). Akin to other RBDs, these domains occur mostly in multiple copies and have been found in eukaryotic, prokaryotic and even viral proteins. These proteins have an essential role in RNA interference, RNA processing, RNA localization, RNA editing and translational repression (Doyle and Jantsch, 2002; Masliah et al., 2012; Saunders and Barber, 2003).

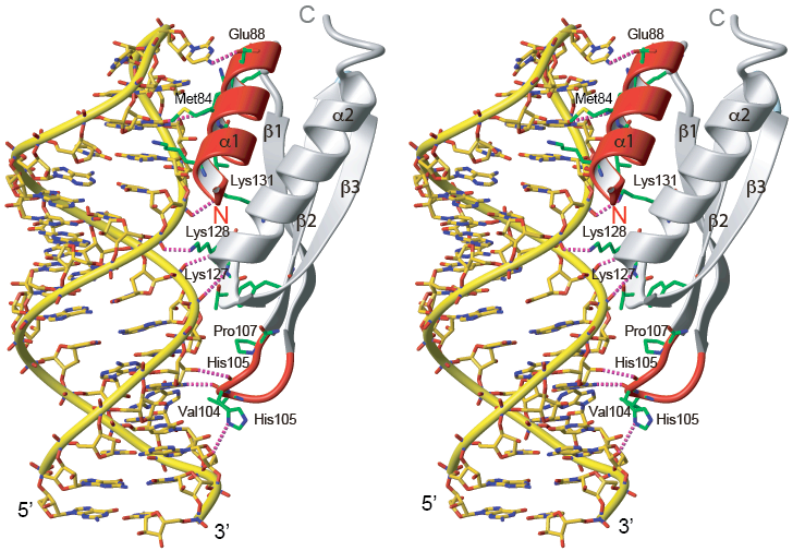


Figure 3. RNA recognition by ADAR2 dsRBM1 (Stefl et al., 2010). Stereo view of the most representative structure of the dsRBM1 bound to USL RNA. The RNA is represented as a yellow stick model and the protein is shown as a ribbon model with residues that contact the RNA shown in green. Helix α1 and the β1-β2 loop that mediate the sequence-specific contacts are colored in red. Hydrogen-bonds are indicated by magenta dotted lines. Figure was generated with MOLMOL (Koradi et al., 1996).

A number of structures of dsRBMs in complex with dsRNA that have been determined using X-ray crystallography or NMR spectroscopy. For example, a 1.9 Å crystal structure of the second dsRBM of Xenopus laevis RNA-binding protein A (Xlrbpa2) bound to two coaxially stacked dsRNA molecules (Ryter and Schultz, 1998), was determined by X-ray crystallography. NMR was used to determine the structure of the third dsRBM from the Drosophila Staufen protein in complex with a symmetrical GC-rich 12-bp duplex capped by a UUCG tetraloop (Ramos et al., 2000), as well as the structure of the dsRBM of Rnt1p (an RNase III homologue from budding yeast) bound to a 14-bp RNA duplex capped by an AGAA tetraloop (Wu et al., 2004). We used NMR to determine the solution structure of the two dsRBMs of ADAR2 bound to a stem-loop

pre-mRNA encoding the R/G editing site of GluR-2 (Stefl et al., 2010; Stefl et al., 2006). These structures provide a precise description of how the dsRBM recognize specifically the A-form helix conformation adopted by double-stranded RNA (dsRNA) (Masliah et al., 2012; Stefl et al., 2005a). All the dsRBM-RNA complexes have the following common features. The dsRBMs interact along one face of the dsRNA through both α-helices and their β1–β2 loop (Figure 3). The contacts with the RNA involve two consecutive minor grooves separated by a major groove. The major groove of the A-form duplex with its specific width is probed by the KKxAK motif of the N-terminal tip of α-helix 2. The α-helix 1 and β1–β2 loop mostly contact the two consecutive minor grooves using non-sequence-specific interactions. These interactions involve 2′-hydroxyls and phosphate oxygens and are perfectly adapted to the shape of an RNA double helix. However, there are some variations. For example, the interactions mediated by the α-helix 1 vary in different dsRBM complexes. In the Xlrbpa2 dsRBM–dsRNA complex, the α-helix 1 interacts nonspecifically with the minor groove of the RNA, with a few contacts to the bases. In the dsRBM of Staufen, the α-helix 1 interacts with a UUCG tetraloop that caps the RNA double helix. Although the UUCG tetraloop is not a natural substrate of Staufen, this finding led to the proposal that the α-helix 1 modulates the specificity of individual dsRBMs (Ramos et al, 2000). Indeed, this was recently confirmed by the structure of the dsRBM of Rnt1p bound to its natural RNA substrate (Fig 1B), in which the α-helix 1 recognizes the specific shape of the minor groove created by the conserved AGNN tetraloop (Wu et al, 2004). The α-helix 1 conformation that is stabilized by an additional carboxy-terminal α-helix 3, is tightly inserted into the RNA minor groove and contacts the sugar-phosphate backbone and the two non-conserved tetraloop bases, whereas the conserved A and G bases are not involved in the interactions (Wu et al, 2004). This structure illustrates how this dsRBM recognizes the specific shape of its RNA target but not its sequence. dsRBMs are highly conserved and have the same structural framework, but are chemically distinct through variations in key residues. The structure of the dsRBM of Rnt1p in complex with RNA highlights the essential role of the α-helix 1 in the recognition of structured elements that deviate from regular dsRNA. Interestingly, the α-helix 1 is the least-conserved secondary structure element among various dsRBMs and seems to have a different spatial arrangement relative to the rest of the domain in different dsRBMs. This variability may be an important element as many biochemical experiments have shown that dsRBM-containing proteins possess binding specificity for a variety of RNA structures, such as stem–loops, internal loops, bulges or helices with mismatches (Doyle & Jantsch, 2002; Fierro-Monti & Mathews, 2000; Ohman et al, 2000; Stephens et al, 2004).

Indeed, the length and different spatial arrangement of the α-helix 1 in dsRBMs of ADAR2 play an essential role for the binding of these domains at a very specific register on the large GluR-2 R/G RNA molecules (Figure 1) (Stefl et al., 2010; Stefl et al., 2005b; Stefl et al., 2006). In this complex, the specific binding is achieved by a direct readout of the RNA sequence in the minor groove of the A-form double helix. Both dsRBMs of ADAR2 use the α-helix 1 and the β1-β2 loop as molecular rulers to find their binding register in the RNA minor groove of the GluR-2 R/G RNA (Stefl et al., 2010). Using the α-helix 1, the side-chain methyl groups of Met84 in dsRBM1 and of Met238 in dsRBM2 are in contact with the H2s of A32 and A18, respectively. Recognition of these two anchoring points in the minor groove, separated by 9 and 8 base-pairs for dsRBM1 and dsRBM2, respectively, illustrates how the two dsRBMs find their sequence-specific binding registers. Quite interestingly, in each complex, one of the two anchoring points involves a mismatched base-pair (the G22-G50 base-pair for dsRBM1 and the A18-C54 base-pair for dsRBM2). It is therefore likely that the highly exposed amino or C2H2 groups of these mismatches in the minor groove further assist the dsRBMs of ADAR2 to find their binding register. These findings suggest a solution to the sequence-specific paradox faced by many dsRBM-containing proteins that are involved in post-transcriptional regulation of gene expression.

The following articles are related to the above topic:

Stefl R, Oberstrass FC, Hood JL, Jourdan M, Zimmermann M, Skrisovska L, Maris C, Peng L, Hofr C, Emeson RB Allain FH (2010) The solution structure of the ADAR2 dsRBM-RNA complex reveals a sequence-specific read out of the minor groove. *Cell*, 143, 225-237. Author contributions (Stefl R): 50%, designed and performed experiments, analyzed the data and wrote the paper. IF=32.406

Stefl R, Xu M, Skrisovska L, Emeson RB, Allain FH. (2006) Structure and specific RNA binding of ADAR2 double-stranded RNA binding motifs. *Structure* 14, 345–355. Author contributions (Stefl R): 60%, designed and performed experiments, analyzed the data and wrote the paper. IF=6.347

Stefl R. and Allain, F.H.-T. (2005) A novel RNA pentaloop fold involved in targeting ADAR2. *RNA* 11, 592-597. Author contributions (Stefl R): 75%, designed and performed experiments, analyzed the data and wrote the paper. IF=5.095

Stefl R, Skrisovska L, Xu M, Emeson RB, Allain FH. (2005) Resonance assignments of the double-stranded RNA-binding domains of adenosine deaminase acting of RNA 2

(ADAR2). *J. Biomol. NMR* 31, 71-72. Author contributions (Stefl R): 80%, designed and performed experiments, analyzed the data and wrote the paper. IF=3.612

4.3 RNA recognition by zinc-binding domains

CCHH-type zinc-finger domains are the most common DNA-binding domain found in eukaryotic genomes. Usually, several fingers are used in a modular fashion to achieve high sequence-specific recognition of DNA (Miller et al., 1985). Each finger displays a ββα protein fold in which a β-hairpin and an α-helix are pinned together by a Zn^{2+} ion. DNA-sequence-specific recognition is achieved by the interactions between protein side-chains of the α-helix (at position –1, 2, 3 and 6, for the canonical arrangement) and the DNA bases in the major groove (Pabo and Nekludova, 2000; Wolfe et al., 2000). However, there is growing evidence that zinc fingers are also used to recognize RNA (Finerty and Bass, 1997; Mendez-Vidal et al., 2002; Picard and Wegnez, 1979; Theunissen et al., 1992). The crystal structure of three zinc fingers (fingers 4–6) of the transcription factor IIIA (TFIIIA) in complex with a 61-nucleotide fragment of the 5S RNA (Lu et al., 2003) provided the first insight into RNA recognition by CCHH-type zinc fingers. In this structure, finger 4 binds to loop E, finger 5 to helix V, and finger 6 to loop A (Figure 4). Finger 4 recognizes loop E by specifically interacting with a bulged guanosine and, similarly, finger 6 recognizes loop A by specifically interacting with two bases (an adenine and a cytosine) that also bulge out from the rest of the RNA. The specific recognition of the RNA by both fingers 4 and 6 is achieved by side-chain contacts from the N-terminal parts of the α-helix (at position –1, 1 and 2). The interaction of finger 5 with helix V differs from the ones made by fingers 4 and 6. In this case, finger 5 recognizes a short RNA double helix by multiple contacts between basic amino acids of the α-helix and the RNA sugar-phosphate backbone.

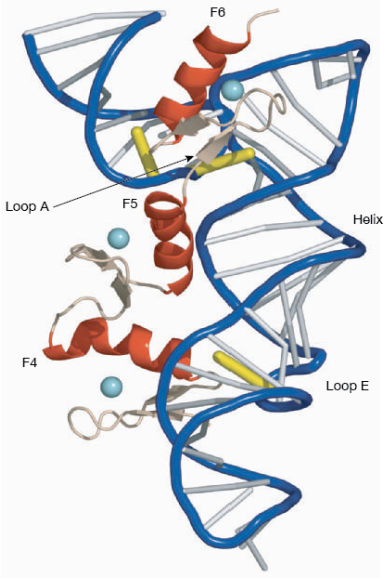


Figure 4. RNA recognition by CCHH-type zinc finger. Overall view of the complex of transcription factor IIIA (TFIIIA) fingers 4–6 (F4–F6) and 61-nucleotide 5S RNA (Lu et al., 2003). The protein and RNA are represented as ribbon models. The bulged bases involved in the recognitions are highlighted in yellow. Cyan balls represent zinc ions. TFIIIA finger 4 (F4) binds to loop E. The α-helix (in red) of the finger 4 specifically interacts with a guanosine base that bulges out (in yellow); the base contacts are made from the side-chain at position –1, 1 and 2 of the α-helix. TFIIIA finger 5 (F5) binds to helix V. The α-helix (in red) of finger 5 recognizes the dsRNA shape by non-sequence-specific contacts to the RNA sugar-phosphate backbone. Figure was generated with pymol (The PyMOL Molecular Graphics System, Schrödinger, LLC.).

In contrast to the above-mentioned CCHH zinc fingers, another class of zinc fingers (CCCH-type) was recently found to adopt a different fold and to recognize sequence-

specifically single-stranded RNA (Hudson et al., 2004). In this NMR structure, sequence-specific RNA recognition is achieved by a network of intermolecular hydrogen bonds between the protein main-chain functional groups and the Watson–Crick edges of the bases (Hudson et al., 2004). Another zinc-binding domain that is able to bind RNA are zinc knuckles. The CCHC-type zinc knuckles are found in eukaryotes and also in the nucleocapsid of retroviruses (Lu et al., 2011). For instance, in the structure of Lin28 in complex with *let-7* microRNA, the two zinc knuckles binds G-rich fragment (GGAG) of *let-7* RNA (Loughlin et al., 2012; Nam et al., 2011). The resulting conformation of the recognized single-stranded RNA resembles the conformation of "K-turn" that often participates in specific protein-RNA interactions (Nam et al., 2011).

These structures reveal that zinc fingers and zinc knuckles bind to RNA differently to the way they do to DNA. The CCHH-type zinc fingers have two modes of RNA binding. First, the zinc fingers interact non-specifically with the backbone of a double helix, and second, the zinc fingers specifically recognize individual bases that bulge out of a structurally rigid element. The CCCH-type zinc fingers and the CCHC-type zinc knuckles show a third mode of RNA binding, in which the single-stranded RNA is recognized in a sequence-specific manner. Zinc binding domains represent a unique class of nucleic-acid-binding proteins that are capable of a direct readout of the DNA sequence within a DNA double helix, a direct readout of the RNA sequence within single-stranded RNA, and an indirect readout of the RNA as they recognize the shape of the RNA rather than its sequence. Of course, more structures of zinc fingers and zinc knuckles in complex with RNA, in particular those with many tandemly arranged repeats, will need to be determined to generalize their mode of RNA recognition.

To study tandemly arranged zinc-binding domains, we set out to delineate the structure and function of Air2 that is an RNA-binding subunit of Trf4/5-Air1/2-Mtr4 polyadenylation complex (TRAMP), which is an essential component of nuclear RNA surveillance in yeast (LaCava et al., 2005; Vanacova and Stefl, 2007; Vanacova et al., 2005; Wyers et al., 2005). Air2 contains five tandemly arranged zinc knuckle motifs of the CCHC-type ($CX_2CX_4HX_4C$). As the preparation of multiple tandemly arranged zinc knuckle motifs is challenging task, we set out to develop a procedure for the preparation of Air2 in large amounts for structural studies. After many unsuccessful trials, we succeeded to prepare Air2 protein through a chaperone-assisted refolding procedure (Holub et al., 2012). With the recombinantly prepared protein we could show that Air2 is a genuine RNA-binding subunit of TRAMP. We identified the zinc knuckles 2, 3 and 4 as the RNA-binding domains, and revealed the essentiality of the zinc knuckle 4 for TRAMP4 polyadenylation activity. Furthermore, we identified Air2 as the

key component of TRAMP4 assembly providing bridging between Mtr4 and Trf4. The former is bound via the N-terminus of Air2, while the latter is bound via ZnK5, the linker between ZnK4 and 5 and the C-terminus of the protein (Holub et al., 2012). We also investigate the structure of ZCCHC9, which is a human nuclear protein with sequence homology to yeast Air2 protein. The ZCCHC9 protein contains four retroviral-type zinc knuckle motifs. So far, we completed the NMR spectral assignment of the zinc knuckle region of ZCCHC9. These data will allow performing NMR structural and RNA-binding studies of ZCCHC9 with the aim to investigate its role in the RNA quality control in mammalian cells (Sanudo et al., 2011).

The following articles are related to the above topic:

Holub P, Lalakova J, Cerna H, Pasulka J, Sarazova M, Hrazdilova K, Arce MS, Hobor F, Stefl R, Vanacova S. (2012). Air2p is critical for the assembly and RNA-binding of the TRAMP complex and the KOW domain of Mtr4p is crucial for exosome activation. *Nucleic Acids Res.* 40, 5679-5693. Author contributions (Stefl R): 10%, designed experiments and wrote the paper. IF=8.026

Sanudo M., Jacko M., Rammelt C., Vanacova S., Stefl R. (2010) 1H, 13C, and 15N chemical shift assignments of ZCCHC9. *Biomol NMR Assign.* 5, 19-21. Author contributions (Stefl R): 25%, designed and performed experiments, and wrote the paper. IF=0.720

Vanacova S, Stefl R. (2007) The exosome and RNA quality control in the nucleus. *EMBO Rep.* 8, 651-657. Author contributions (Stefl R): 50%, Analyzed data and wrote the paper. IF=7.355

4.4 RNA recognition by SAM domains

Sterile alpha motif (SAM) domain is one of most abundant protein domain found in eukaryotes. It is mostly involved in transcription regulation and signal-transduction cascades (Qiao and Bowie, 2005). For example, the SAM domain of protein TEL serves as a self-association domain that can either activate kinase catalytic domain or repress transcription (Jousset et al., 1997). A subset of SAM domains also bind nucleic acids. The RuvABC complex is involved in processing of the Holliday junction intermediate, with the SAM domain of RuvA binding the B-form of double helical DNA. Furthermore, some SAM domains have been shown to bind RNA (Aviv et al., 2003; Dilcher et al., 2001; Green et al., 2003). For example, the SAM domain of *S. cerevisiae* Vts1 bind specifically an RNA stem-loop with high affinity (Aviv et al., 2003). The recognized RNA, Smaug response element (SRE), contains a pentaloop with a CNGGN consensus (where N is can be any nucleotide).

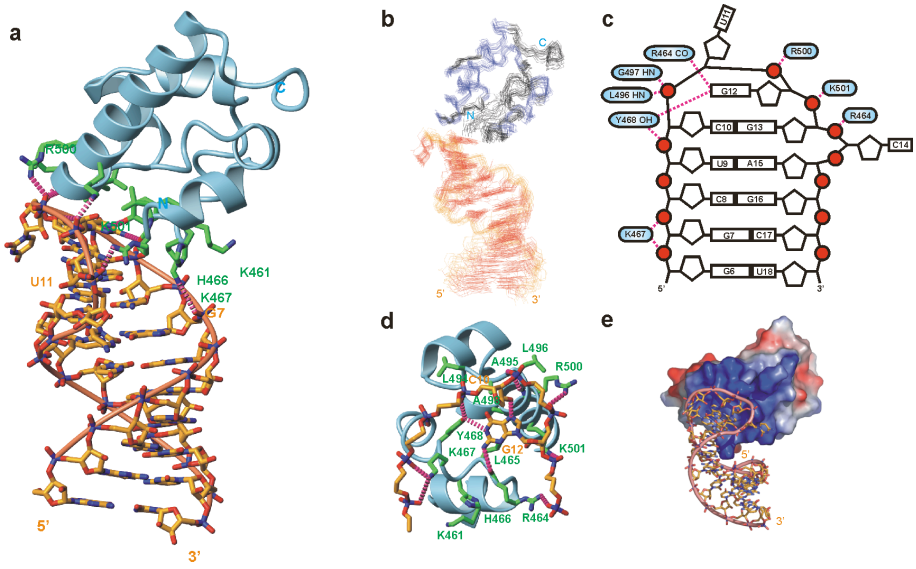


Figure 5. RNA recognition by Vts1 SAM domain (Oberstrass et al., 2006). (a) The solution structure of Vts1 SAM domain bound to SRE RNA complex. The side-chains important for recognition are shown in green. The dashed lines in magenta indicate possible hydrogen bonds. (b) The ensemble of the 20 lowest-energy structures calculated using NMR restraints. (c) Schematic representation of the intermolecular interactions. (d) Close-up showing all the interactions important for recognition. (e) Electrostatic surface representation of the protein in the complex (blue, positive; red, negative). RNA is shown in stick representation. Figure was generated with MOLMOL (Koradi et al., 1996).

To understand the structure and function of SAM-RNA recognition, we determined the solution structure of the SAM domain of Vts1 in complex with the Smaug response element (SRE) stem-loop RNA (Figure 5). We have also determined the structure of both protein and RNA in isolation to better understand the recognition mechanism. Interestingly, the SAM domain of Vts1 and the SRE RNA adopt virtually identical structures in the complex and in their free forms. This indicates that the binding

mechanism occurs via rigid body docking. The SAM domain of Vts1 binds the apical pentaloop, contacting the major groove surface of the RNA. The only base that is specifically recognized is G12, situated on top of the apical RNA loop. This guanine base occupies a small hydrophobic cavity and its Watson-Crick edge is specifically recognized by the SAM domain. In the Vts1 SAM–SRE RNA complex, the majority of protein-RNA contacts involve the RNA sugar-phosphate backbone of the apical loop. As the conformation of the RNA loop is unchanged upon complex formation, the Vts1 SAM–SRE RNA recognition is rather shape-specific than sequence-specific, with the exception of G12 recognition. Mechanistically, the shape of the CUGGC pentaloop creates a unique electrostatic fingerprint that is recognized by the basic surface of Vts1 SAM domain (Figure 5) (Oberstrass et al., 2006). We confirmed the functional importance of the Vts1 SAM–SRE RNA interaction using microarray gene profiling in yeast, as we identified a number of genes in *S. cerevisiae* that are strongly upregulated in the *vts1Δ* strain. Characterization of these genes by bioinformatics showed that many of them contain one or more copies of the SRE RNA. Interestingly, virtually identical results were obtained in parallel by others (Aviv et al., 2006; Johnson and Donaldson, 2006).

The following articles are related to the above topic:

Oberstrass F.C., Lee A., Stefl R., Janis M., Chanfreau G. and Allain F.H.-T (2006) Shape-specific recognition in the structure of the Vts1p SAM domain with RNA. *Nat Struct. Mol. Biol.* 13, 160-167. Author contributions (Stefl R): 10%, designed experiments, interpreted data and performed structural calculations. IF=12.712

5. Protein-protein interaction modules of RNA-binding proteins

In addition to RNA, RNA-binding proteins often associate with other protein factors. The protein-protein interactions can either be mediated by canonical protein-protein interaction modules or by RNA-binding domains. The RNA-dependent association of RNA-binding domains is used to achieve higher affinity, through cooperative binding. Alternatively, the RNA-binding domains can associate in order to form a specific conformation of the RNA substrate required for its function, for example to create RNA loops (Oberstrass et al., 2005). The association of RNA-binding domains can also prevent RNA binding, as the mutual interaction can mask the RNA-interaction surface (Bae et al., 2007).

The canonical protein-protein interaction modules of RNA-binding proteins are often important for the recruitment of RNA-binding proteins to the site of action (for example to the site of transcription), and/or are involved in the assembly of a larger architecture that is required for its function. A typical example of such modular arrangement is found in Nrd1, which contains an RRM and a CTD-interacting domain (CID). Nrd1 is required for the Nrd1-dependent termination pathway, used at small nuclear/nucleolar RNAs (sn/snoRNAs), cryptic unstable transcripts (CUTs), and other short RNAP II transcripts (Arigo et al., 2006; Jasnovidova and Stefl, 2012; Kubicek et al., 2012; Steinmetz et al., 2001; Thiebaut et al., 2006; Vanacova and Stefl, 2007; Vasiljeva et al., 2008). The CID of Nrd1 recruits the entire Nrd1 complex to the site of transcription in the early elongation stage of the transcriptional cycle. This interaction is spatially and temporally regulated by the phosphorylation at serine 5 of the C-terminal domain (CTD) of RNA polymerase II (Munoz et al., 2010; Vasiljeva et al., 2008). The CTD consists of tandem repeats with the consensus sequence $Y_1S_2P_3T_4S_5P_6S_7$ (Figure 6). The number of canonical heptad repeats and the level of their conservation varies across different organisms (Jasnovidova and Stefl, 2012). For instance, budding yeast possess 26 heptad repeats, whereas humans have a CTD that is twice as long, consisting of 52 repeats (Buratowski, 2003; Egloff and Murphy, 2008; Meinhart et al., 2005).

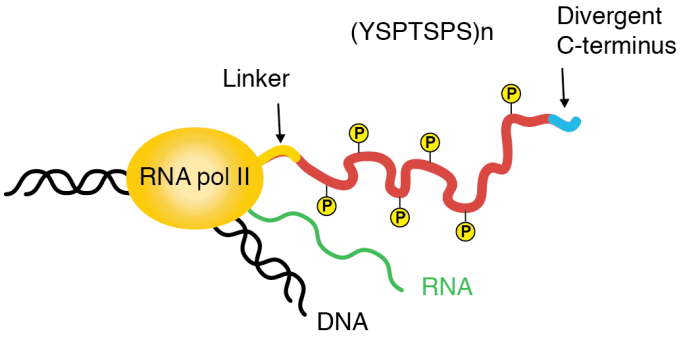


Figure 6. Schematic representation of the transcribing RNA polymerase II with its C-terminal domain that undergoes phosphorylations.

CIDs are the most studied family of the CTD-binding domains. CIDs are found in proteins that are involved in RNA processing pathways, such as the subunit of yeast Rat1 exonuclease complex Rtt103, human RNA processing factor SCAF8, the subunit of yeast cleavage factor IA Pcf11, and the subunit of yeast poly A-independent transcription termination complex Nrd1 (Jasnovidova and Stefl, 2012). CIDs consist of eight α-helices arranged into right-handed superhelical arrangement, where helices 2, 4 and 7 create a groove that interacts with the CTD. Structural studies have revealed that the CIDs bind from eight to eleven residues of the CTD. In all of the structures of the CID-CTD complexes, the CTD peptide adopts a classical β-turn conformation that is accommodated in the binding groove of the CID (Figure 7). The β-turn is formed by $S_{2b}P_{3b}T_{4b}S_{5b}$ and is always stabilized by three intramolecular H-bonds regardless of the phosphorylation pattern. The recognition of the CTD peptide conformation is achieved by many specific contacts between the CID-containing proteins and CTD peptides. These contacts include H-bonds between the CID and P_{6a}, S_{7a}, Y_{1b} and S_{5b}. The side-chain hydroxyl group of Y_{1b} forms a H-bond with a conserved aspartate of CID, whereas Y_{1b} and P_{3b} are tightly accommodated in the hydrophobic groove of the CID. NMR studies have shown that the recognition motif that is important for Pcf11 CID binding lies within two canonical repeats (Jasnovidova and Stefl, 2012; Noble et al., 2005).

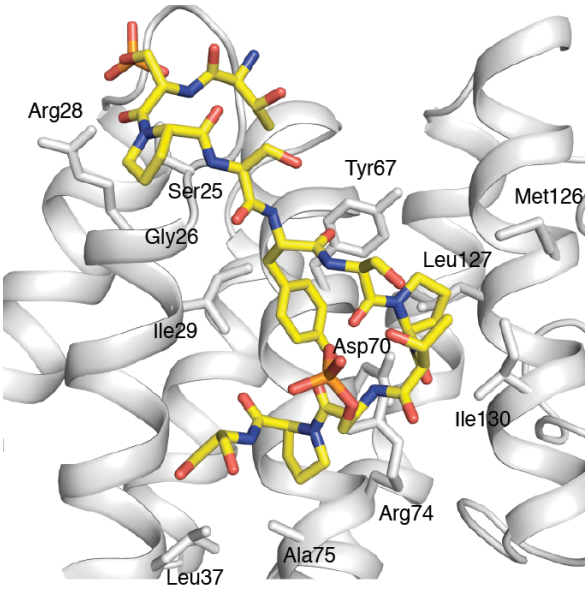


Figure 7. Recognition of the pSer5 CTD by Nrd1. Structure showing contacts between the Nrd1 CID and the CTD peptide. Protein residues that form hydrogen bonds and hydrophobic contacts to the CTD peptide are shown in white sticks.

Interestingly, the Rtt103 CID binds the pS_2 CTD with a higher affinity than the Pcf11 CID (Lunde et al., 2010; Meinhart and Cramer, 2004; Noble et al., 2005). This is explained by the presence of a conserved arginine residue in Rtt103, and also SCAF8, which

forms a salt-bridge interaction with the phosphate group of pS_{2b}. Mutations of the arginine residue in Rtt103 and SCAF8 decreased their affinities to the pS_2 CTD. Interestingly, a low affinity binding of Pcf11 to the pS_2 CTD was improved by the introduction of this arginine residue at the corresponding position in Pcf11 (Lunde et al., 2010). ChIP studies showed that the arginine mutations in Rtt103 impaired its recruitment to the 3' end processing site in vivo (Jasnovidova and Stefl, 2012; Lunde et al., 2010).

In all complexes, the S-P peptidyl-prolyl bonds are in the *trans* conformation with the exception of the structure of the Nrd1 CID that is bound to the pS_5 CTD (Figure 7). Nrd1 binds more residues upstream of the β-turn compared with the other CID-CTD complexes. It uses a conserved region of the CID at the N-terminal tip of helix α2 to specifically recognize pS_{5a} via H-bonding that is mediated by serine and arginine (Kubicek et al., 2012). The structure of the Nrd1 CID in a complex with the pS_5 CTD shows that the specific recognition of pS_{5a} is facilitated by the *cis* conformation of the pS_{5a}-P_{6a} peptidyl-prolyl bond. The *cis* conformation maximizes the intermolecular contacts and prevents the peptide from clashing with the α1-α2 loop of Nrd1 (Kubicek et al., 2012). Mutations at the complex interface that are associated with the *cis* conformer selection and phosphoserine interactions diminish the binding affinity and impair the processing or degradation of the non-coding RNAs. These findings demonstrate the interplay between the covalent and non-covalent changes in the CTD structure that constitute the CTD code (Jasnovidova and Stefl, 2012).

All of the CID-CTD complexes contain highly conserved structural features involving the Y_1 residue: the Y_1 hydroxyl group forms a H-bond with a conserved aspartate of the CID, and the aromatic ring of Y_1 is tightly accommodated in the hydrophobic pocket of the CID. Therefore, it has long been speculated that the phosphorylation or other modification of Y_1 may have detrimental consequences to the binding with CIDs. Indeed, recently it was shown that the phosphorylation of Y_1 impairs the binding to all three yeast CID-containing proteins, Nrd1, Pcf11, and Rtt103, yet it stimulates the binding of elongation factor Spt6 (Mayer et al., 2012).

The CTD phosphorylation is not the only post-translational modifications of the CTD that is important for its function. The CTD can also be glycosylated at serine and threonine (Jasnovidova and Stefl, 2012). In mammals, non-consensus repeats of the CTD contain two arginine and seven lysine substitutions that predominantly occur at position seven of the heptad motif. Lysine can be mono-, di- or trimethylated, while arginine can be mono- or dimethylated, and the arginine dimethylation can be

asymmetrical or symmetrical. It was shown that Arg1810 of the human CTD is methylated by the co-activator-associated arginine methyltransferase 1 (CARM1) and that the CTD methylation facilitates the expression of snRNAs and snoRNAs (Sims et al., 2011).

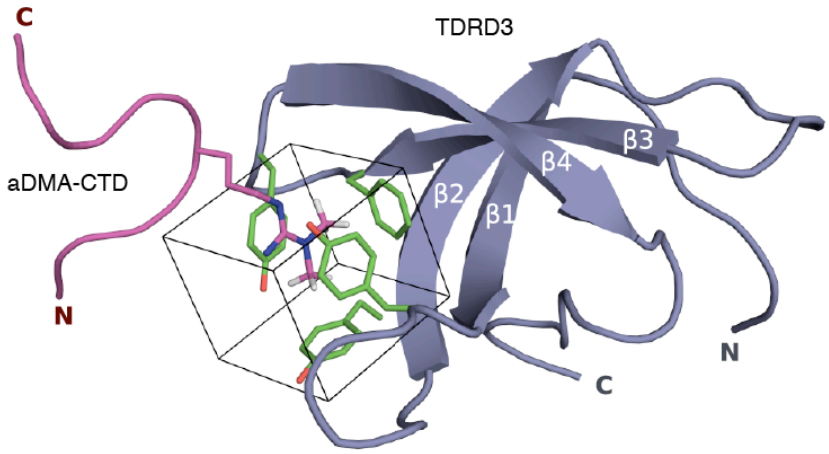


Figure 8. NMR structure of the TDRD3 Tudor–aDMA-CTD complex. Structure of the human TDRD3 Tudor domain bound to the aDMA-CTD peptide. The aDMA-CTD peptide is represented as a magenta ribbon with the aDMA residue in sticks (only methyl protons are shown) and the protein is shown as a blue ribbon model. Residues forming the aromatic cavity (Y567, Y573, F591, and Y594) are shown in green sticks and the cube-shaped cavity is highlighted by a box.

In order to reveal the structural basis of the selective recognition of the aDMA marks by TDRD3, we have determined the solution structure of the Tudor domain of TDRD3 in complex with asymmetrically dimethylated CTD. In contrast to other Tudor domain-containing proteins, Tudor domain-containing protein 3 (TDRD3) associates selectively with the aDMA marks but not with other methylarginine motifs. Our determined structure and mutational analysis provide a molecular basis for how TDRD3 recognizes the aDMA mark. The β1-β2 and β3-β4 loops of the Tudor domain contain aromatic residues (Y566, Y573, F591, and Y594). These residues form a partially exposed aromatic cavity of rectangular cuboid shape (Figure 8). Upon binding of TDRD3 to the CTD peptide, the aromatic cavity accommodates aDMA, that is placed parallel between two tyrosines (Y566 and Y594). This binding results from cation-π and stacking interactions between positively charged guanidinium group of aDMA and aromatic rings of the two parallel tyrosine residues. Furthermore, CH-π interactions between the methyl groups of aDMA and two remaining aromatic residues (Y573 and F591) stabilize the interaction. One guanidinomethyl group faces Y573 whereas the second guanidinomethyl group faces F591 at the back wall of the cavity. Our NMR data show no intermolecular contacts between the aromatic cavity and neighboring residues of aDMA. This suggests that those residues are flexible and do not interact with the

TDRD3 Tudor domain, and that the CTD peptide with the aDMA mark is recognized in a sequence-independent manner. Furthermore, phosphorylations of aDMA-CTD (at Ser2 and Ser5) showed no effect on the binding affinity to TDRD3. This work contributes to the understanding of substrate selectivity rules of the Tudor aromatic cavity, which is an important structural motif for reading of methylation marks (Sikorsky et al., 2012).

The following articles are related to the above topic:

Sikorsky T, Hobor F, Krizanova E, Pasulka J, Kubicek K, Stefl R. (2012) Recognition of asymmetrically dimethylated arginine by TDRD3. *Nucleic Acids Res*. doi: 10.1093/nar/gks929. Author contributions (Stefl R): 20%, designed experiments and wrote the paper. IF=8.026

Kubicek K, Cerna H, Holub P, Pasulka J, Hrossova D, Loehr F, Hofr C, Vanacova S, Stefl R (2012) Serine phosphorylation and proline isomerization in RNAP II CTD control recruitment of Nrd1. *Genes Dev.* 26, 1891-1896. Author contributions (Stefl R): 20%, designed and performed experiments, and wrote the paper. IF=11.659

Kubicek K, Pasulka J, Cerna H, Löhr F, Stefl R. (2011) 1H, 13C, and 15N resonance assignments for the CTD-interacting domain of Nrd1 bound to Ser5-phosphorylated CTD of RNA polymerase II. *Biomol NMR Assign.* 5, 203-205. Author contributions (Stefl R): 10%, designed experiments and wrote the paper. IF=0.720

Jasnovidova O, Stefl R. (2012) The CTD Code of RNA Polymerase II: a Structural View WIREs RNA. doi: 10.1002/wrna.1138. Author contributions (Stefl R): 50%, wrote the paper.

6. Concept of modularity in protein-RNA interactions

Posttranscriptional control of gene expression relies on a highly sophisticated regulatory network of interactions. The concept of modularity is the key feature for these interactions that involve protein-RNA and protein-protein interactions. The multidomain RNA-binding proteins can bind RNA with higher specificity and affinity when compared to individual RNA-binding domains, which can bind only short RNA stretches with weak affinities. At the same time, the modular complexes are easier to disassemble when compared to the single high-affinity one (Lunde et al., 2010), which is important for the dynamics in RNPs. The linker regions tethering individual RNA-binding domains are often important for the specific recognition of RNA targets. Long linkers are usually disordered and allow multiple domains to recognize a diverse set of RNA targets, whereas short linkers impose the domains to bind to a continuous stretch of RNA. The linker can also become ordered upon RNA-binding, e.g., forming a short α-helix 1 that positions the two domains relative to one another and sometimes make additional contacts to RNA substrate (Allain et al., 2000a; Deo et al., 1999; Handa et al., 1999; Lunde et al., 2010; Perez-Canadillas, 2006). In the case, where the linker regions do not participate in the interaction with RNA, the length of the linker is an important parameter. It plays a role in restricting local diffusion, as it defines a maximum distance among tethered domains. In other words, for two tandemly arranged RNA-binding domains, which do not interact with each other, and behave like balls on a string, the RNA binding event of one domain yields an increased local concentration for the other. The effects of the linker length for the resulting affinity was investigated for tandemly arranged RNA-binding domains and RNA containing adjacent recognition sites (Shamoo et al., 1995). The individual binding affinities of two RNA-binding domains were found to be additive in the case of an infinitely long linker, or multiplicative in the case of a very short linker (Shamoo et al., 1995).

7. Different modes of protein-RNA recognitions

The enormous diversity of interactions observed in protein–RNA complexes indicates that a simple recognition code is unlikely to exist in the world of protein–RNA interaction. However, two unifying themes may be inferred from the known complexes: (i) the recognition of the primary RNA sequence, and/or (ii) the recognition of the RNA shape by individual RNA-binding proteins (Stefl et al., 2005a).

7.1 Single-stranded RNA sequence-specific recognitions

The solved structures of RNA-binding proteins bound to single-stranded RNA molecules suggest that there are two basic modes of how sequence-specificity can be achieved. First, the identity of RNA bases can be probed through hydrogen bonds mediated by the protein main-chain carbonyl and amide groups with the Watson-Crick edges of the RNA bases. This mechanism of RNA sequence-specific recognition is manifested by the solution structure of two tandemly arranged CCCH-type zinc fingers of TIS11d bound to RNA. In this complex, each zinc finger recognizes a UAUU sequence (Hudson et al., 2004). In such an arrangement, the fold of these zinc fingers is the only determinant for the specificity and the possibility to modulate sequence-specificity of these domains is rather unlikely.

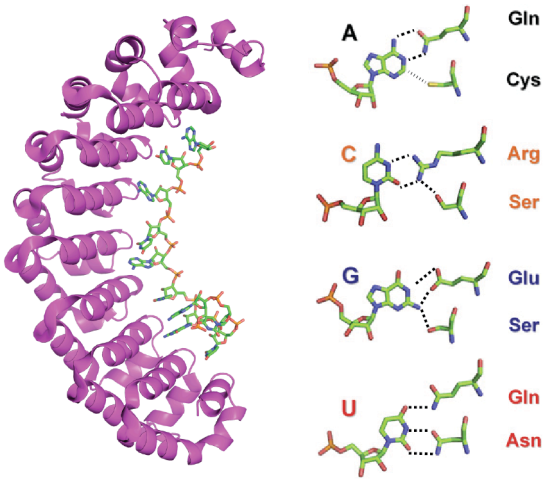


Figure 9. RNA recognition code of PUF proteins. The crystal structure of an engineered PUF domain in complex with RNA (on the left) (Dong et al., 2011). On the right, the RNA recognition code of PUF domain.

In contrast, there is an RNA-binding protein called PUF (PUMILIO and fem-3 binding factor (FBF) homology), where sequence-specificity is exclusively provided by hydrogen bonds between the protein side-chains and the Watson-Crick edges of the RNA bases.

Interestingly, the recognition code for all four RNA bases has recently been revealed (Dong et al., 2011; Filipovska et al., 2011). PUF proteins contain an array of PUF domains (conserved 36-amino acids) that form an extended arc. The single-stranded RNA runs antiparallel to the protein and binds to the inner concave surface created by the multiple PUF repeats (Figure 9) (Dong et al., 2011; Filipovska et al., 2011; Wang et al., 2002). Each PUF repeat recognizes one nucleotide using three amino acids in a base-specific manner. The amino acid side-chain at position 13 in the repeat creates stacking interactions with the aromatic ring of the RNA base. The Watson-Crick edge of the RNA base is recognized by a specific combination of two amino acids (at positions 12 and 16) (Figure 9). The PUF domain can now be engineered to design RNA binders *de novo*, which may be of extraordinary use in both basic research and medical applications as was done for DNA-binding proteins (Klug, 2005).

Historically, the RRMs have been the most studied RNA-binding domains, with more than 20 high-resolution structures solved by X-ray crystallography or NMR spectroscopy. The solved structures rationalize how certain single-stranded sequences can be recognized in a sequence-specific manner. The binding surface of the domain usually contains aromatic residues in order to maximize intermolecular contacts with the RNA bases. The bases are usually spread on the surface of the domain and the phosphates point away to the solvent. The sequence-specific contacts are mediated using a combination of the main-chains and side-chains that form hydrogen bonds with the Watson-Crick edges of the bases. This arrangement has an inherent preference for certain bases at specific positions, which is dictated by the fold of the domain and the presence of side-chains. It is therefore difficult to engineer RRM to modulate its RNA-binding specificity (Auweter et al., 2006b).

7.2 Double-stranded RNA shape- and sequence-specific recognitions

Bi-directional pervasive transcription produces enormous amount of double-stranded RNA in the cell. Secondary structures also occur in protein-coding transcripts as a result of intramolecular folding. These double-stranded RNA (dsRNA) architectures are primarily bound and processed by proteins, which contain dsRBMs (see above, chapter 4.2). The fact that dsRBMs bind to any dsRNA but not to dsDNA, regardless of its base composition, indicates that the domains recognize the specific shape of dsRNA. Indeed, structural studies revealed that the specific shape of the A-form dsRNA is recognized by dsRBMs, involving the contacts with the A-form specific minor and major grooves of the double helix. However, many dsRBM-containing proteins and enzymes act only on certain substrates (e.g., ADAR (RNA editing), DICER (RNAi), Staufen (RNA localization), PKR (signal transduction), TRBP (miRNA) and many others involved in all

post-transcriptional gene regulatory processes), which contrasts with the concept of sequence-independent recognition of dsRNA by dsRBMs.

Adenosine deaminase that acts on RNA 2, ADAR2, is one of such dsRBM-containing enzymes that is highly specific for certain RNA substrates. To address the key question of the RNA editing process mediated by ADAR2—how are the RNA substrates selected for site-specific editing—we solved the solution structure of the two dsRBMs of ADAR2 bound to GluR-2 R/G RNA. The structure provided a molecular basis for how dsRBMs recognize the shape, and also more surprisingly, the sequence of the dsRNA. The unexpected direct readout of the RNA primary sequence by dsRBMs is achieved via the minor groove of the dsRNA and this recognition is critical for both editing and binding affinity at the R/G site of GluR-2 (Stefl et al., 2010). The dsRBM1 and dsRBM2 of ADAR2 preferentially recognize G-X_9-A and G-X_8-A sequences, respectively (Figure 10). The length and the positioning of helix α1 relative to the dsRBM fold appear to be the key structural elements that determine the register length of the different dsRBMs. These structures demonstrate the ability of these domains to recognize RNA sequence–specifically and its functional relevance provide a rationale into how proteins of this family can discriminate substrates based on their secondary structures and also their sequences (Stefl et al., 2010).

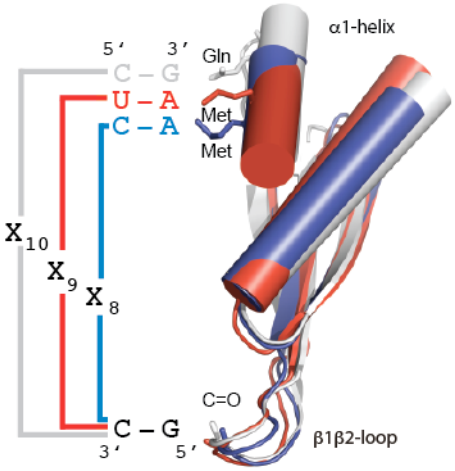


Figure 10. RNA recognition code of various dsRBMs. Overlay of the ADAR2 dsRBM1 (in blue), ADAR2 dsRBM2 (in red), and Aquifex aeolicus RNaseIII dsRBM (in gray) structures highlights the variability of helix α1 within the dsRBM fold and its importance for the determination of the register length between the two specific contacts on the RNA helix.

The following articles are related to the above topic:

Stefl R, Skrisovska L, Allain, FH. (2005) RNA sequence- and shape-dependent recognition by proteins in the ribonucloeprotein particle. *EMBO Rep.* 6:33-38. Author contributions (Stefl R): 33%, analyzed the data and wrote the paper. IF=7.355

Stefl R, Oberstrass FC, Hood JL, Jourdan M, Zimmermann M, Skrisovska L, Maris C, Peng L, Hofr C, Emeson RB Allain FH (2010) The solution structure of the ADAR2 dsRBM-RNA complex reveals a sequence-specific read out of the minor groove. *Cell*, 143, 225-237. Author contributions (Stefl R): 50%, designed and performed experiments, analyzed the data and wrote the paper. IF=32.406

8. Future prospects

Proteins that contain RNA-binding domains and their interactions with RNA have important roles in all aspects of gene expression and regulation. The enormous diversity of interactions observed in the determined structures of protein–RNA complexes indicates that a simple recognition code is unlikely to exist in the world of protein–RNA interactions. From these structures we learned the molecular details about how individual RNA-binding domains recognize RNA substrates, but many of these proteins require multiple copies of one of several RNA-binding domains to function. It is hence essential to understand how multiple RNA-binding modules bind RNA, how they achieve cooperativity and combinatorial binding, and how the modular nature of these RNA-binding proteins determines their biological function. Although more multi-domain protein-RNA complexes still need to be determined, it might soon be possible to predict which RNA-binding protein binds to which RNA, and how it recognizes its target. Furthermore, future studies aimed at unraveling the role of conformational dynamics involved in the multi-domain protein-RNA recognition are expected. As a consequence, gene expression and its regulation could be understood and controlled at the atomic level.

9. References

Allain, F.H., Bouvet, P., Dieckmann, T., and Feigon, J. (2000a). Molecular basis of sequence-specific recognition of pre-ribosomal RNA by nucleolin. EMBO J 19, 6870-6881.

Allain, F.H., Gilbert, D.E., Bouvet, P., and Feigon, J. (2000b). Solution structure of the two N-terminal RNA-binding domains of nucleolin and NMR study of the interaction with its RNA target. J Mol Biol 303, 227-241.

Arigo, J.T., Eyler, D.E., Carroll, K.L., and Corden, J.L. (2006). Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. Mol Cell 23, 841-851.

Auweter, S.D., Fasan, R., Reymond, L., Underwood, J.G., Black, D.L., Pitsch, S., and Allain, F.H. (2006a). Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. EMBO J 25, 163-173.

Auweter, S.D., Oberstrass, F.C., and Allain, F.H. (2006b). Sequence-specific binding of single-stranded RNA: is there a code for recognition? Nucleic Acids Res 34, 4943-4959.

Aviv, T., Lin, Z., Ben-Ari, G., Smibert, C.A., and Sicheri, F. (2006). Sequence-specific recognition of RNA hairpins by the SAM domain of Vts1p. Nat Struct Mol Biol 13, 168-176.

Aviv, T., Lin, Z., Lau, S., Rendl, L.M., Sicheri, F., and Smibert, C.A. (2003). The RNA-binding SAM domain of Smaug defines a new family of post-transcriptional regulators. Nat Struct Biol 10, 614-621.

Bae, E., Reiter, N.J., Bingman, C.A., Kwan, S.S., Lee, D., Phillips, G.N., Jr., Butcher, S.E., and Brow, D.A. (2007). Structure and interactions of the first three RNA recognition motifs of splicing factor prp24. J Mol Biol 367, 1447-1458.

Bahadur, R.P., Zacharias, M., and Janin, J. (2008). Dissecting protein-RNA recognition sites. Nucleic Acids Res 36, 2705-2716.

Bass, B.L. (2002). RNA editing by adenosine deaminases that act on RNA. Annu Rev Biochem 71, 817-846.

Buratowski, S. (2003). The CTD code. Nat Struct Biol 10, 679-680.

Clery, A., Blatter, M., and Allain, F.H. (2008). RNA recognition motifs: boring? Not quite. Curr Opin Struct Biol 18, 290-298.

Cooper, T.A., Wan, L., and Dreyfuss, G. (2009). RNA and disease. Cell 136, 777-793.

Cramer, P., Armache, K.J., Baumli, S., Benkert, S., Brueckner, F., Buchen, C., Damsma, G.E., Dengl, S., Geiger, S.R., Jasiak, A.J., et al. (2008). Structure of eukaryotic RNA polymerases. Annu Rev Biophys 37, 337-352.

Crowder, S.M., Kanaar, R., Rio, D.C., and Alber, T. (1999). Absence of interdomain contacts in the crystal structure of the RNA recognition motifs of Sex-lethal. Proc Natl Acad Sci U S A 96, 4892-4897.

Deo, R.C., Bonanno, J.B., Sonenberg, N., and Burley, S.K. (1999). Recognition of polyadenylate RNA by the poly(A)-binding protein. Cell 98, 835-845.

Dilcher, M., Kohler, B., and von Mollard, G.F. (2001). Genetic interactions with the yeast Q-SNARE VTI1 reveal novel functions for the R-SNARE YKT6. J Biol Chem 276, 34537-34544.

Ding, J., Hayashi, M.K., Zhang, Y., Manche, L., Krainer, A.R., and 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.

Dominguez, C., and Allain, F.H. (2006). NMR structure of the three quasi RNA recognition motifs (qRRMs) of human hnRNP F and interaction studies with Bcl-x G-tract RNA: a novel mode of RNA recognition. Nucleic Acids Res 34, 3634-3645.

Dominguez, C., Fisette, J.F., Chabot, B., and Allain, F.H. (2010). Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs. Nat Struct Mol Biol 17, 853-861.

Dominguez, C., Schubert, M., Duss, O., Ravindranathan, S., and Allain, F.H. (2011). Structure determination and dynamics of protein-RNA complexes by NMR spectroscopy. Prog Nucl Magn Reson Spectrosc 58, 1-61.

Dong, S., Wang, Y., Cassidy-Amstutz, C., Lu, G., Bigler, R., Jezyk, M.R., Li, C., Hall, T.M., and Wang, Z. (2011). Specific and modular binding code for cytosine recognition in Pumilio/FBF (PUF) RNA-binding domains. J Biol Chem 286, 26732-26742.

Doyle, M., and Jantsch, M.F. (2002). New and old roles of the double-stranded RNA-binding domain. J Struct Biol 140, 147-153.

Dreyfuss, G., Kim, V.N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. Nat Rev Mol Cell Biol 3, 195-205.

Duss, O., Lukavsky, P.J., and Allain, F.H. (2012). Isotope Labeling and Segmental Labeling of Larger RNAs for NMR Structural Studies. Adv Exp Med Biol 992, 121-144.

Egloff, S., and Murphy, S. (2008). Cracking the RNA polymerase II CTD code. Trends Genet 24, 280-288.

Esteller, M. (2011). Non-coding RNAs in human disease. Nat Rev Genet 12, 861-874.

Filipovska, A., Razif, M.F., Nygard, K.K., and Rackham, O. (2011). A universal code for RNA recognition by PUF proteins. Nat Chem Biol 7, 425-427.

Finerty, P.J., Jr., and Bass, B.L. (1997). A Xenopus zinc finger protein that specifically binds dsRNA and RNA-DNA hybrids. J Mol Biol 271, 195-208.

Font, J., and Mackay, J.P. (2010). Beyond DNA: zinc finger domains as RNA-binding modules. Methods Mol Biol 649, 479-491.

Fribourg, S., Gatfield, D., Izaurralde, E., and Conti, E. (2003). A novel mode of RBD-protein recognition in the Y14-Mago complex. Nat Struct Biol 10, 433-439.

Golovanov, A.P., Hautbergue, G.M., Tintaru, A.M., Lian, L.Y., and Wilson, S.A. (2006). The solution structure of REF2-I reveals interdomain interactions and regions involved in binding mRNA export factors and RNA. RNA 12, 1933-1948.

Green, J.B., Gardner, C.D., Wharton, R.P., and Aggarwal, A.K. (2003). RNA recognition via the SAM domain of Smaug. Mol Cell 11, 1537-1548.

Grishaev, A., Wu, J., Trewhella, J., and Bax, A. (2005). Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. J Am Chem Soc 127, 16621-16628.

Handa, N., Nureki, O., Kurimoto, K., Kim, I., Sakamoto, H., Shimura, Y., Muto, Y., and Yokoyama, S. (1999). Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein. Nature 398, 579-585.

Hobor, F., Pergoli, R., Kubicek, K., Hrossova, D., Bacikova, V., Zimmermann, M., Pasulka, J., Hofr, C., Vanacova, S., and Stefl, R. (2011). Recognition of transcription termination signal by the nuclear polyadenylated RNA-binding (NAB) 3 protein. J Biol Chem 286, 3645-3657.

Holub, P., Lalakova, J., Cerna, H., Pasulka, J., Sarazova, M., Hrazdilova, K., Arce, M.S., Hobor, F., Stefl, R., and Vanacova, S. (2012). Air2p is critical for the assembly and RNA-binding of the TRAMP complex and the KOW domain of Mtr4p is crucial for exosome activation. Nucleic Acids Res 40, 5679-5693.

Hudson, B.P., Martinez-Yamout, M.A., Dyson, H.J., and Wright, P.E. (2004). Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. Nat Struct Mol Biol 11, 257-264.

Jacquier, A. (2009). The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet 10, 833-844.

Jasnovidova, O., and Stefl, R. (2012). The CTD code of RNA polymerase II: a structural view. Wiley Interdiscip Rev RNA. doi: 10.1002/wrna.1138

Jinek, M., and Doudna, J.A. (2009). A three-dimensional view of the molecular machinery of RNA interference. Nature 457, 405-412.

Johansson, C., Finger, L.D., Trantirek, L., Mueller, T.D., Kim, S., Laird-Offringa, I.A., and Feigon, J. (2004). Solution structure of the complex formed by the two N-terminal RNA-binding domains of nucleolin and a pre-rRNA target. J Mol Biol 337, 799-816.

Johnson, J.M., Edwards, S., Shoemaker, D., and Schadt, E.E. (2005). Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. Trends Genet 21, 93-102.

Johnson, P.E., and Donaldson, L.W. (2006). RNA recognition by the Vts1p SAM domain. Nat Struct Mol Biol 13, 177-178.

Jousset, C., Carron, C., Boureux, A., Quang, C.T., Oury, C., Dusanter-Fourt, I., Charon, M., Levin, J., Bernard, O., and Ghysdael, J. (1997). A domain of TEL conserved in a subset of ETS proteins defines a specific oligomerization interface essential to the mitogenic properties of the TEL-PDGFR beta oncoprotein. EMBO J 16, 69-82.

Klinge, S., Voigts-Hoffmann, F., Leibundgut, M., and Ban, N. (2012). Atomic structures of the eukaryotic ribosome. Trends Biochem Sci 37, 189-198.

Klug, A. (2005). Towards therapeutic applications of engineered zinc finger proteins. FEBS Lett 579, 892-894.

Koradi, R., Billeter, M., and Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 14, 51-55, 29-32.

Kubicek, K., Cerna, H., Holub, P., Pasulka, J., Hrossova, D., Loehr, F., Hofr, C., Vanacova, S., and Stefl, R. (2012). Serine phosphorylation and proline isomerization in RNAP II CTD control recruitment of Nrd1. Genes Dev 26, 1891-1896.

LaCava, J., Houseley, J., Saveanu, C., Petfalski, E., Thompson, E., Jacquier, A., and Tollervey, D. (2005). RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. Cell 121, 713-724.

Loughlin, F.E., Gebert, L.F., Towbin, H., Brunschweiger, A., Hall, J., and Allain, F.H. (2012). Structural basis of pre-let-7 miRNA recognition by the zinc knuckles of pluripotency factor Lin28. Nat Struct Mol Biol 19, 84-89.

Lu, D., Searles, M.A., and Klug, A. (2003). Crystal structure of a zinc-finger-RNA complex reveals two modes of molecular recognition. Nature 426, 96-100.

Lu, K., Heng, X., and Summers, M.F. (2011). Structural determinants and mechanism of HIV-1 genome packaging. J Mol Biol 410, 609-633.

Lukong, K.E., Chang, K.W., Khandjian, E.W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. Trends Genet 24, 416-425.

Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nat Rev Mol Cell Biol 8, 479-490.

Lunde, B.M., Reichow, S.L., Kim, M., Suh, H., Leeper, T.C., Yang, F., Mutschler, H., Buratowski, S., Meinhart, A., and Varani, G. (2010). Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain. Nat Struct Mol Biol 17, 1195-1201.

Macbeth, M.R., Schubert, H.L., Vandemark, A.P., Lingam, A.T., Hill, C.P., and Bass, B.L. (2005). Inositol hexakisphosphate is bound in the ADAR2 core and required for RNA editing. Science 309, 1534-1539.

Mackereth, C.D., Madl, T., Bonnal, S., Simon, B., Zanier, K., Gasch, A., Rybin, V., Valcarcel, J., and Sattler, M. (2011). Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. Nature 475, 408-411.

Mackereth, C.D., and Sattler, M. (2012). Dynamics in multi-domain protein recognition of RNA. Curr Opin Struct Biol 22, 287-296.

Madl, T., Gabel, F., and Sattler, M. (2011a). NMR and small-angle scattering-based structural analysis of protein complexes in solution. J Struct Biol 173, 472-482.

Madl, T., Guttler, T., Gorlich, D., and Sattler, M. (2011b). Structural analysis of large protein complexes using solvent paramagnetic relaxation enhancements. Angew Chem Int Ed Engl 50, 3993-3997.

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

Maris, C., Dominguez, C., and Allain, F.H. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. Febs J 272, 2118-2131.

Masliah, G., Barraud, P., and Allain, F.H. (2012). RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. Cell Mol Life Sci. doi: 10.1007/s00018-012-1119-x

Mayer, A., Heidemann, M., Lidschreiber, M., Schreieck, A., Sun, M., Hintermair, C., Kremmer, E., Eick, D., and Cramer, P. (2012). CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. Science 336, 1723-1725.

Mazza, C., Segref, A., Mattaj, I.W., and Cusack, S. (2002). Large-scale induced fit recognition of an m(7)GpppG cap analogue by the human nuclear cap-binding complex. EMBO J 21, 5548-5557.

Meinhart, A., and Cramer, P. (2004). Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors. Nature 430, 223-226.

Meinhart, A., Kamenski, T., Hoeppner, S., Baumli, S., and Cramer, P. (2005). A structural perspective of CTD function. Genes Dev 19, 1401-1415.

Mendez-Vidal, C., Wilhelm, M.T., Hellborg, F., Qian, W., and Wiman, K.G. (2002). The p53-induced mouse zinc finger protein wig-1 binds double-stranded RNA with high affinity. Nucleic Acids Res 30, 1991-1996.

Messias, A.C., and Sattler, M. (2004). Structural basis of single-stranded RNA recognition. Acc Chem Res 37, 279-287.

Miller, J., McLachlan, A.D., and Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. EMBO J 4, 1609-1614.

Moore, M.J. (2005). From birth to death: the complex lives of eukaryotic mRNAs. Science 309, 1514-1518.

Munoz, M.J., de la Mata, M., and Kornblihtt, A.R. (2010). The carboxy terminal domain of RNA polymerase II and alternative splicing. Trends Biochem Sci 35, 497-504.

Muto, Y., and Yokoyama, S. (2012). Structural insight into RNA recognition motifs: versatile molecular Lego building blocks for biological systems. Wiley Interdiscip Rev RNA 3, 229-246.

Nam, Y., Chen, C., Gregory, R.I., Chou, J.J., and Sliz, P. (2011). Molecular basis for interaction of let-7 microRNAs with Lin28. Cell 147, 1080-1091.

Noble, C.G., Hollingworth, D., Martin, S.R., Ennis-Adeniran, V., Smerdon, S.J., Kelly, G., Taylor, I.A., and Ramos, A. (2005). Key features of the interaction between Pcf11 CID and RNA polymerase II CTD. Nat Struct Mol Biol 12, 144-151.

Oberstrass, F.C., Auweter, S.D., Erat, M., Hargous, Y., Henning, A., Wenter, P., Reymond, L., Amir-Ahmady, B., Pitsch, S., Black, D.L., et al. (2005). Structure of PTB bound to RNA: specific binding and implications for splicing regulation. Science 309, 2054-2057.

Oberstrass, F.C., Lee, A., Stefl, R., Janis, M., Chanfreau, G., and Allain, F.H. (2006). Shape-specific recognition in the structure of the Vts1p SAM domain with RNA. Nat Struct Mol Biol 13, 160-167.

Pabo, C.O., and Nekludova, L. (2000). Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? J Mol Biol 301, 597-624.

Perez-Canadillas, J.M. (2006). Grabbing the message: structural basis of mRNA 3'UTR recognition by Hrp1. EMBO J 25, 3167-3178.

Picard, B., and Wegnez, M. (1979). Isolation of a 7S particle from Xenopus laevis oocytes: a 5S RNA-protein complex. Proc Natl Acad Sci U S A 76, 241-245.

Price, S.R., Evans, P.R., and Nagai, K. (1998). Crystal structure of the spliceosomal U2B"-U2A' protein complex bound to a fragment of U2 small nuclear RNA. Nature 394, 645-650.

Qiao, F., and Bowie, J.U. (2005). The many faces of SAM. Sci STKE 2005, re7.

Ramos, A., Grunert, S., Adams, J., Micklem, D.R., Proctor, M.R., Freund, S., Bycroft, M., St Johnston, D., and Varani, G. (2000). RNA recognition by a Staufen double-stranded RNA-binding domain. EMBO J 19, 997-1009.

Ryter, J.M., and Schultz, S.C. (1998). Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. EMBO J 17, 7505-7513.

Sanudo, M., Jacko, M., Rammelt, C., Vanacova, S., and Stefl, R. (2011). 1H, 13C, and 15N chemical shift assignments of ZCCHC9. Biomol NMR Assign 5, 19-21.

Saunders, L.R., and Barber, G.N. (2003). The dsRNA binding protein family: critical roles, diverse cellular functions. FASEB J 17, 961-983.

Shamoo, Y., Abdul-Manan, N., and Williams, K.R. (1995). Multiple RNA binding domains (RBDs) just don't add up. Nucleic Acids Res 23, 725-728.

Sickmier, E.A., Frato, K.E., Shen, H., Paranawithana, S.R., Green, M.R., and Kielkopf, C.L. (2006). Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65. Mol Cell 23, 49-59.

Sikorsky, T., Hobor, F., Krizanova, E., Pasulka, J., Kubicek, K., and Stefl, R. (2012). Recognition of asymmetrically dimethylated arginine by TDRD3. Nucleic Acids Res. doi: 10.1093/nar/gks929

Simon, B., Madl, T., Mackereth, C.D., Nilges, M., and Sattler, M. (2010). An efficient protocol for NMR-spectroscopy-based structure determination of protein complexes in solution. Angew Chem Int Ed Engl 49, 1967-1970.

Sims, R.J., 3rd, Rojas, L.A., Beck, D., Bonasio, R., Schuller, R., Drury, W.J., 3rd, Eick, D., and Reinberg, D. (2011). The C-terminal domain of RNA polymerase II is modified by site-specific methylation. Science 332, 99-103.

Skrisovska, L., Bourgeois, C.F., Stefl, R., Grellscheid, S.N., Kister, L., Wenter, P., Elliott, D.J., Stevenin, J., and Allain, F.H. (2007). The testis-specific human protein RBMY recognizes RNA through a novel mode of interaction. EMBO Rep 8, 372-379.

Stefl, R., Oberstrass, F.C., Hood, J.L., Jourdan, M., Zimmermann, M., Skrisovska, L., Maris, C., Peng, L., Hofr, C., Emeson, R.B., et al. (2010). The solution structure of the

ADAR2 dsRBM-RNA complex reveals a sequence-specific readout of the minor groove. Cell 143, 225-237.

Stefl, R., Skrisovska, L., and Allain, F.H. (2005a). RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. EMBO Rep 6, 33-38.

Stefl, R., Skrisovska, L., Xu, M., Emeson, R.B., and Allain, F.H. (2005b). Resonance assignments of the double-stranded RNA-binding domains of adenosine deaminase acting on RNA 2 (ADAR2). J Biomol NMR 31, 71-72.

Stefl, R., Xu, M., Skrisovska, L., Emeson, R.B., and Allain, F.H. (2006). Structure and specific RNA binding of ADAR2 double-stranded RNA binding motifs. Structure 14, 345-355.

Steinmetz, E.J., Conrad, N.K., Brow, D.A., and Corden, J.L. (2001). RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts. Nature 413, 327-331.

Theunissen, O., Rudt, F., Guddat, U., Mentzel, H., and Pieler, T. (1992). RNA and DNA binding zinc fingers in Xenopus TFIIIA. Cell 71, 679-690.

Thiebaut, M., Kisseleva-Romanova, E., Rougemaille, M., Boulay, J., and Libri, D. (2006). Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. Mol Cell 23, 853-864.

Tintaru, A.M., Hautbergue, G.M., Hounslow, A.M., Hung, M.L., Lian, L.Y., Craven, C.J., and Wilson, S.A. (2007). Structural and functional analysis of RNA and TAP binding to SF2/ASF. EMBO Rep 8, 756-762.

Tugarinov, V., Hwang, P.M., and Kay, L.E. (2004). Nuclear magnetic resonance spectroscopy of high-molecular-weight proteins. Annu Rev Biochem 73, 107-146.

Tugarinov, V., Kanelis, V., and Kay, L.E. (2006). Isotope labeling strategies for the study of high-molecular-weight proteins by solution NMR spectroscopy. Nat Protoc 1, 749-754.

Vanacova, S., and Stefl, R. (2007). The exosome and RNA quality control in the nucleus. EMBO Rep 8, 651-657.

Vanacova, S., Wolf, J., Martin, G., Blank, D., Dettwiler, S., Friedlein, A., Langen, H., Keith, G., and Keller, W. (2005). A new yeast poly(A) polymerase complex involved in RNA quality control. PLoS Biol 3, e189.

Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S., and Meinhart, A. (2008). The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. Nat Struct Mol Biol 15, 795-804.

Wahl, M.C., Will, C.L., and Luhrmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. Cell 136, 701-718.

Wang, X., McLachlan, J., Zamore, P.D., and Hall, T.M. (2002). Modular recognition of RNA by a human pumilio-homology domain. Cell 110, 501-512.

Wang, Y.X., Zuo, X., Wang, J., Yu, P., and Butcher, S.E. (2010). Rapid global structure determination of large RNA and RNA complexes using NMR and small-angle X-ray scattering. Methods 52, 180-191.

Wolfe, S.A., Nekludova, L., and Pabo, C.O. (2000). DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct 29, 183-212.

Wu, H., Henras, A., Chanfreau, G., and Feigon, J. (2004). Structural basis for recognition of the AGNN tetraloop RNA fold by the double-stranded RNA-binding domain of Rnt1p RNase III. Proc Natl Acad Sci U S A 101, 8307-8312.

Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.C., Dufour, M.E., Boulay, J., Regnault, B., Devaux, F., Namane, A., Seraphin, B., et al. (2005). Cryptic pol II

transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. Cell 121, 725-737.

10. List of publications included in this habilitation work

Sikorsky T, Hobor F, Krizanova E, Pasulka J, Kubicek K, Stefl R. (2012) Recognition of asymmetrically dimethylated arginine by TDRD3. *Nucleic Acids Res*. doi: 10.1093/nar/gks929.

Jasnovidova O, Stefl R. (2012) The CTD Code of RNA Polymerase II: a Structural View WIREs RNA. doi: 10.1002/wrna.1138.

Kubicek K, Cerna H, Holub P, Pasulka J, Hrossova D, Loehr F, Hofr C, Vanacova S, Stefl R. (2012) Serine phosphorylation and proline isomerization in RNAP II CTD control recruitment of Nrd1. *Genes Dev.* 26, 1891-1896.

Porrua O, Hobor F, Boulay J, Kubicek K, D'Aubenton-Carafa Y, Gudipati RK, Stefl R, Libri D. (2012) In vivo SELEX reveals novel sequence and structural determinants of Nrd1-Nab3-Sen1-dependent transcription termination. *EMBO J.* 31, 3935-3948.

Holub P, Lalakova J, Cerna H, Pasulka J, Sarazova M, Hrazdilova K, Arce MS, Hobor F, Stefl R, Vanacova S. (2012). Air2p is critical for the assembly and RNA-binding of the TRAMP complex and the KOW domain of Mtr4p is crucial for exosome activation. *Nucleic Acids Res.* 40, 5679-5693.

Hobor F, Pergoli R, Kubicek K, Hrossova D, Bacikova V, Zimmermann M, Pasulka J, Hofr C, Vanacova S, Stefl R. (2011) Recognition of transcription termination signal by the nuclear polyadenylated RNA-binding (Nab)3 protein. *J. Biol. Chem.* 286, 3645-3657.

Kubicek K, Pasulka J, Cerna H, Löhr F, Stefl R. (2011) 1H, 13C, and 15N resonance assignments for the CTD-interacting domain of Nrd1 bound to Ser5-phosphorylated CTD of RNA polymerase II. *Biomol. NMR Assign.* 5, 203-205.

Stefl R, Oberstrass FC, Hood JL, Jourdan M, Zimmermann M, Skrisovska L, Maris C, Peng L, Hofr C, Emeson RB, Allain FH. (2010) The solution structure of the ADAR2 dsRBM-RNA complex reveals a sequence-specific read out of the minor groove. *Cell* 143, 225-237.

Sanudo M, Jacko M, Rammelt C, Vanacova S, Stefl R. (2010) 1H, 13C, and 15N chemical shift assignments of ZCCHC9. *Biomol. NMR Assign.* 5, 19-21.

Pergoli R, Kubicek K, Hobor F, Pasulka J, Stefl R. (2010) 1H, 13C, and 15N chemical shift assignments for the RNA recognition motif of Nab3. *Biomol. NMR Assign.* 4, 119-121.

Vanacova S, Stefl R. (2007) The exosome and RNA quality control in the nucleus. *EMBO Rep.* 8, 651-657.

Skrisovska L, Bourgeois CF, Stefl R, Grellscheid SN, Kister L, Wenter P, Elliott DJ, Stevenin J, Allain FH. (2007) The testis-specific human protein RBMY recognizes RNA through a novel mode of interaction. *EMBO Rep.* 8, 372-379.

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.

Stefl R, Xu M, Skrisovska L, Emeson RB, Allain FH. (2006) Structure and specific RNA binding of ADAR2 double-stranded RNA binding motifs. *Structure* 14, 345–355.

Stefl R, Allain FH. (2005) A novel RNA pentaloop fold involved in targeting ADAR2. *RNA* 11, 592-597.

Stefl R, Skrisovska L, Allain, FH. (2005) RNA sequence- and shape-dependent recognition by proteins in the ribonucloeprotein particle. *EMBO Rep.* 6:33-38.

Stefl R, Skrisovska L, Xu M, Emeson RB, Allain FH. (2005) Resonance assignments of the double-stranded RNA-binding domains of adenosine deaminase acting of RNA 2 (ADAR2). *J. Biomol. NMR* 31, 71-72.

11. Summary

The present habilitation thesis summarizes the author's contribution to the structural studies of protein-RNA interactions. These interactions have important roles in all aspects of gene expression and regulation. Proteins that mediate protein-RNA interactions are called RNA-binding proteins. They have modular structures and contain multiple copies of RNA-binding domains, which are arranged in a various fashions to convey their diverse functional requirements. This habilitation work describes a number of basic RNA-binding domains and the structural basis of how these domains recognize the sequence and/or the shape of RNA. The work also discusses how RNA-binding proteins are recruited to the site of action, e.g. to the site of transcription, through protein-protein interaction modules. The concept of modularity in protein-RNA interactions is described, focusing on how multiple RNA-binding modules bind RNA via cooperativity and combinatorial binding.

12. Shrnutí

Tato habilitační práce shrnuje příspěvek autora ke strukturnímu studiu interakcí mezi proteiny a RNA. Tyto interakce hrají důležitou úlohu ve všech aspektech exprese genů a jejich regulací. Proteiny které zprostředkovávají vazbu s RNA se nazývají RNA vazebné proteiny. Tyto proteiny mají modulární struktury a obsahují mnoho kopií RNA vazebných domén, které jsou uspořádány v mnoha konfiguracích tak, aby mohly zprotředkovat různé funkce. Habilitační práce popisuje několik základních RNA vazebných modulů a strukturní podstatu toho, jak tyto domény rozpoznávají sekvenci RNA nebo její specifický tvar. Práce dále popisuje jak jsou RNA vazebné proteiny rekrutovány do místa určení, například do místa transkripce, s pomocí modulů zprostředkovávající interakce mezi proteiny. V této práci je popsán koncept modularity v interakcích mezi proteiny a RNA se zaměřením na to, jak RNA vazebné moduly interagují s RNA a využívají k tomu mechanismy vazby kooperativní a kombinatorické.

13. Abbreviations

NMR	Nuclear magnetic resonance
RNA	Ribonucleic acid
CTD	C-terminal domain
RNAPII	RNA polymerase II
RNP	Ribonucleoprotein particle
hnRNP	Heterogeneous nuclear ribonucleoprotein
mRNA	Messenger RNA
RRM	RNA recognition motid
dsRBM	Double-stranded RNA-binding motif
RBP	RNA-binding protein
KH	K-homology
OB	Oligonucleotide/oligosaccharide-binding
PIWI	P-element induced wimpy testis
SAM	Sterile alpha motif
CSD	Cold shock domain
ADAR	Adenosine deaminase that act on RNA
K_D	Dissociation constant
NOE	Nuclear Overhauser effect
RDC	Residual dipolar coupling
PRE	Paramagnetic relaxation enhancement
SAS	Small angle scattering
SAXS	Small angle X-ray scattering
SANS	Small angle neutron scattering
FBF	Fem-3 binding factor
PUF	PUMILIO and FBF homology
UTR	Untranslated region
CBP	Cap binding protein
RBMY	RNA-binding motif gene on Y chromosome
SF	Splicing factor
ASF	Alternative splicing factor
REF	RNA binding and export factor
TRAMP	Trf4/Air2/Mtr4p polyadenylation
SRE	Smaug response element
CUT	Cryptic unstable transcript
CID	CTD-interacting domain
Nrd	Nuclear pre-mRNA down-regulation
Pcf	Protein 1 of cleavage and polyadenylation factor
Rtt	Regulator of Ty1 transposition

ChIP	Chromatin immunoprecipitation
CARM	coactivator-associated arginine methyltransferase
snRNA	small nuclear RNA
snoRNA	Small nucleolar RNA
TDRD	Tudor domain-containing protein
aDMA	Asymmetric dimethylarginine
sDMA	Symmetric dimethylarginine

Nucleic Acids Research Advance Access published October 12, 2012

Nucleic Acids Research, 2012, 1–8
doi:10.1093/nar/gks929

Recognition of asymmetrically dimethylated arginine by TDRD3

Tomas Sikorsky, Fruzsina Hobor, Eva Krizanova, Josef Pasulka, Karel Kubicek and Richard Stefl*

CEITEC—Central European Institute of Technology, Masaryk University, CZ-62500 Brno, Czech Republic

Received June 27, 2012; Revised August 9, 2012; Accepted September 14, 2012

ABSTRACT

Asymmetric dimethylarginine (aDMA) marks are placed on histones and the C-terminal domain (CTD) of RNA Polymerase II (RNAP II) and serve as a signal for recruitment of appropriate transcription and processing factors in coordination with transcription cycle. In contrast to other Tudor domain-containing proteins, Tudor domain-containing protein 3 (TDRD3) associates selectively with the aDMA marks but not with other methylarginine motifs. Here, we report the solution structure of the Tudor domain of TDRD3 bound to the asymmetrically dimethylated CTD. The structure and mutational analysis provide a molecular basis for how TDRD3 recognizes the aDMA mark. The unique aromatic cavity of the TDRD3 Tudor domain with a tyrosine in position 566 creates a selectivity filter for the aDMA residue. Our work contributes to the understanding of substrate selectivity rules of the Tudor aromatic cavity, which is an important structural motif for reading of methylation marks.

INTRODUCTION

Arginine methylation is a frequent post-translational modification of proteins that regulates a variety of cellular processes, including transcriptional regulation, RNA processing, trafficking, signal transduction and DNA repair (1–3). There are three major forms of methylated arginine identified in mammals, monomethylarginine (MMA), asymmetric dimethylarginine (aDMA) and symmetric dimethylarginine (sDMA). These methylation marks are introduced by the protein arginine methyltransferases' (PRMTs) family, in which, type I PRMTs (PRMT1, 2, 3, 4, 6 and 8) generate MMA and aDMA modifications, whereas type II PRMTs (PRMT5 and 7) produce MMA and sDMA modifications. The MMA modifications introduced by both type I and type II PRMTs are likely generated as 'intermediates' on the way to dimethylarginines. PRMTs methylate a large number of protein targets, involved in various aspects of regulation of gene expression (2).

The co-activator-associated arginine methyltransferase 1 (CARM1/PRMT4) deposits an asymmetric dimethylation at the R2 and R17 sites of histone H3 (H3R17 and H3R2) and at the R3 site of histone H4 (H4R3) (4). It also introduces the aDMA mark at R1810 of the mammalian carboxy-terminal domain (CTD) of RNA Polymerase II (RNAP II) (5). The CTD is an important region of RNAP II that undergoes structural remodeling throughout the transcriptional cycle, which allows the association and dissociation of a multitude of effector molecules (6–11). These temporal and spatial interactions couple transcription with most, if not all, pre-mRNA processing steps (12–24). The CTD methylation at R1810 is present on the hyperphosphorylated CTD *in vivo*, and it facilitates the expression of small nuclear and nucleolar RNAs (snRNAs and snoRNAs) (5).

Pull-down experiments showed that Tudor domain-containing protein 3 (TDRD3) displays a specific interaction with the asymmetrically dimethylated R1810-containing CTD peptide (aDMA-CTD) and histones H3 and H4, but not with unmodified and monomethylated or symmetrically dimethylated peptides (4,5). Consistently, using fluorescence polarization, it was reported that TDRD3 preferentially binds the aDMA marks compared with other methylarginine species (25). Other members of the Tudor family, such as the Survival of Motor Neuron (SMN), survival of motor neuron-related splicing factor 30 (SPF30), staphylococcal nuclease domain-containing protein 1 (SND1) and Tur11 do not discriminate between the aDMA- and sDMA-containing peptides (26–28). They are capable of binding to both dimethylarginine isoforms, with slightly higher affinity to sDMA, which also represents their physiological ligand (29,30). The structures of the Tudor domains of SMN, SPF30, SND1 and Tur11 bound to their ligands have been reported (26–28).

In order to reveal the structural basis of selective recognition of the aDMA marks by TDRD3 that has recently been implicated in transcription activation and RNA processing (4,5), we have determined the solution structure of

*To whom correspondence should be addressed. Tel: +42 054 94 92 436; Fax: +42 054 94 92 556; Email: richard.stefl@ceitec.muni.cz

the Tudor domain of TDRD3 in complex with asymmetrically dimethylated CTD.

MATERIALS AND METHODS

Cloning, expression and purification of Tudor domain of TDRD3 (residues 554–608)

The Tudor domain of TDRD3 (residues 554–608) was cloned into pET22b expression vector to generate C-terminal 6x His-tagged protein. The protein was over-expressed in *Escherichia coli* BL21-Codon Plus (DE3)-RIPL (Stratagene) overnight at 16°C after induction by 1 mM IPTG and purified by affinity chromatography on Ni-NTA resin (Qiagen) and further purified on Superdex75 gel filtration column. For nuclear magnetic resonance (NMR) experiments, the protein was concentrated into a buffer containing 50 mM Na_2HPO_4, pH 8.0, 150 mM NaCl and 10 mM β-mercaptoethanol. Mutant constructs were prepared by QuikChange™ Site-directed Mutagenesis Kit with complementary sense (S) and antisense (AS) oligonucleotide primers as follows:

Y566F (S), 5′-cctggagatgaatgttttgcacttttttgggaagacaaca-3′
(AS), 5′-tgttgtcttcccaaaaaagtgcaaaacattcatctc cagg-3′;
Y566W (S), 5′-gaaacctggagatgaatgttttgcactttggtgggaag acaacaag-3′
(AS), 5′-cttgttgtcttcccaccaaagtgcaaaacattcatctc caggtttc-3′;
W567S (S), 5′-cctggagatgaatgttttgcactttattcggaagacaac aagtt-3′
(AS), 5′-aacttgttgtcttccgaataaagtgcaaaacattcat ctccagg-3′;
D569A (S), 5′-gcactttattgggaagcgaacaagttttaccgggc-3′
(AS) 5′-gcccggtaaaacttgttcgcttcccaataaagtgc-3′.

The mutant proteins were over-expressed in the same *E. coli* line and purified using the same procedure as described above. Peptides for NMR measurements were purchased from Clonestar peptide service (Brno, CZ).

Nuclear magnetic resonance

All spectra for the backbone and side-chain assignments of ~1.8 mM uniformly ^{15}N,^{13}C-labeled TDRD3 Tudor in 50 mM sodium phosphate buffer (pH 8.0), 150 mM NaCl, 10 mM β-mercaptoethanol (90% H_2O/10% D_2O) were recorded on Bruker AVANCE 600 MHz spectrometer equipped with a cryoprobe at a sample temperature of 293.15 K. The spectra were processed using NMRPipe package (31) and the protein resonances were assigned manually, using Sparky software (Goddard T.G. and Kellner D.G., University of California, San Francisco, USA). The ^{1}H, ^{13}C and ^{15}N chemical shifts of TDRD3–aDMA-CTD complex were assigned using standard triple resonance experiments (32). All distance constraints were derived from the 3D ^{15}N- and ^{13}C-separated nuclear Overhauser effect spectroscopy (NOESYs) (with mixing time of 110 ms) collected on a 600 MHz spectrometer. Intermolecular distance constraints were obtained from the 3D F_1-^{13}C/^{15}N-filtered NOESY-[^{13}C,^{1}H]-HSQC experiment (33,34), with mixing time of 150 ms on a 600 MHz spectrometer. In the 3D NOESY spectra, intermolecular constraints were semi-quantitatively classified, based on their peak volumes divided by the number of ^{1}H spins involved in nuclear Overhauser effect (NOE) cross-relaxation.

Structure calculation

Structure determination of the TDRD3 Tudor–aDMA-CTD complex was performed with the NOE assignment algorithm implemented in the CYANA program (35). This automated NOE assignment procedure is a re-implementation of the former CANDID algorithm (36) on the basis of a probabilistic treatment of the NOE assignment. CYANA carries out automated assignments, distance calibration of NOE intensities, removal of erroneous restraints, structure calculations with torsion angle dynamics. The resultant NOE cross-peak assignments were subsequently confirmed by visual inspection of the spectra. In the next step, CYANA-generated restraints along with manually assigned TDRD3–aDMA-CTD intermolecular restraints were used for further refinement of 20 preliminary structures with AMBER 11 software (37,38). These calculations employed a modified version (AMBER ff99SB) of the force field described by Cornell *et al.* (39) and an explicit water solvent. The non-standard aDMA residue was parameterized with Gaussian 09 using restrained electrostatic potential charge approach at HF/6-31G(d) level of theory (40). The compound name for aDMA is DA2. Structural quality was assessed using PROCHECK (41) and WHAT IF (42).

Fluorescence anisotropy

The equilibrium binding of Tudor domain of TDRD3 to the CTD peptides was analyzed by fluorescence anisotropy. The CTD peptides were N-terminally labeled with 5,6-carboxyfluorescein (FAM). The measurements were conducted on a FluoroMax-4 spectrofluorometer (Horiba Jobin-Yvon Edison, NJ, USA). Samples were excited with vertically polarized light at 495 nm, and both vertical and horizontal emissions were recorded at 535 nm. All measurements were conducted at 293.15 K in 50 mM phosphate buffer (pH 8.0) containing 150 mM NaCl. A fixed delay of 30 s was set between each aliquot addition and start of the measurement to allow the reaction to reach equilibrium. This delay was sufficient, as no further change in anisotropy was observed. Each data point is an average of three measurements. The data were analyzed using Gnuplot. The experimental isotherms were fit to a single-site binding model using non-linear least squares regression according to Heyduk and Lee (43).

NMR binding experiments

^{1}H-^{15}N Heteronuclear single quantum coherence (HSQC) spectra were recorded at same conditions as NOESY experiments. To determine the affinity of different TDRD3 mutants with aDMA- and sDMA-containing peptides, a series of ^{1}H-^{15}N HSQC spectra were recorded. Increasing

amounts of the unlabeled dimethylarginine-containing peptides were added to ^{15}N-labeled proteins. Combined chemical shift (CCS) is defined as the normalized length of a vector E_j, whose components are chemical shift differences δ_{ji} between observed chemical shift and reference experiment (free form). Index j represents the atom type within the primary sequence of the protein

$$|E_j| = \sqrt{\sum_{i=H,N} w_i \delta_{ji}^2}.$$

Weight factors for each atom type $w_H = 1$ and $w_N = 0.15$ were used. CCS values of three best resolved residues for each titration were used to construct a binding isotherm of 1:1 stoichiometry. Parameters for binding isotherms were found by non-linear least-square regression with Gnuplot. The errors for fitted parameters were calculated within a 95% confidence interval (95% CI).

Quantum chemical calculations

The geometry of aromatic cavity was taken from the deposited NMR structure. The truncated model was made of aDMA and four residues forming the aromatic cavity. The C_α atoms were substituted with methyl groups that were kept frozen during the optimization to account for backbone covalent interactions. The geometry of aDMA and four aromatic cage amino acids was then optimized *in vacuo* using TPSS-D3/def2-TZVP level of theory with Turbomole 6.3 (44,45). For interaction energy calculations, we used a Hartree–Fock wave function of the monomers as a reference from which, we obtained interaction energy terms in first two orders at symmetry-adapted perturbation theory (SAPT0) approximation using PSI4 suite of codes. Density fitting of 4-index integrals and Laplace transformations of energy denominators were used as described in the following article (46). Reference wavefunction in SAPT analysis was calculated using aug-cc-pVDZ basis set.

RESULTS

Solution structure of TDRD3 in complex with aDMA–CTD peptide

Initial NMR titration experiments suggested that two regions of the Tudor domain of TDRD3 (residues 566–573 and 589–598) were significantly perturbed upon binding to the aDMA-CTD peptide (Figure 1). Analysis of NMR titration data under fast exchange regime allowed to evaluate the binding affinity of the TDRD3 Tudor with the sDMA- and aDMA-containing CTD peptides. TDRD3 binds aDMA-CTD with a K_D of 770 ± 30 μM, more than one order of magnitude stronger, compared with the symmetrically methylated isoform (Figure 1D). Similar binding affinity of TDRD3 to fluorescently labeled aDMA-CTD peptide was determined in a quantitative binding assay using fluorescence anisotropy (K_D = 900 ± 200 μM; Supplementary Figures S1 and S2).

To understand the structural basis for molecular recognition of aDMA-containing CTD by TDRD3, we determined the solution structure of a reconstituted complex consisting of the Tudor domain of TDRD3 (residues 555–610) and a 13-mer peptide aDMA-CTD [YSPSSP(aDMA)YTPQSP; Figure 1A] (Figure 2A and B; Supplementary Figure S2B and C; Supplementary Table S1). The Tudor domain of TDRD3 forms a four-stranded β-barrel fold and is similar to the structure of TDRD3 in free form (Supplementary Figure S2B) (25).

The β1–β2 and β3–β4 loops of the Tudor domain contain aromatic residues (Y566, Y573, F591 and Y594). These residues form a partially exposed aromatic cavity of rectangular cuboid shape that is side-walled by N596 (Figure 2C). Upon binding of TDRD3 to the CTD peptide, the aromatic cavity accommodates aDMA that is placed parallel between two tyrosines (Y566 and Y594). This binding results from cation–π and stacking interactions between positively charged guanidinium group of aDMA and aromatic rings of the two parallel tyrosine residues. Furthermore, CH–π interactions between the methyl groups of aDMA and two remaining aromatic residues (Y573 and F591) stabilize the interaction. One guanidinomethyl group faces Y573, whereas the second guanidinomethyl group faces F591 at the back wall of the cavity (Figure 2C). Interestingly, conserved N596 does not form a H-bond to aDMA-CTD, akin to the structures of s/aDMA bound to the canonical Tudor domains of SPF30 and SMN (26). This conserved asparagine in the structures of extended Tudor domains forms an H-bond to sDMA, altering its role in this Tudor subfamily (27,28).

Our NMR data show no intermolecular contacts between the aromatic cavity and neighboring residues of aDMA (Supplementary Figure S2C). This suggests that those residues are flexible and do not interact with the TDRD3 Tudor domain, and that the CTD peptide with the aDMA mark is recognized in a sequence-independent manner. Furthermore, phosphorylations of aDMA-CTD (at Ser2 and Ser5) showed no effect on the binding affinity to TDRD3.

TDRD3 Y566 mutants loose specificity to aDMA-CTD peptides

It was shown that substitution of any of the four aromatic residues of the cavity with a non-aromatic amino acid, abrogates dimethylarginine binding (26). In the present study, we have investigated the effect of aromatic substitutions of the least conserved residue within the aromatic cavity (Figures 2C and 3A). In TDRD3, Y566 is a unique residue, whereas SMN, SPF30, SND1 and Tur11 contain tryptophan or phenylalanine in this position. Y566F substitution diminishes binding to aDMA-CTD and it does not increase binding affinity to sDMA-CTD (Figure 3B and Supplementary Figure S5). On the other hand, Y566W substitution promotes complex formation with sDMA-CTD, yet it has similar binding affinity to aDMA-CTD as the wild-type protein. This indicates that both phenylalanine and tryptophan substitutions at position Y566 abrogate TDRD3 selectivity for aDMA- against sDMA-CTD peptides. Residue Y566 is thus the

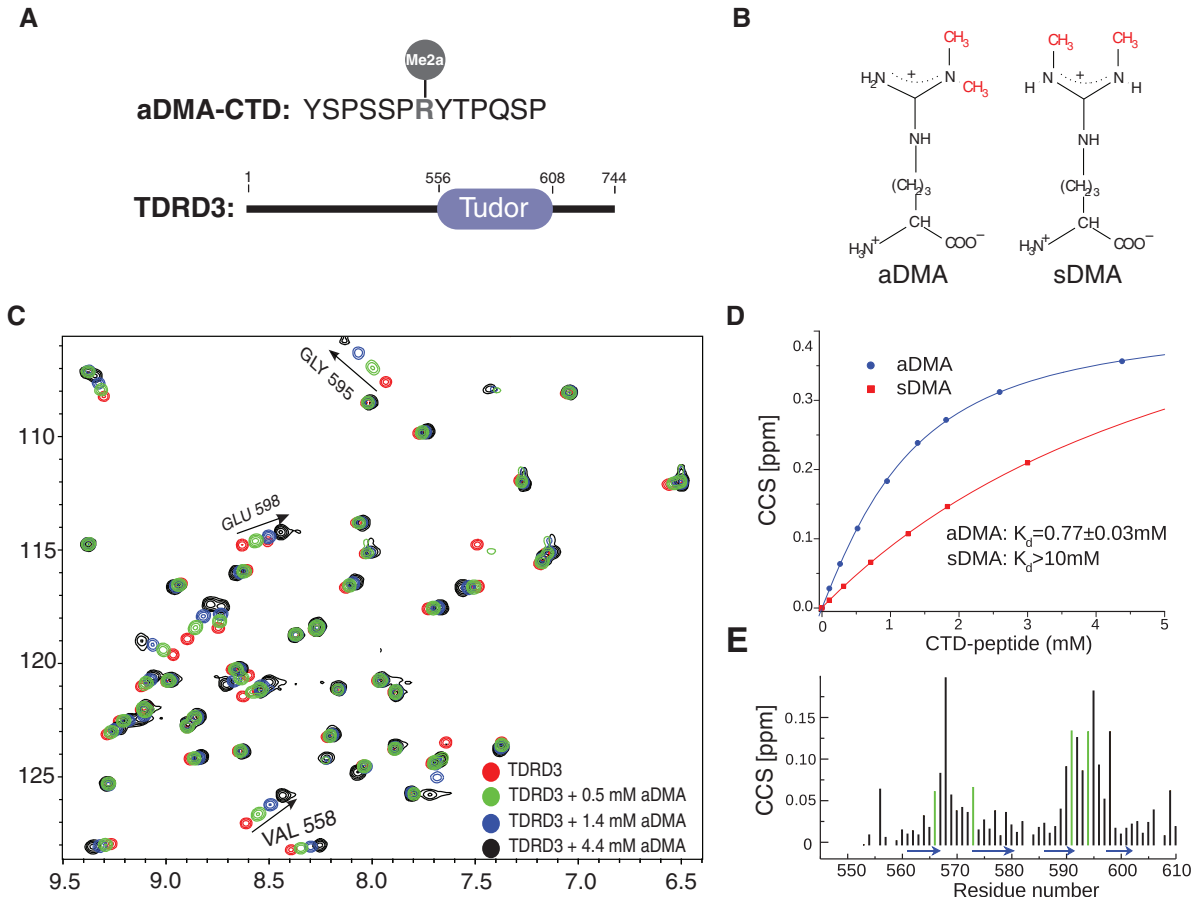


Figure 1. Interaction of the TDRD3 Tudor with aDMA-CTD. (**A**) Sequence of a 13 amino acid peptide with aDMA, used as a mimic of methylated CTD; and domain organization of TDRD3. (**B**) Scheme of aDMA and sDMA. (**C**) A 2D ^{1}H-^{15}N HSQC spectra showing four representative steps of NMR titration of 1.8 mM TDRD3 with aDMA-CTD peptide. Trajectories for the three best resolved signals that were used for construction of binding isotherm are highlighted. (**D**) Estimation of the TDRD3 Tudor–aDMA-CTD (in blue) and TDRD3 Tudor–sDMA-CTD (in red) dissociation constants from NMR titration experiments. CCS values derived from ^{1}H-^{15}N HSQCs are plotted against the CTD peptide concentration. Errors are denoted as 95% CI. (**E**) Quantification of chemical shift perturbations of the TDRD3 Tudor upon addition of aDMA-CTD. The combined chemical shift perturbations are plotted versus the amino-acid residue number with β-sheets regions shown as blue arrows. Large changes occur in the regions involved in binding of the aDMA-CTD peptide. Green lines represent the residues forming hydrophobic cavity.

key element, which determines the specificity of TDRD3 toward aDMA-containing peptides.

Comparison of the TDRD3–aDMA-CTD complex with previously determined structures of SPF30/SMN–a/sDMA complexes show variations in accommodation of the substrates in the aromatic cavities (Supplementary Figure S3A). The sDMA substrates are inserted at a different angle than aDMA, so that they maximize stacking interactions with the tryptophan aromatic residue in the position 566 (TDRD3 numbering). This underlines the importance of the aromatic residue type (Y/W/F; Figure 3A) in this position for dimethylarginine recognition.

Pyramidalization of aDMA amino group promotes hydrogen bond formation with TDRD3

Force field approximations along with limited resolution of experimental data often imperfectly describe molecular interactions. In our structure of the TDRD3–aDMA-CTD complex, we identified a possibility of H-bond between the aDMA amino group and the hydroxyl group of Y566. Such H-bond would require a distortion of the aDMA amino group planarity. Studies of high-resolution x-ray structures have shown amino group pyramidalization allowing hydrogen bond formation in nucleic acids (47). However, empirical force fields used for the structural calculations enforce the planarity of the amino group; therefore, we had to resort to other methods. To probe the existence of this H-bond, we performed a dispersion-corrected density functional theory (DFT-D) study of aDMA bound to the aromatic cavity of TDRD3 (for details, see 'Materials and Methods' section). Geometry of the DFT-D optimized model is very close to the NMR structure with heavy atom RMSD of 0.9 Å (Supplementary Figure S3B). Most importantly, the planarity of aDMA amino group was slightly distorted (due to partial sp^3 hybridization), allowing the H-bond formation with the hydroxyl group of Y566 (Figure 3D). As the cavity is formed by hydrophobic residues and its interior is not accessible to solvents, it is likely that this H-bond has significant energy. Y566F substitution in TDRD3 or the lack of

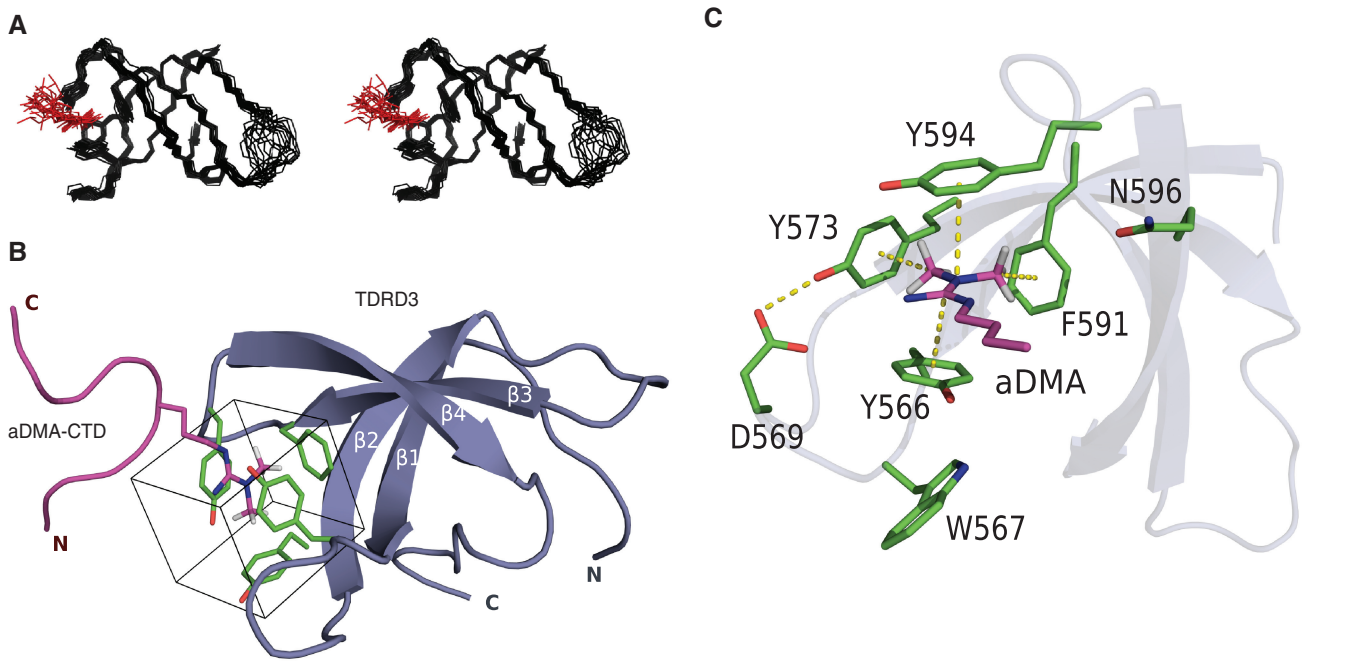


Figure 2. NMR structure of the TDRD3 Tudor–aDMA-CTD complex. (**A**) Overlay of the 20 lowest energy structures of the TDRD3 Tudor–aDMA-CTD complex shown in stereo view. Backbone and side-chains of the aromatic cavity (Y566, Y573, F591 and Y594) are shown in wire representation. For clarity, only aDMA residue of the bound peptide is shown in wire representation. The TDRD3 Tudor domain and aDMA are shown in black and red, respectively. (**B**) Structure of the human TDRD3 Tudor domain bound to the aDMA-CTD peptide. The aDMA-CTD peptide is represented as a magenta ribbon with the aDMA residue in sticks (only methyl protons are shown) and the protein is shown as a blue ribbon model. Residues forming the aromatic cavity (Y566, Y573, F591 and Y594) are shown in green sticks and the cube-shaped cavity is highlighted by a square. (**C**) aDMA recognition by the TDRD3 Tudor domains. The hydrogen bond and aDMA interactions with the aromatic cavity are shown with yellow dotted lines. Only aDMA and the side-chain of amino acids that form, or interact with the aromatic cavity are shown (starting from C_α atoms).

amino group in sDMA prevents formation of this H-bond (Figures 1D and 3B). This hydrogen bond explains why, during titration experiments, Y566F substitution abrogated TDRD3 binding with aDMA and why sDMA did not bind to wild-type TDRD3.

Dispersion interactions play a major role in complex formation

To understand the forces contributing to the complex formation, we calculated the interaction energy of previously optimized DFT-D model of TDRD3–aDMA complex using SAPT calculations (48). In this perturbation approach, the interaction energy is expressed as a sum of physically well defined parts. Electrostatic and exchange-repulsion terms represent the first order contributions, whereas the second order terms are represented by induction and dispersion energy. Calculated interaction energy terms are summarized in Figure 3C.

The 'electrostatic' term is stabilizing. This is consistent with our picture that $N-H\cdots O$ hydrogen bond and cation–π interactions contribute to complex formation. However, the 'exchange-repulsion' over-compensates the electrostatic attraction, as many electron pairs come close when aDMA and aromatic residues form the complex. Therefore, the first order contributions are repulsive. The second order contributions showed that 'induction' term plays a minor role for the overall binding energy. In general, cation–π interactions are dominated by electrostatic forces and cation-induced polarization of π systems (49). This suggests that induction mainly represents an interaction between a positive guanidinium group of aDMA and π orbitals of parallel tyrosines. The 'dispersion' interaction is the most significant stabilizing term. It should be noted that all non-covalent interactions contain some degree of a dispersion-type component (49). However, the role of dispersion in cation–π and hydrogen bonds is modest (50). As the dispersion energy is almost as large as the total interaction energy, we conclude that the dispersion originates predominantly from stacking interactions between aDMA and two parallel tyrosines (Y566 and Y594).

Surrounding amino acids stabilize the aromatic cavity

The structure of the TDRD3–aDMA-CTD complex reveals that D569 of the β1–β2 loop forms the H-bond with the hydroxyl group of Y573 and thus, stabilizes the geometry of the aromatic cavity (Figure 2C). If the H-bond formation is impaired by D569A substitution, the [^{1}H,^{15}N]-HSQC spectrum is significantly altered when compared with that of the wild-type (Supplementary Figure S4A). The residues of the aromatic cavity and their nearest neighbors display large chemical shift perturbations, suggesting that the geometry of the cavity is disturbed (Supplementary Figure S4B). Furthermore, D569A substitution abrogates binding to the aDMA-CTD peptide, indicating the importance of D569 for

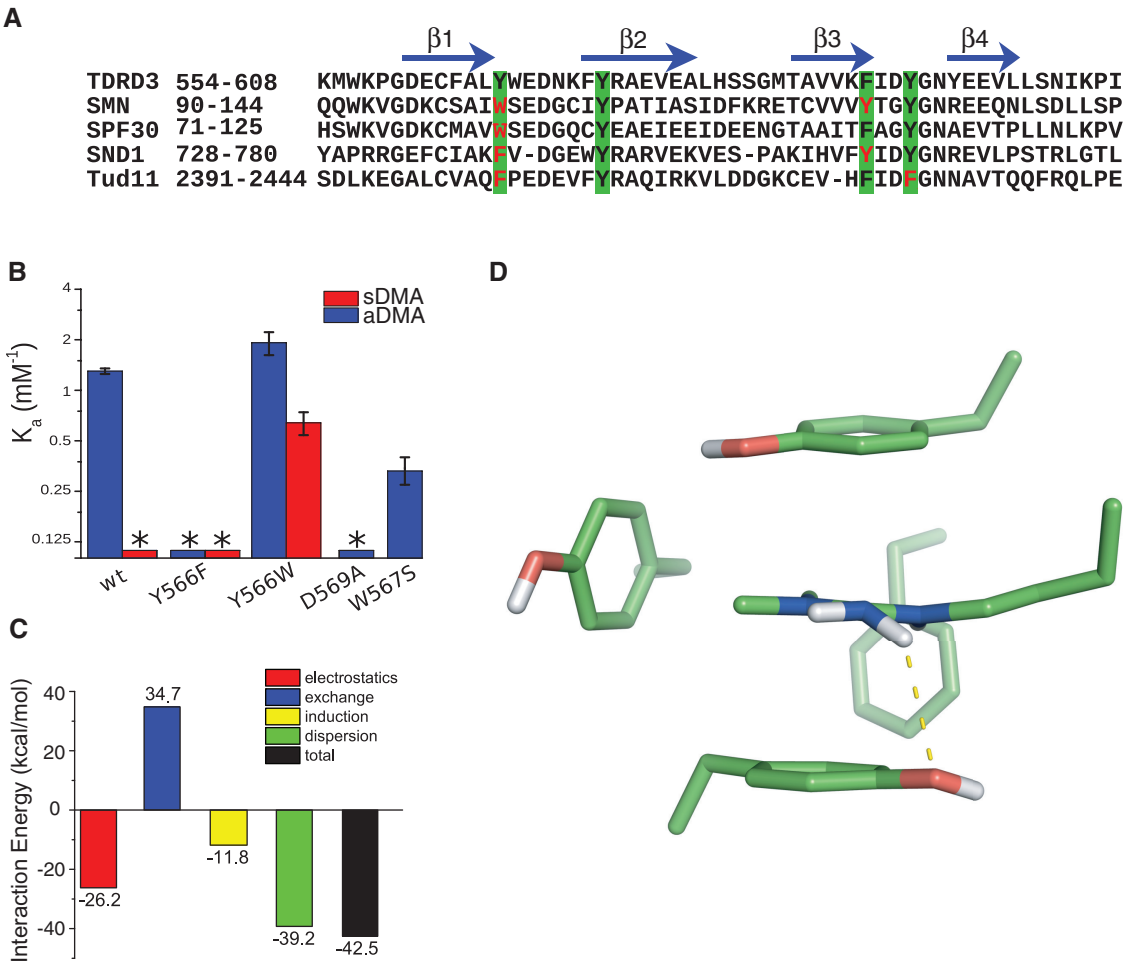


Figure 3. Recognition of aDMA by the TDRD3 Tudor. (**A**) Sequence alignment of the human TDRD3 Tudor domain with other dimethylarginine binding Tudor domains of SMN (*Homo sapiens*), SPF30 (*H. sapiens*), SND1 (*H. sapiens*), and the 11th Tudor domain of Tud (*Drosophila melanogaster;* Tud11). Residues forming an aromatic cavity are highlighted in green squares, variable residues are shown in red. The β-sheet regions (β1, β2, β3 and β4) of human TDRD3 Tudor are shown with blue arrows. (**B**) Bar plot of the NMR-derived association constants (K_a) of various TDRD3 mutants with the aDMA-CTD peptide (blue) and sDMA-CTD peptide (red) in a logarithmic scale. Asterisk indicates that the binding constant could not be determined. (**C**) Bar plot showing decomposed interactions' energies between the aromatic cavity and aDMA calculated by SAPT. The exchange-repulsion (blue bar) accounts for an interaction caused by tunneling of the electrons between interacting systems and electron–electron repulsion due to the Pauli exclusion principle. The induction interaction (yellow bar) is a second-order energy contribution, which originates from mutual distortion of electron density distribution of interacting molecules. The dispersion interaction (green) arises from the correlated electron fluctuations in the interacting molecules (48). (**D**) Pyramidalization of the aDMA amino group as predicted by DFT-D theory. The hydrogen bond (2.7 Å) that is responsible for aDMA recognition is shown with yellow dotted line. Only heavy atoms and non-polar hydrogens are shown.

the stabilization of the aromatic cavity (Supplementary Figure S4B).

Conserved W567 residue stacks with Y566, stabilizing the geometry of aromatic cavity (Figure 2C and Supplementary Figure S2A). W567S substitution of TDRD3 Tudor yields lower affinity to aDMA-CTD (Figure 3B). These results demonstrate the importance of residues surrounding the cavity for the recognition of aDMA by TDRD3.

DISCUSSION

Available structures of the SMN, SPF30, SND1 and Tud11 Tudor domains in complex with their corresponding dimethylarginine ligands (26–28), together with the structure of the TDRD3 Tudor domain bound to aDMA-CTD, define structural determinants for the recognition of sDMA- and aDMA-containing peptides. The extended Tudor domains (SND1 and Tur11) contain an additional α-helix and two β-strands at aminoterminal and several α-helices and β-strands at the carboxy-terminal to the canonical Tudor core. They preferably interact with sDMA-containing peptides, because of hydrogen bond formation between sDMA and asparagine residue (Supplementary Figure S6). In contrast, the structures of the canonical Tudor domains bound to dimethylarginine ligands lack this hydrogen bond (Supplementary Figure S7). They utilize a different strategy for the dimethylarginine recognition. The SMN and SPF30 canonical Tudor domains contain tryptophan at position 566 (Figure 3A). The presence of the fused-ring heterocyclic amino acid alters the stacking interactions

between aromatic cavity and ligand. As a result, both SMN and SPF30 preferably bind to sDMA but also possess a comparable affinity to aDMA.

The presence of tyrosine at position 566 in the canonical Tudor domain of TDRD3, creates unique binding properties of the cavity as it selectively binds the aDMA mark but not the sDMA mark. The TDRD3–aDMA complex formation is driven by a combination of cation–π, CH–π and stacking interactions between aromatic cavity of Tudor domain and aDMA ligand. The complex is further stabilized by the $N-H\cdots O$ hydrogen bond between the aDMA amino group and the hydroxyl group of Y566. Y566F variant of TDRD3, which causes the loss of this hydrogen bond, abrogates the interaction between TDRD3 and aDMA. Interestingly, Y566W variant allows TDRD3 to form a complex with sDMA-CTD and also has a similar affinity to aDMA-CTD (Figure 3B). The presence of tryptophan in this position provides an additional dispersion interaction that is non-specific to the dimethylarginine motifs. Y566W mutant of the TDRD3 Tudor domain is therefore, able to bind both sDMA and aDMA peptides even though the $N-H\cdots O$ hydrogen bond is broken by this substitution. We conclude that the tyrosine residue at position 566 of the TDRD3 Tudor domain works as a selectivity filter for the aDMA-containing peptides.

The results presented in this study extend our understanding of how the methylarginine binding Tudor domains coordinate their ligands. A comparison of the ligand-bound TDRD3 structure with other Tudor domains, supported by the mutational analysis and quantum chemical calculations provide a framework to understand dimethylarginine recognition at a molecular level.

ACCESSION NUMBERS

The atomic coordinates for the NMR ensemble of the TDRD3–aDMA-CTD complex have been deposited in the Protein Data Bank under accession code 2lto. The chemical shift assignments have been deposited in the Biological Magnetic Resonance Data Bank under accession code 18490.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary Figures 1–7.

ACKNOWLEDGEMENTS

We thank Dr Frank Löhr for the NMR pulse programs for the three-dimensional F_1-$^{13}C/^{15}N$-filtered NOESY-[^{13}C,^{1}H]-HSQC experiments. T.S. did quantum chemical calculations, carried out NMR experiments and performed structure calculations. F.H. and E.K. carried out sample preparations, performed and analyzed the FA experiments. J.P. contributed to structure calculations. K.K. contributed to NMR experiments. R.S. conceived and designed the project and contributed to NMR experiments. All authors discussed the results and wrote the manuscript.

FUNDING

CEITEC-Central European Institute of Technology [CZ.1.05/1.1.00/02.0068] from European Regional Development Fund; Czech Science Foundation [P305/12/G034 and P305/10/1490]; Brno City Municipality Scholarship for Talented Ph.D. Students (to F.H.). Funding for open access charge: Czech Science Foundation.

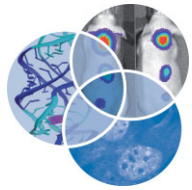
Conflict of interest statement. None declared.

REFERENCES

1. Bedford,M.T. and Richard,S. (2005) Arginine methylation an emerging regulator of protein function. *Mol. Cell*, **18**, 263–272.
2. Bedford,M.T. and Clarke,S.G. (2009) Protein arginine methylation in mammals: who, what, and why. *Mol. Cell*, **33**, 1–13.
3. Siomi,M.C., Mannen,T. and Siomi,H. (2010) How does the royal family of Tudor rule the PIWI-interacting RNA pathway? *Genes Dev.*, **24**, 636–646.
4. Yang,Y., Lu,Y., Espejo,A., Wu,J., Xu,W., Liang,S. and Bedford,M.T. (2010) TDRD3 is an effector molecule for arginine-methylated histone marks. *Mol. Cell*, **40**, 1016–1023.
5. Sims,R.J. 3rd, Rojas,L.A., Beck,D., Bonasio,R., Schuller,R., Drury,W.J. 3rd, Eick,D. and Reinberg,D. (2011) The C-terminal domain of RNA polymerase II is modified by site-specific methylation. *Science*, **332**, 99–103.
6. Phatnani,H.P. and Greenleaf,A.L. (2006) Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev.*, **20**, 2922–2936.
7. Corden,J.L. (2007) Transcription. Seven ups the code. *Science*, **318**, 1735–1736.
8. Chapman,R.D., Heidemann,M., Hintermair,C. and Eick,D. (2008) Molecular evolution of the RNA polymerase II CTD. *Trends Genet.*, **24**, 289–296.
9. Egloff,S. and Murphy,S. (2008) Cracking the RNA polymerase II CTD code. *Trends Genet.*, **24**, 280–288.
10. Buratowski,S. (2003) The CTD code. *Nat. Struct. Biol.*, **10**, 679–680.
11. Buratowski,S. (2009) Progression through the RNA polymerase II CTD cycle. *Mol. Cell*, **36**, 541–546.
12. Hirose,Y. and Manley,J.L. (2000) RNA polymerase II and the integration of nuclear events. *Genes Dev.*, **14**, 1415–1429.
13. Maniatis,T. and Reed,R. (2002) An extensive network of coupling among gene expression machines. *Nature*, **416**, 499–506.
14. Meinhart,A., Kamenski,T., Hoeppner,S., Baumli,S. and Cramer,P. (2005) A structural perspective of CTD function. *Genes Dev.*, **19**, 1401–1415.
15. Viladevall,L., St Amour,C.V., Rosebrock,A., Schneider,S., Zhang,C., Allen,J.J., Shokat,K.M., Schwer,B., Leatherwood,J.K. and Fisher,R.P. (2009) TFIIH and P-TEFb coordinate transcription with capping enzyme recruitment at specific genes in fission yeast. *Mol. Cell*, **33**, 738–751.
16. Ghosh,A., Shuman,S. and Lima,C.D. (2011) Structural insights to how mammalian capping enzyme reads the CTD code. *Mol. Cell*, **43**, 299–310.
17. de la Mata,M. and Kornblihtt,A.R. (2006) RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20. *Nat. Struct. Mol. Biol.*, **13**, 973–980.
18. Munoz,M.J., de la Mata,M. and Kornblihtt,A.R. (2010) The carboxy terminal domain of RNA polymerase II and alternative splicing. *Trends Biochem. Sci.*, **35**, 497–504.
19. David,C.J., Boyne,A.R., Millhouse,S.R. and Manley,J.L. (2011) The RNA polymerase II C-terminal domain promotes splicing

activation through recruitment of a U2AF65-Prp19 complex. *Genes Dev.*, **25**, 972–983.
20. de Almeida,S.F., Grosso,A.R., Koch,F., Fenouil,R., Carvalho,S., Andrade,J., Levezinho,H., Gut,M., Eick,D., Gut,I. *et al.* (2011) Splicing enhances recruitment of methyltransferase HYPB/Setd2 and methylation of histone H3 Lys36. *Nat. Struct. Mol. Biol.*, **18**, 977–983.
21. Ahn,S.H., Kim,M. and Buratowski,S. (2004) Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3′ end processing. *Mol. Cell*, **13**, 67–76.
22. Egloff,S., O'Reilly,D., Chapman,R.D., Taylor,A., Tanzhaus,K., Pitts,L., Eick,D. and Murphy,S. (2007) Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. *Science*, **318**, 1777–1779.
23. Johnson,S.A., Kim,H., Erickson,B. and Bentley,D.L. (2011) The export factor Yra1 modulates mRNA 3′ end processing. *Nat. Struct. Mol. Biol.*, **18**, 1164–1171.
24. MacKellar,A.L. and Greenleaf,A.L. (2011) Cotranscriptional association of mRNA export factor Yra1 with C-terminal domain of RNA polymerase II. *J. Biol. Chem.*, **286**, 36385–36395.
25. Liu,K., Guo,Y., Liu,H., Bian,C., Lam,R., Liu,Y., Mackenzie,F., Rojas,L.A., Reinberg,D., Bedford,M.T. *et al.* (2012) Crystal structure of TDRD3 and methyl-arginine binding characterization of TDRD3, SMN and SPF30. *PLoS One*, **7**, e30375.
26. Tripsianes,K., Madl,T., Machyna,M., Fessas,D., Englbrecht,C., Fischer,U., Neugebauer,K.M. and Sattler,M. (2011) Structural basis for dimethylarginine recognition by the Tudor domains of human SMN and SPF30 proteins. *Nat. Struct. Mol. Biol.*, **18**, 1414–1420.
27. Liu,K., Chen,C., Guo,Y., Lam,R., Bian,C., Xu,C., Zhao,D.Y., Jin,J., MacKenzie,F., Pawson,T. *et al.* (2010) Structural basis for recognition of arginine methylated Piwi proteins by the extended Tudor domain. *Proc. Natl Acad. Sci. USA*, **107**, 18398–18403.
28. Liu,H., Wang,J.Y., Huang,Y., Li,Z., Gong,W., Lehmann,R. and Xu,R.M. (2010) Structural basis for methylarginine-dependent recognition of Aubergine by Tudor. *Genes Dev.*, **24**, 1876–1881.
29. Cote,J. and Richard,S. (2005) Tudor domains bind symmetrical dimethylated arginines. *J. Biol. Chem.*, **280**, 28476–28483.
30. Chen,C., Jin,J., James,D.A., Adams-Cioaba,M.A., Park,J.G., Guo,Y., Tenaglia,E., Xu,C., Gish,G., Min,J. *et al.* (2009) Mouse Piwi interactome identifies binding mechanism of Tdrkh Tudor domain to arginine methylated Miwi. *Proc. Natl Acad. Sci. USA*, **106**, 20336–20341.
31. Delaglio,F., Grzesiek,S., Vuister,G.W., Zhu,G., Pfeifer,J. and Bax,A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, **6**, 277–293.
32. Bax,A. and Grzesiek,S. (1993) Methodological advances in protein NMR. *Accounts Chem. Res.*, **26**, 131–138.
33. Peterson,R.D., Theimer,C.A., Wu,H. and 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.
34. Zwahlen,C., Legault,P., Vincent,S.J.F., Greenblatt,J., Konrat,R. and Kay,L.E. (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.
35. Guntert,P. (2004) Automated NMR structure calculation with CYANA. *Methods Mol. Biol.*, **278**, 353–378.
36. Herrmann,T., Guntert,P. and 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.
37. Case,D.A., Darden,T.A., Cheatham,T.E. III, Simmerling,C.L., Wang,J., Duke,R.E., Luo,R., Crowley,M., Walker,R.C., Zhang,W. *et al.* (2008) *AMBER 10*. University of California, San Francisco.
38. Case,D.A., Cheatham,T.E. 3rd, Darden,T., Gohlke,H., Luo,R., Merz,K.M. Jr, Onufriev,A., Simmerling,C., Wang,B. and Woods,R.J. (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.*, **26**, 1668–1688.
39. Cornell,W.D., Cieplak,P., Bayly,C.I., Gould,I.R., Merz,K.M., Ferguson,D.M., Spellmeyer,D.C., Fox,T., Caldwell,J.W. and Kollman,P.A. (1995) A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. *J. Am. Chem. Soc.*, **117**, 5179–5197.
40. Bayly,C.I., Cieplak,P., Cornell,W.D. and Kollman,P.A. (1993) A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges - the Resp Model. *J. Phys. Chem.*, **97**, 10269–10280.
41. Laskowski,R.A., Rullmannn,J.A., MacArthur,M.W., Kaptein,R. and Thornton,J.M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR*, **8**, 477–486.
42. Vriend,G. (1990) What If - a molecular modeling and drug design program. *J. Mol. Graphics*, **8**, 52–56.
43. Heyduk,T. and Lee,J.C. (1990) Application of fluorescence energy transfer and polarization to monitor Escherichia coli cAMP receptor protein and lac promoter interaction. *Proc. Natl Acad. Sci. USA*, **87**, 1744–1748.
44. Grimme,S., Antony,J., Ehrlich,S. and Krieg,H. (2010) A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.*, **132**, 154104.
45. Tao,J., Perdew,J.P., Staroverov,V.N. and Scuseria,G.E. (2003) Climbing the density functional ladder: nonempirical meta-generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.*, **91**, 146401.
46. Turney,J.M., Simmonett,A.C., Parrish,R.M., Hohenstein,E.G., Evangelista,F.A., Fermann,J.T., Mintz,B.J., Burns,L.A., Wilke,J.J., Abrams,M.L. *et al.* (2012) Psi4: an open-source ab initio electronic structure program. *WIREs: Comput. Mol. Sci.*, **2**, 556–565.
47. Luisi,B., Orozco,M., Sponer,J., Luque,F.J. and Shakked,Z. (1998) On the potential role of the amino nitrogen atom as a hydrogen bond acceptor in macromolecules. *J. Mol. Biol.*, **279**, 1123–1136.
48. Jeziorski,B., Moszynski,R. and Szalewicz,K. (1994) Perturbation-theory approach to intermolecular potential-energy surfaces of Van-Der-Waals complexes. *Chem. Rev.*, **94**, 1887–1930.
49. Cubero,E., Luque,F.J. and Orozco,M. (1998) Is polarization important in cation-pi interactions? *Proc. Natl Acad. Sci. USA*, **95**, 5976–5980.
50. Kim,D., Hu,S., Tarakeshwar,P., Kim,K.S. and Lisy,J.M. (2003) Cation-pi interactions: A theoretical investigation of the interaction of metallic and organic cations with alkenes, arenes, and heteroarenes. *J. Phys. Chem. A*, **107**, 1228–1238.

Advanced Review



The CTD code of RNA polymerase II: a structural view

Olga Jasnovidova and Richard Stefl*

RNA polymerase II (RNA pol II) is not only the fundamental enzyme for gene expression but also the central coordinator of co-transcriptional processing. RNA pol II associates with a large number of enzymes and protein/RNA-binding factors through its C-terminal domain (CTD) that consists of tandem repeats of the heptapeptide consensus $Y_1S_2P_3T_4S_5P_6S_7$. The CTD is posttranslationally modified, yielding specific patterns (often called the CTD code) that are recognized by appropriate factors in coordination with the transcription cycle. Serine phosphorylations are currently the best characterized elements of the CTD code; however, the roles of the proline isomerization and other modifications of the CTD remain poorly understood. The dynamic remodeling of the CTD modifications by kinases, phosphatases, isomerases, and other enzymes introduce changes in the CTD structure and dynamics. These changes serve as structural switches that spatially and temporally regulate the binding of processing factors. Recent structural studies of the CTD bound to various proteins have revealed the basic rules that govern the recognition of these switches and shed light on the roles of these protein factors in the assemblies of the processing machineries.

How to cite this article:
WIREs RNA 2012. doi: 10.1002/wrna.1138

INTRODUCTION

RNA polymerase II (RNA pol II) transcribes not only protein-coding genes but also a variety of small non-coding RNAs, including small nuclear and nucleolar (sn/sno) RNAs, cryptic unstable transcripts (CUTs), stable unannotated transcripts (SUTs), and Xrn1-dependent unstable transcripts (XUTs).[1–4] In contrast to RNA polymerases I and III,[5] the largest subunit of RNA pol II, Rpb1, contains a long, flexible C-terminal domain (CTD). The CTD can be divided into three parts: (1) a flexible linker region, (2) a region consisting of tandem repeats of the consensus sequence $Y_1S_2P_3T_4S_5P_6S_7$ (the short stretch of the repetitive portion of the CTD is termed the CTD peptide throughout this review), and (3) a divergent C-terminal portion (Figure 1(b)). The CTD of yeast spans up to 650 Å in an extended conformation and is positioned near the RNA exit channel of Rpb1. In this manner, both the location and flexible nature of the CTD allow for the binding of many protein factors in the proximity of the nascent transcript.[6] The CTD is dynamically modified and serves as a platform for the recruitment of enzymes and regulatory proteins that are involved in transcription and RNA processing.[7–10] The association of specific posttranslational modifications of the CTD with particular events of the transcription cycle gave rise to the concept of the CTD code.[11] Key components of the CTD code are posttranslational modifications of the CTD that are generated by CTD-modifying enzymes and CTD-associated protein factors that read these modification marks. The proper combination of these factors leads to spatially and temporally controlled events that are involved in transcription and RNA processing. As new structural data accumulates, it is becoming clear that not only the posttranslational modification but also the recognition of particular conformations of the CTD by the regulatory proteins are important parts of the CTD code.

*Correspondence to: richard.stefl@ceitec.muni.cz
CEITEC-Central European Institute of Technology, Masaryk University, Brno, Czech Republic

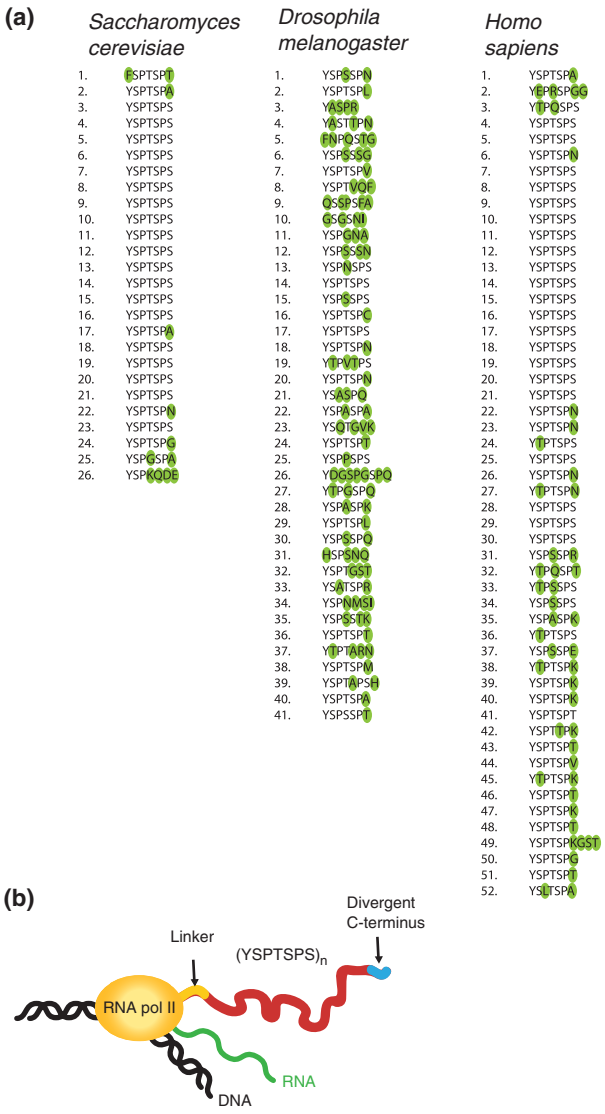


FIGURE 1 | The C-terminal domain (CTD) of RNA polymerase II. (a) The conservation of amino acid residues within the canonical heptad of the CTD. The amino acid sequences of the CTD from *Sacchoromyces cerevisiae* (shown are residues 1542–1723 from total 1733, UniProt ID P04050), *Drosohila melanogaster* (shown are residues 1598–1881 from total 1887, UniProt ID P04052), and *Homo sapience* (shown are residues 1593–1960 from total 1970, UniProt ID P24928). The amino acid residues shown in green do not possess the CTD consensus. Numbers on the left side indicate the sequential number of the heptads. (b) Schematic representation of RNA polymerase II largest subunit.

C-TERMINAL DOMAIN OF RNA POLYMERASE II

Canonical CTD Repeat: Tyr_1-Ser_2-Pro_3-Thr_4-Ser_5-Pro_6-Ser_7

The CTD consists of tandem repeats with the consensus sequence $Y_1S_2P_3T_4S_5P_6S_7$ (Figures 1(a) and 2(a)). The number of canonical heptad repeats and the level of their conservation vary across different organisms.[12,13] For instance, budding yeast possesses 26 heptad repeats, whereas humans have a CTD that is twice as long, consisting of 52 repeats.

The conservation of repeats within the CTD is different (Figure 1(a)). For example, the CTD of yeast is one of the most conserved, with only 6 out of 26 repeats differing from the consensus in several positions. In humans, only 21 CTD repeats out of 52 possess the full consensus. In comparison, the CTD of the fruit fly is highly divergent, having only two canonical CTD repeats.[12,13] It has been established that not all positions of the heptad are equally important among different species. In budding yeast, substitutions of T_4 and S_7 to alanine have little to no impact on cell growth,[14] whereas the conversion of S_2 and S_5 to glutamate or alanine residues and that of Y_1 to phenylalanine are lethal.[15,16] In contrast, fission yeast is also able to tolerate the S2A mutation during normal growth, but the same mutation causes mating defects and results in yeast sterility.[16,17] Despite the viabilities of the S7A mutants, there is evidence of the phosphorylation of S_7 and its importance to yeast cells.[18–21] Similarly, the S_7 position was shown to be involved in snRNA gene transcription and processing in metazoans.[22–25]

To assess the length requirements of the CTD repeats, a yeast growth analysis showed that as few as 8 repeats of the conserved consensus are sufficient for yeast viability, whereas 10 heptad repeats grow similar to the wild type.[15] Surprisingly, yeast mutants with low numbers of wild-type CTD repeats were able to recover from deletions and reconstructed the full-length CTDs by reduplication. However, those possessing low numbers of mutated CTD repeats were not able to reduplicate the full-length CTD.[15] This observation suggests that the function of the CTD may lie not only in the number of repeats but also in the overall length of the CTD sequence. The attainment of the proper length is necessary to allow for the optimal spatial organization of protein factors on the CTD. The separation of the heptad pairs in yeast via the insertion of alanine tracts showed that the minimal functional unit of the CTD lies within two heptads.[26] Increased distances between pairs of heptads resulted in low viabilities and reduced levels of CTD phosphorylation by various CTD kinases.[13,27]

Posttranslational Modifications of CTD

Dynamic posttranslational modifications of the CTD are key to CTD functioning. Phosphorylation is the most frequent modification of the CTD, and it can occur at five out of seven amino acid residues (Y_1, S_2, T_4, S_5, and S_7) of the heptad (Figure 3(a)). The

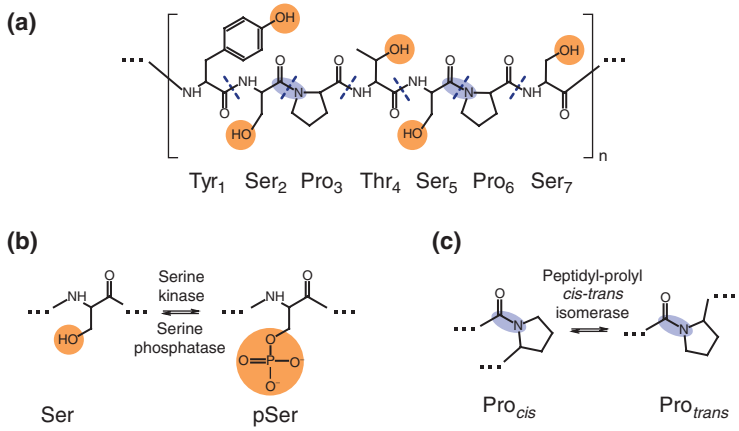


FIGURE 2 | Structural representation of the C-terminal domain (CTD) heptad and its possible modifications. (a) Structural formula of the canonical heptapeptide of the CTD. (b) Representation of the serine phosphorylation. (c) Representation of the serine–proline peptidyl–prolyl bond isomerization. Small orange circles indicate hydroxyl group that can be phosphorylated or modified otherwise, big orange circle indicates phosphorylation, blue ellipses indicate the serine–proline peptidyl–prolyl bonds that can have two isomeric conformations, blue dashed lines show visual separation between amino acid residues.

phosphorylation marks are introduced by various kinases and can be removed by specific phosphatases[7] (Figure 3(b)). Serine phosphorylations (pS_2, pS_5, and pS_7) are relatively well-understood and crucial for many processing events of coding and non-coding RNAs.[28,29] T_4 phosphorylation has recently been implicated in histone mRNA 3′ end processing[30] and transcription elongation,[31] but its overall significance remains in question. Similarly, Y_1 can also be phosphorylated,[32,33] and this phosphorylation stimulates the binding of the elongation factor Spt6. Interestingly, Y_1 phosphorylation impairs the binding of factors that are involved in early and late RNA processing events.[33] In addition to phosphorylation, the CTD can be glycosylated at serine and threonine.[34,35] Recently, the glycosylations of positions S_5 and S_7 were implicated in the formation of pre-initiation complexes in higher eukaryotes.[36] In mammals, nonconsensus repeats of the CTD contain two arginine and seven lysine substitutions that predominantly occur at position 7 of the heptad motif (Figure 1(a)). Lysine can be mono-, di-, or trimethylated, while arginine can be mono- or dimethylated, and the arginine dimethylation can be asymmetrical or symmetrical. It was shown that Arg1810 of the human CTD is methylated by the coactivator-associated arginine methyltransferase 1 (CARM1) and that the CTD methylation facilitates the expression of snRNAs and snoRNAs.[37] Furthermore, six of seven lysines (Lys1859, Lys1866, Lys1873, Lys1887, Lys1908, and Lys1922) of the mice CTD were shown to be ubiquitinylated by Wwp2, which is a mouse HECT domain ubiquitin E3 ligase.[38] Lysine can also be a target for acetylation and sumoylation.

Phosphorylations and other modifications not only alter the chemical structure of the CTD but also enrich or restrict the conformational variability on the domain and effect recognition by other factors. For example, the phosphate group of phosphoserine

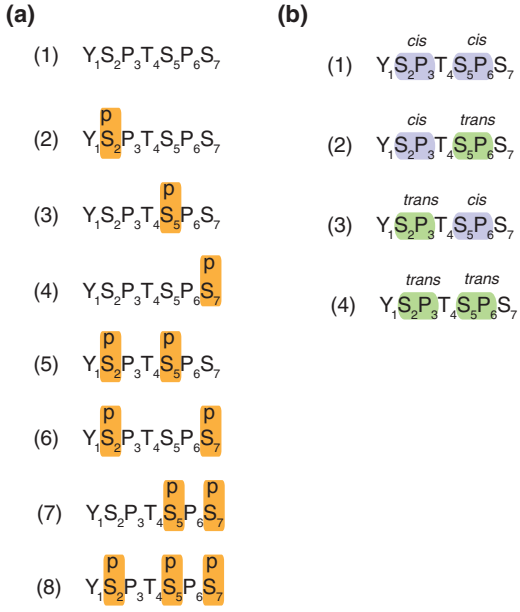


FIGURE 3 | Example of modification patterns creating 'the CTD code' (C-terminal domain). (a) Possible patterns of the serine phosphorylations within a single heptad repeat. (b) Possible patterns of the serine–proline peptidyl–prolyl bond isomerization within a single heptad repeat.

carries a double negative charge (Figure 2(b)) and is capable of establishing multiple hydrogen bonds and salt bridges. The introduction of the methyl group to arginine or lysine has the opposite effect; it eliminates the possibility for the formation of H-bonds but promotes hydrophobic interactions. Phosphorylation and glycosylation are accomplished using the same hydroxyl groups of serine and threonine residues, causing the modifications to be mutually exclusive. Glycosylation introduces a relatively large sugar moiety to the peptide chain and is predicted to function as a steric block to prevent aberrant phosphorylation.[34,35] Ubiquitination and

sumoylation confer even more drastic changes to the CTD structure. The additions of carbohydrate, ubiquitin, and SUMO likely block access to the neighboring amino acids and prevent or regulate the dynamic exchange of other posttranslational modifications and binding factors. Detailed mechanisms and the functional relevance that associated with these modifications are currently poorly understood.

Any posttranslational modifications result in the formation or disruption of specific protein–protein interactions. These types of protein–protein interactions usually involve specialized interaction domains on the partner protein that recognize the phosphorylated segment of the CTD peptide. As different positions can be modified simultaneously, it may lead to the combinatorial explosion of modification patterns (Figure 3). If we take into account that only serine positions are phosphorylated, this would result in eight possible combinations of phosphorylation, including non-phosphorylated peptides (Figure 3(a)). Multiple phosphorylations of the CTD create not only new recognition sites but also change the overall structure of the CTD. Because of the charge repulsion between phosphate groups, the CTD structure is believed to be becoming more extended upon phosphorylation.[6]

The isomerization of S_2–P_3 and S_5–P_6 peptide bonds is another important element of the CTD code[39–41] (Figure 2(c)). The serine–proline peptidyl–prolyl bonds can adopt two distinct *cis* and *trans* isomers (Figure 2(c)). In equilibrium, the fraction of X-Pro peptide bonds in the *cis* conformation ranges from 10 to 40% depending on the preceding amino acid and the structural context.[39] The majority of peptide bonds in the *cis* conformation occur in surface-accessible bend, coil, or turn motifs.[39] Switches between the *cis* and *trans* isomers induce large structural changes, introducing sharp turns into the backbone that disrupt previous interactions and also introducing new epitopes for recognition. The CTD canonical heptad has two bonds (S_2–P_3 and S_5–P_6) that can exist in different isomeric forms, resulting in four possible heptapeptide conformations (Figure 3(b)). A spontaneous interconversion between the two isomeric forms is very slow because the *cis* and *trans* conformations are separated by a high energy barrier.[39] The *cis* and *trans* isomers create two distinct populations of CTD peptides with specific conformations that can establish a new set of protein–protein interactions. Furthermore, the conversion between the two conformers can be catalyzed by a class of specific enzymes, peptidyl–prolyl *cis–trans* isomerases (PPIases),[42] that control the timing of isomer specific protein–protein interactions.[39–41]

CTD MODIFICATIONS ARE BELIEVED TO ORCHESTRATE TRANSCRIPTION

The coordinated exchange of the CTD modification marks orchestrates mechanisms involved in the regulation of transcription and RNA processing, which involves controlled interplay among dozens of protein factors. Distinct modification patterns during different stages of the transcription cycle gave rise to the idea of the CTD code,[9,11] which involves PPIases in coordination with CTD kinases and phosphatases in determining the phosphorylation and conformation of the CTD, thereby directing the dynamic association of the factors during transcription.[43]

In the canonical model of the CTD phosphorylation cycle,[7,9,10,44] the hypophosphorylated CTD interacts with the multi-subunit mediator complex and general transcription factors at the beginning of the transcription cycle.[40,41] Thus, RNA pol II is recruited to the promoter region of a gene.[45,46] Upon the formation of the pre-initiation complex, the S_5 position becomes phosphorylated by TFIIH (CDK7/Kin28) and the mediator-associated kinases (CDK8/Srb10) that induce the dissociation of the mediator complex from the CTD.[47] Then, the S_5-phosphorylated CTD recruits mRNA capping enzymes, histone-modifying enzymes (Set2-containing complex), and RNA processing factors.[8,48,49] As the transcription cycle proceeds, the CTD loses most of its S_5 phosphorylations and becomes phosphorylated by the positive transcription elongation factor b (pTEF-b) complex at position S_2, yielding a mixture of doubly phosphorylated CTD repeats at S_2 and S_5 in the center of the gene.[50] As RNA pol II approaches the 3′ end of the gene, S_5 phosphorylation levels decrease, but high levels of S_2 phosphorylation prompt the recruitment of cleavage, polyadenylation, and termination factors.[51–53] After the release of transcripts and termination of transcription, the CTD is recycled by phosphatases to reenter the transcription cycle.

This universal model of the CTD phosphorylation cycle has been challenged by two recent genome-wide studies in yeast, which suggested that the CTD phosphorylation cycle may not be uniform at all genes but rather be gene-class specific.[20,21] However, two other works have suggested that the phosphorylation marks are placed and removed similarly at all genes as a function of the distances from the transcription start sites and termination sites, respectively, which is consistent with the universal model.[54,55] There is also considerable debate regarding the levels of S_7 phosphorylation during the transcription cycle. While its importance has been clearly implicated in the transcription termination of snRNAs in humans,[23] it is unclear whether in yeast the S_7 marks are deposited in

peaks, first by Kin28 and later by Bur1,[21] or uniformly along genes.[54]

STRUCTURAL DIVERSITY OF CTD

Owing to its flexibility, the CTD has not been observed in the crystal structures of RNA pol II.[56] Nuclear magnetic resonance (NMR) spectroscopy evidence suggests that the free CTD is largely flexible, although it has shown some residual structure and a tendency to form β-turns at two SPXX motifs ($S_2P_3T_4S_5$ and $S_5P_6S_7Y_1$).[57–60] A number of models for the overall CTD conformation have been proposed, including random coil[56] and β-spiral models.[13,61] On the basis of the crystal structure of Pcf11, which was bound to the two-heptad repeats of the pS_2 CTD, it has been proposed that these spiral architectures can be very compact as a result of the β-turn stabilization by the protein binding.[62]

The enormous diversity of interactions that have been observed with CTD–protein complexes suggests that the simple rules of the CTD code are unlikely to exist. Currently, available structures of bound CTD peptides show how complex and diverse the recognition of basically the same peptide sequence may be (Table 1, Figure 4). The specificity of binding originates in different modification isoforms, modification patterns, lengths of bound CTD peptides, and *cis* or *trans* conformations of the phosphoserine–proline peptidyl–prolyl bonds. In these complexes, the CTD peptide backbone adopts distinct types of loops when it is bound to the nucleotidyltransferase (NT) domain of Cgt1,[63] the CTD-interacting domain (CID),[62,64,65] the BRCT domain of SCP1,[66] and Ssu72,[67,68] and forms a spiral shape upon binding with the PIN1 WW domain (Figure 4). The stabilization of the peptide conformation and specific recognition is achieved by a combination of steric constrains, intramolecular and intermolecular hydrogen bonds (H-bonds), van der Waals (VDW) forces, and electrostatic and stacking interactions. As a result, each protein factor has its own minimal requirement for the functional unit of a CTD peptide, which can be as long as three repeats, similar to Cgt1,[63] or as short as four residues, similar to SCP1.[66]

The CTD code cannot be explained solely by the phosphorylation patterns, which only make up one portion of the recognition signal. The CTD peptide with the same phosphorylation pattern can be a target for different binding partners (e.g., the interaction of $pS_2P_3T_4pS_5$ with PIN1[71] and SCP1[66]). Even within the same family of protein binding domains, variation can exist in the recognition of the same peptide conformation and modification pattern (discussed below: CID domains, capping enzymes). The examples below display how initially transiently disordered conformations of the CTD are modulated following posttranslational modifications and specific binding. A combination of the flexible nature of the CTD coupled with posttranslation modifications creates an ideal structural mechanism for the regulation of cellular processes.

CTD-Interacting Domains: Twist by β-Turn

CIDs are the most studied family of the CTD-binding domains. CIDs are found in proteins that are involved in RNA processing pathways, such as the subunit of yeast Rat1 exonuclease complex Rtt103,[65,72] human RNA processing factor SCAF8,[64,73] the subunit of yeast cleavage factor IA Pcf11,[62,74,75] and the subunit of yeast poly A-independent transcription termination complex Nrd1.[76,77]

CIDs consist of eight α-helices arranged into right-handed superhelical arrangement, where helices 2, 4, and 7 create a groove that interacts with the CTD. Structural studies have revealed that the CIDs bind from 8 to 11 residues of the CTD (Table 1). In all the structures of the CID–CTD complexes, the CTD peptide adopts a classical β-turn conformation that is accommodated in the binding groove of the CID (Figure 5(a)). The β-turn is formed by $S_{2b}P_{3b}T_{4b}S_{5b}$ and is always stabilized by three intramolecular H-bonds regardless of the phosphorylation pattern[64] (Figure 5(b)). The recognition of the CTD peptide conformation is achieved by many specific contacts between the CID-containing proteins and CTD peptides (Figure 5(b)). These contacts include H-bonds between the CID and $P_{6a,b}$, S_{7a}, Y_{1b}, and S_{5b}. The side-chain hydroxyl group of Y_{1b} forms an H-bond with a conserved aspartate of CID, whereas Y_{1b} and P_{3b} are tightly accommodated in the hydrophobic groove of the CID. NMR studies have shown that the recognition motif that is important for Pcf11 CID binding lies within two canonical repeats.[78] If more than two such recognition sites are present in the CTD, the binding between the Pcf11 CID and the pS_{2b} and pS_{2c} sites is in fast exchange on the NMR time scale,[78] indicating the possible mobility of the CID along the CTD if more than one pS_2 site is present.

Interestingly, the Rtt103 CID binds the pS_2 CTD with a higher affinity than the Pcf11 CID.[62,64,78] This is explained by the presence of a conserved arginine residue in Rtt103, and also SCAF8, which forms a salt-bridge interaction with the phosphate group of pS_{2b}[78] (Figure 5(a)). Mutations of the arginine residue in Rtt103 and SCAF8 decreased their affinities to the pS_2 CTD.[65,77] Interestingly, a

TABLE 1 | Published Protein Structures of Protein–CTD Complexes

Protein, CTD Binding Domain	Protein Origin	CTD Peptide Sequence Captured in the Structure	pdb Code	References
Cgt1, NT domain	*Candida albicans*	$T4_a$**$pS5_a$**$P6_aS7_aY1_bS2_bP3_bT4_b$**$pS5_b$**$P6_bS7_bY1_cS2_c$ $P3_cT4_c$**$pS5_c$**$P6_c$, $T4_a$**$pS5_a$**$P6_aS7_aY1_bS2_bP3_bT4_b$ **$pS5_b$**	1P16	Fabrega et al.[63]
MCE1, NT domain	*Mus musculus*	**$pS5_a$**$P6_aS7_aY1_b$**$pS2_b$**$P3_bT4_b$**$pS5_b$**$P6_bS7_bY1_c$**$pS2_c$**$P3_c$	3RTX	Ghosh et al.[69]
Rtt103, CID domain	*Saccharomyces cerevisiae*	$P6_aS7_aY1_b$**$pS2_b$**$P3_bT4_bS5_bP6_bS7_b$	2L0I	Lunde et al.[65]
SCAF8, CID domain	*Homo sapiens*	$Y1_a$**$pS2_a$**$P3_aT4_apS5_aP6_aS7_aY1_b$, $Y1_a$**$pS2_a$**$P3_aT4_aS5_aP6_aS7_aY1_b$**$pS2_b$**$P3_b$, $P6_aS7_aY1_bS2_bP3_bT4_b$**$pS5_b$**$P6_bS7_b$, $P6_a$**$pS7_a$**$Y1_b$ **$pS2_b$**$P3_bT4_bS5_bP6_b$, $P6_aS7_aY1_bS2_bP3_bT4_bS5_b$ $P6_bS7_b$, $P6_aS7_aY1_b$**$pS2_b$**$P3_bT4_b$**$pS5_b$**$P6_b$	3D9K, 3D9L, 3D9M, 3D9N, 3D9O, 3D9P	Becker et al.[64]
Pcf11, CID domain	*Saccharomyces cerevisiae*	$P6_aS7_aY1_b$**$pS2_b$**$P3_bT4_bS5_bP6_bS7_b$	1SZA	Meinhart and Cramer[62]
Nrd1, CID domain	*Saccharomyces cerevisiae*	$T4_a$**$pS5_a$**$P6_aS7_aY1_bS2_bP3_bT4_b$**$pS5_b$**$P6_bS7_b$	2IO6	Kubicek et al.[70]
SCP1, BRCT domain	*Homo sapiens*	$P6_aS7_aY1_b$**$pS2_b$**$P3_bT4_b$**$pS5_b$**$P6_bS7_b$, $S7_aY1_bS2_bP3_bT4_b$**$pS5_b$**$P6_bS7_b$	2GHQ, 2GHT	Zhang et al.[66]
PIN1, WW domain	*Homo sapiens*	$Y1_a$**$pS2_a$**$P3_aT4_a$**$pS5_a$**$P6_aS7_a$	1F8A	Verdecia et al.[71]
Ssu72	*Homo sapiens*	$P3_aT4_a$**$pS5_a$**$P6_aS7_aY1_b$	3O2Q	Xiang et al.[68]
Ssu72	*Drosophila melanogaster*	$P3_aT4_a$**$pS5_a$**$P6_aS7_aY1_b$	3P9Y	Werner-Allen et al.[67]

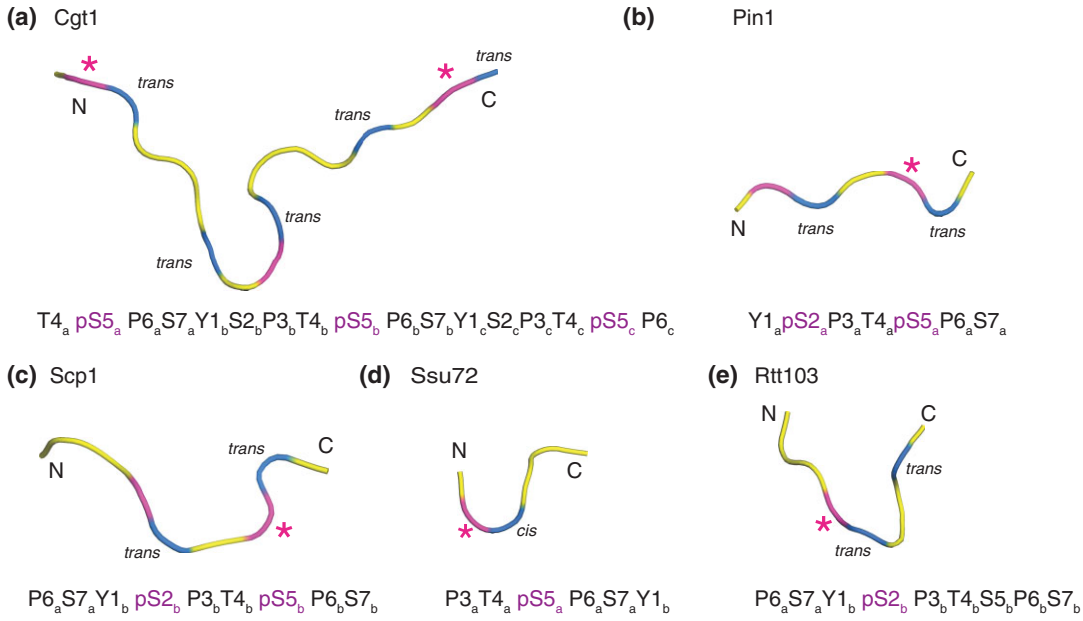


FIGURE 4 | Diversity of the C-terminal domain (CTD) backbone conformations. Cartoon representation of the CTD peptide (yellow) secondary structure upon binding with (a) Cgt1, (b) PIN1, (c) SCP1, (d) Ssu72, and (e) Rtt103 (PDB ID 1P16, 1F8A, 2GHQ, 3P9Y, and 2L0I, respectively). Primary sequence of the bound peptide is given below the cartoon representation, alphabetical subscripts indicate the sequential number of the heptads, the positions of the phosphorylated serine residue are shown in purple, asterisks indicate phosphate that is specifically recognized by the protein, the proline residues are shown in blue, *cis* and *trans* indicate the isomeric states of the serine–proline peptidyl–prolyl bonds. N, N-terminal end; C, C-terminal end.

low affinity binding of Pcf11 to the pS_2 CTD was improved by the introduction of this arginine residue at the corresponding position in Pcf11.[65] chromatin immunoprecipitation (ChIP) studies showed that the arginine mutations in Rtt103 impaired its recruitment to the 3′ end processing site *in vivo*.[65]

Fluorescence anisotropy measurements of Pcf11 and Rtt103 with CTD peptides showed that two binding events are observed for the CTD with four pS_2 repeats. The first has a higher affinity than that which is achieved for the CTD with two repeats, while the second binding event has a similar affinity as that which occurs with just two CTD repeats. NMR data showed that two copies of the Rtt103 CID and four repeats of the pS_2 CTD are tumbling as a single complex in the solution, and hence, the increased affinity can be a result of cooperative binding.[65] Several residues that are involved in domain–domain interactions were identified in Rtt103, and mutations in the residues that are affected by binding were able to reduce the affinity of the second binding event and decrease recruitment to the 3′ processing sites *in vivo*.[65] Cooperative interactions may act as an additional mechanism for CTD binding and its regulation.

In all complexes, the S–P peptidyl–prolyl bonds are in the *trans* conformation with the exception of the structure of the Nrd1 CID that is bound to the pS_5 CTD (Figure 5(c) and (d)). Nrd1 binds more residues upstream of the β-turn compared with the other CID–CTD complexes. It uses a conserved region of the CID at the N-terminal tip of helix $\alpha 2$ to specifically recognize pS_{5a} via H-bonding that is mediated by serine and arginine.[70] The structure of the Nrd1 CID in a complex with the pS_5 CTD shows that the specific recognition of pS_{5a} is facilitated by the *cis* conformation of the pS_{5a}–P_{6a} peptidyl–prolyl bond (Figure 5(d)). The *cis* conformation maximizes the intermolecular contacts and prevents the peptide from clashing with the $\alpha 1$–$\alpha 2$ loop of Nrd1.[70] Mutations at the complex interface that are associated with the *cis* conformer selection and phosphoserine interactions diminish the binding affinity and impair the processing or degradation of the non-coding RNAs. These findings demonstrate the interplay between the covalent and non-covalent changes in the CTD structure that constitute the CTD code.[70]

All of the CID–CTD complexes contain highly conserved structural features involving the Y_1 residue: the Y_1 hydroxyl group forms an H-bond with a conserved aspartate of the CID, and the aromatic ring of Y_1 is tightly accommodated in the hydrophobic pocket of the CID. Therefore, it has long been speculated that the phosphorylation

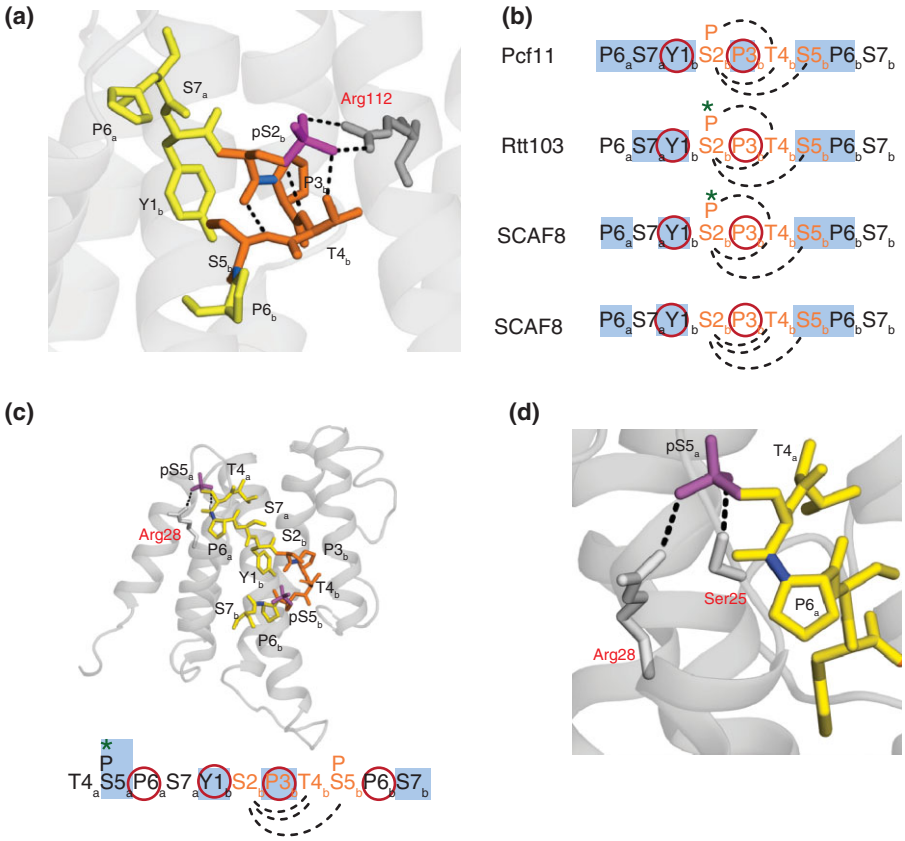


FIGURE 5 | C-terminal domain (CTD) recognition by the CTD-interacting domain (CID) domains. (a) A close-up of the CTD recognition by the SCAF8 CID (PDB ID 3D9L). (b) Schematic comparison of the CTD recognition by the CID domains of Pcf11 (PDB ID 1SZA), Rtt103 (PDB ID 2L0I), and SCAF8 (PDB ID 3D9P, 3D9O). (c) The structure of the Nrd1 CID–CTD complex (PBD ID 2IO6), the primary sequence of the bound peptide is given below the cartoon representation. (d) A close-up of the pSer5 recognition by the Nrd1 CID. The CTD sequence and the critical Arg and Ser residues of the CIDs are shown in stick representation (in gray); the CIDs are shown in transparent cartoon representation. The CTD residues forming the β-turn conformation are highlighted in orange, the phosphate group of Ser is shown in magenta, the serine–proline peptide bonds are highlighted in blue, and the dashed lines indicate H-bonds. Blue boxes indicate residues involved in the intermolecular H-bonds, dashes lines indicate residues forming the intramolecular H-bonds, green asterisks indicate a direct recognition of the phosphorylated serine, red circles indicate other types of electrostatic interactions contributing to the binding. Alphabetical subscripts indicate the sequential number of the heptads.

or other modification of Y_1 may have detrimental consequences to the binding with CIDs. Indeed, recently it was shown that the phosphorylation of Y_1 impairs the binding to all three yeast CID-containing proteins, Nrd1, Pcf11, and Rtt103, yet it stimulates the binding of elongation factor Spt6.[33]

FCP Homology Domains: Sharp Contact with β-Turn

The fission yeast Fcp1 is an essential CTD-specific serine phosphatase,[79,80] while the human small CTD phosphatase, SCP1, is involved in the regulation of neuronal gene silencing and may have other targets than the CTD.[81,82] Both SCP1 and Fcp1 belong to the family of Mg^{2+}-dependent phosphoserine/phosphothreonine-specific phosphatases.[66,83] Both enzymes are able to dephosphorylate the pS_2 and pS_5 CTD phosphoisoforms; however, SCP1 preferentially acts on the pS_5 position, whereas Fcp1 dephosphorylates mainly pS_2.[79,81] The enzymes largely differ in the overall architectures: Fcp1 consists of FCPH (FCP homology), the BRCT (BRCA1 C-terminal) domain and the TFIIF interacting helix, and all three domains interact and form a deep canyon;[83] SCP1 possesses only a single FCPH domain.[66]

The preference of SCP1 for pS_5 dephosphorylation may be rationalized by the crystal structure of the catalytically inactive SCP1 that is bound to the phosphorylated CTD.[66] Similar to the CID–CTD complexes, the CTD peptide (with pS_5 and/or pS_2) adopts a β-turn-like structure that is formed by the residues $S_{2b}P_{3b}T_{4b}pS_{5b}$ (Figure 6). In this structure,

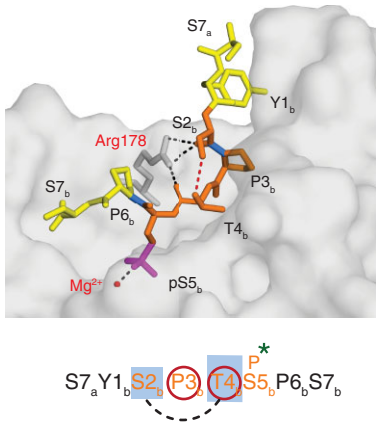


FIGURE 6 | C-terminal domain (CTD) recognition by SCP1. A close-up of the structure of the SCP1–CTD complex (PBD ID 2GHT), where the CTD sequence and the critical Arg residue of SCP1 (gray) are shown in stick representation, SCP1 is shown in transparent surface representation. The phosphate group of the Ser5 residue is shown in magenta, the serine–proline peptidyl–prolyl bonds are highlighted in blue, black dashed lines indicate the intermolecular H-bonds, red dashed lines indicate the intramolecular H-bond. The primary sequence of the bound CTD is shown below the structure. Blue boxes indicate residues involved in the intermolecular H-bonds, dashed line connects residues forming the intramolecular H-bonds, red circles indicate other types of electrostatic interactions contributing to the binding, green asterisk indicates the phosphate group that coordinates Mg^{2+} ion.

only one intramolecular H-bond between the hydroxyl groups of S_{2b} and T_{4b} is present. P_{3b} is stabilized in a hydrophobic pocket, while the S_{2b} and T_{4b} backbone carbonyls form H-bonds with Arg178. A sharp contact surface places the phosphate group of pS_5 in a position to coordinate the catalytically important Mg^{2+} ion.

The exact position of the CTD peptide within the Fcp1 crystal structure is unknown.[83] A minimally effective CTD substrate for Fcp1 consists of $S_{5a}P_{6a}S_{7a}Y_{1b}pS_{2b}P_{3b}T_{4b}$ which, in contrast with SCP1, Y_{1b}, and P_{3b} residues flanking pS_{2b}, are important for the phosphatase activity.[84] Surprisingly, the corresponding CTD interaction surface of SCP1 is blocked in the Fcp1 structure by a helical domain; thus, the fission yeast enzyme should have an alternative interaction surface. It has been proposed that the CTD peptide threads through the deep canyon of Fcp1 to access the active site.[84]

mRNA Capping Enzymes: Alternative Recognition of Same Sequence

The addition of the m7GpppN 5′ cap is the first modification of the nascent mRNA during transcription, and this capping event is tightly connected with the transcription cycle.[56,85,86] The capping enzyme interacts with the pS_5 CTD, and this requirement is conserved from yeast to mammals.[87] However, there are some differences in architecture between capping enzymes. For example, mammals encode a bifunctional enzyme possessing an N-terminal triphosphatase domain and C-terminal guanylyltransferase domain, whereas yeasts encode the triphosphatase and guanylyltransferase enzymes separately.[48,88]

There are two crystal structures capturing the interaction of the capping enzyme with the CTD, which is phosphorylated at position S_5: mouse capping enzyme MCE1[69] and *Candida albicans* RNA guanylyltransferase Cgt1.[63] The tertiary structures of both proteins consist of NT domains and oligonucleotide-binding (OB) domains, in which only the NT is interacting with the CTD. Binding to the CTD containing pS_5 is also known to stimulate the enzymatic activities of these enzymes.[69,89,90] Comparisons of the two structures show that both capping enzymes bind the CTD peptides in an extended β-like conformation. However, the enzymes use distinct CTD-binding interfaces to read the same pattern of CTD modifications. This fact corresponds to the moderate conservation of the CTD-binding site for these two capping enzymes.[69]

The structure of Cgt1 contains the longest bound stretch of the CTD peptide that has been captured to date; the crystal structure includes almost three heptads with pS_5 in each.[63] The CTD peptide spans from one side of the NT domain to the other (Figure 7(a)). The terminal phosphoserines (pS_{5a} and pS_{5c}) are anchored in two positively charged pockets, whereas the central pS_{5b} is not recognized. The majority of the CTD peptide lies in a hydrophobic cleft, and the middle heptad forms an exposed loop. It is likely that more heptads can form this loop that do not establish contacts with Cgt1 but instead serve, for example, as a platform for other CTD-interacting proteins. $Y_{1b,c}$, P_{3b}, P_{6a} and the terminal $pS_{5a,c}$ side chains interact with the binding surface, while S_{2c}, $T_{4a,b,c}$, S_{7a}, and the middle pS_{5b} are oriented away from the surface of the cleft. Because the long CTD stretch is coordinated by multiple contacts that are distributed along the Cgt1 surface, distinct single amino acid residues are unlikely to determine the specificity of the pS_5 CTD recognition. Paired mutants from both docking sites resulted in the most severe phenotypes, highlighting the necessity of and connection between the two docking sites.[63]

In contrast with Cgt1, the mammalian capping enzyme MCE1 interacts with a short doubly phosphorylated $pS_{5a}P_{6a}S_{7a}Y_{1b}pS_{2b}P_{3b}$ peptide (Figure 7(b)). The peptide is localized on a different surface of the protein compared with Cgt1.[69] Although MCE1

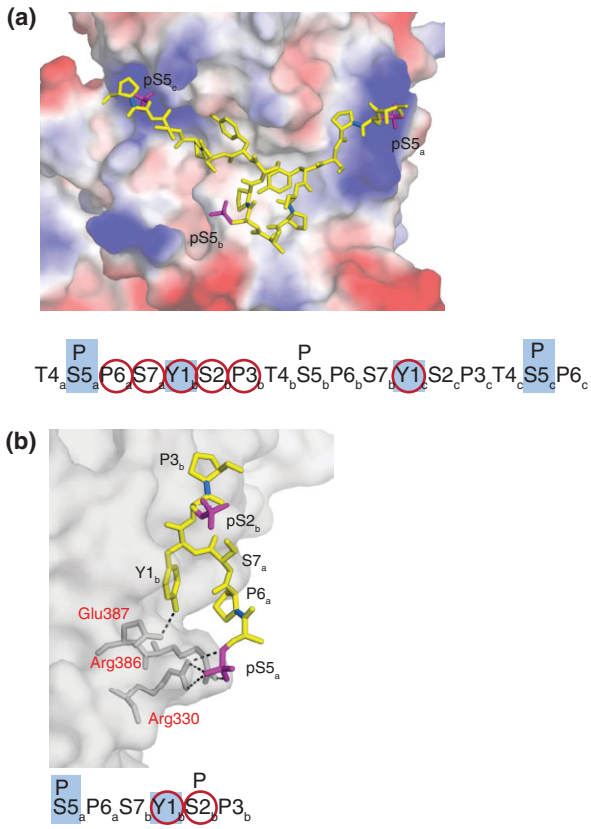


FIGURE 7 | C-terminal domain (CTD) recognition by capping enzymes. A close-up view of the structure (a) of the Cgt1–CTD (PDB ID 1P16) and (b) of the MCE1–CTD complexes (PDB ID 3RTX). The CTD residues (yellow) are shown in stick representation in which the phosphate groups of Ser5 are shown in magenta, the serine–proline peptidyl–prolyl bonds are highlighted in blue, dashed lines indicate H-bonds. Cgt1 is shown in electrostatic surface representation, MCE1 is shown in surface representation. The interacting residues of MCE1 (gray) are shown in stick representation. The primary sequences of the bound CTD are given below the corresponding structures in which blue boxes indicate residues involved in the intermolecular H-bonds, red circles indicate other types of electrostatic interactions contributing to the binding. Alphabetical subscripts indicate the sequential number of the heptads.

binds the short CTD peptide, the minimal stretch of pS_5 CTD peptide that stimulates guanylyltransferase activity *in vitro* is 18 amino acids long. The level of stimulation of the enzymatic activity increases with the number of CTD repeats.[69] Similar to the Cgt1 structure, the most important residues that are involved in the binding are Y_{1b} and pS_{5a}. pS_5 is fixed in a positively charged pocket that is formed by side chains of arginine residues, forming several H-bonds. Y_{1b} is positioned in a hydrophobic pocket, and its side-chain hydroxyl group forms an H-bond. P_{6a} does not interact with MCE1 directly but forms intramolecular VDW contacts with the Y_{1b} side chain. Correspondingly, Y1A substitution in the CTD abolishes both affinity and GTase activity stimulation, whereas P3A and P6A mutations only decrease the level of affinity and ability to stimulate GTase activity.[87]

MCE1 is able to interact with both the pS_2 CTD and pS_5 CTD, but only the pS_5 CTD stimulates enzymatic activity.[69,89] MCE1 that is bound to the doubly phosphorylated $pS_{5a}P_{6a}S_{7a}Y_{1b}pS_{2b}P_{3b}$ peptide has a pS_2 side chain that is solvent-exposed and not involved in binding.[69] Unfortunately, the crystal structure was not able to capture an alternative binding site for the pS_2 CTD that has been proposed based on competition experiments between the pS_2 CTD and pS_5 CTD peptides.[89] Interestingly, a catalytic mutant of MCE1 has shown reduced affinity to the pS_5 CTD, which can indicate a reciprocal relationship between the stimulation of the catalysis and the CTD binding by MCE1.[69]

WW Domain of PIN1: Compact Recognition of pS-P Motif

A human protein PIN1 and its yeast homolog Ess1 are peptidyl–prolyl *cis–trans* isomerases that are involved in many cellular processes, such as cell cycle regulation and pre-mRNA processing.[91–94] Moreover, PIN1 is associated with human disorders, including cancer and Alzheimer's disease.[95–97] It is known to interact with a number of phosphorylated substrates, such as Cdc25C,[98,99] microtubule associated protein tau,[100] and the CTD.[101]

PIN1 function is separated between two domains, the catalytic PPI domain and protein binding WW domain. The catalytic domain of PIN1 is able to bind pS_2- and/or pS_5-containing peptides on its own; however, the binding affinity increases by 10- to 30-fold in the presence of the WW domain.[71] WW domains are small domains that consist of approximately 40 amino acids, forming a compact triple-stranded antiparallel β-sheet.[102,103] WW domains are known for binding to proline-rich peptide sequences and are divided into several groups based on substrate preference. The IV class of the WW domain recognizes peptide sequences with phosphoserine or phosphothreonine followed by a proline residue.[104]

A crystal structure of the N-terminal WW domain of human peptidyl–prolyl *cis–trans* isomerase PIN1 bound to the CTD peptide was the first to capture the CTD interaction upon protein binding.[71] In the crystal structure, the WW domain is bound to one canonical heptad repeat of the CTD, which is phosphorylated at positions pS_{2a} and pS_{5a} (Figure 8),

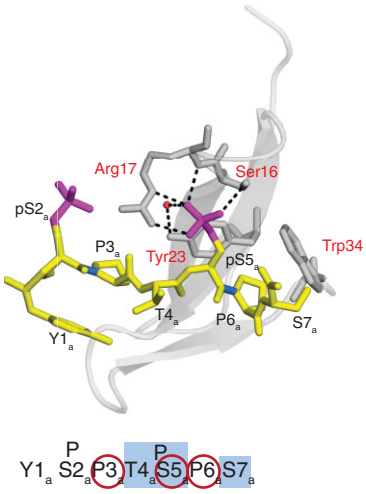


FIGURE 8 | C-terminal domain (CTD) recognition by the WW domain of human PIN1. A close-up of the structure of the PIN1–CTD complex (PBD ID 1F8A) in which the CTD sequence (yellow) and the interacting amino acid residues of PIN1 (gray) are shown in stick representation, the WW domain of PIN1 is shown in cartoon representation. The phosphate groups of the phosphorylated Ser2 and Ser5 residues are shown in magenta, the serine–proline peptidyl–prolyl bonds are highlighted in blue, dashed lines indicate H-bonds. The primary sequence of the bound CTD is shown below the structure in which blue boxes indicate residues involved in the intermolecular H-bonds, red circles indicate other types of electrostatic interactions contributing to the binding. Alphabetical subscripts indicate the sequential number of the heptads.

where only pS_{5a} may be directly recognized. The peptide is positioned in the positively charged channel between the PPIase catalytic domain and the WW domain, interacting only with the latter. The WW domain has an extended contact surface that can recognize up to five residues. However, the most extensive contacts originate from P_{3a}, pS_{5a}, and P_{6a}. A comparison of PIN1 affinities to different naturally occurring peptides revealed a consensus sequence PX(pS/pT)P, thereby confirming the contacts that were observed in the crystal structure.[71,104] The specificity of pS_5 recognition is achieved by the formation of several H-bonds between the phosphate group and Arg17 and Ser16 of the WW domain (Figure 8). R17A and S16A mutations decrease affinity by six- and threefold, respectively.[71] The lack of conservation of the Arg and Ser residues in the WW domain family suggests that only a set of WW domains can bind pS-P similar to PIN1.

Ssu72: Importance of *cis* Isomerization for Phosphatase Activity

Ssu72 belongs to the family of low molecular weight protein tyrosine phosphatases (LMW PTP). LMW PTPs are metal-independent phosphohydrolases that act via cysteinyl phosphoenzyme intermediates.[105] Ssu72 is a pS_5 CTD phosphatase that is implicated in numerous steps of the transcription cycle.[106–108]

Two recent crystal structures of human and fruit fly Ssu72 showed that Ssu72 binds to the CTD via a conformational selection mechanism in which only the *cis* conformation of the pS_5–P_6 bond of the CTD can be accommodated in the catalytic site of Ssu72[67,68] (Figure 9). Similar to the CTD binding of Nrd1, this conformational selection mechanism opens a new level of complexity for the CTD code. Owing to the *cis* conformation, the backbone of the peptide sharply turns and settles into the phosphate group of pS_{5a} inside of the enzymatic pocket. The phosphate group is strongly anchored in the active site with multiple H-bonds. Interactions include intramolecular H-bonds between T_{4a} and P_{6a}, H-bonds of pS_{5a} and S_{7a} backbone amides and Ssu72 residues. VDW, electrostatic and stacking interactions contribute to the binding of the P_{3a}, T_{4a}, P_{6a}, and Y_{1b} residues. The phosphatase activity of yeast Ssu72 was shown to be abolished when P_6 was substituted with alanine, drastically reduced with T4A and Y1A mutations, and was not effected in any other CTD mutant.[109] NMR measurements of proline *cis/trans*-isomer populations

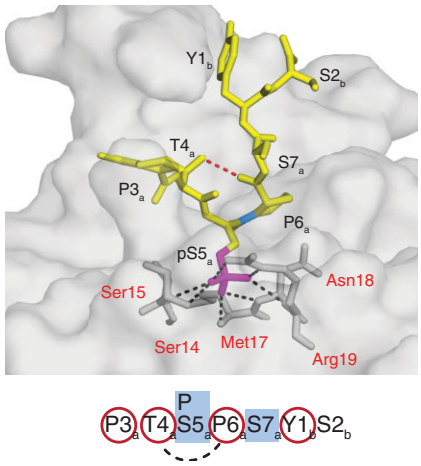


FIGURE 9 | C-terminal domain (CTD) recognition by Ssu72. A close-up of the fruit fly structure of the Ssu72–CTD complex (PBD ID 3P9Y) in which the CTD sequence (yellow) and the interacting amino acid residues of Ssu72 (gray) are shown in stick representation, Ssu72 is shown in surface representation. The phosphate group of Ser5 is shown in magenta, the serine–proline peptidyl–prolyl bonds are highlighted in blue, black dashed lines indicate the intermolecular H-bonds, red dashed lines indicate the intramolecular H-bond. The primary sequence of the bound CTD is shown below the structure in which blue boxes indicate residues involved in the intermolecular H-bonds, dashes lines indicate residues involved in the intramolecular H-bonds, red circles indicate other types of electrostatic interactions contributing to the binding. Alphabetical subscripts indicate the sequential number of the heptads.

for the NH2-$T_4pS_5P_6S_7Y_1$–COOH peptide showed that only approximately 12% of the pS_5–P_6 motifs are in the *cis* conformation,[67] greatly limiting substrate availability for Ssu72. Isomerases highlight the regulatory function of the *cis–trans* isomerization. Studies of enzyme kinetics using synthetic peptides and full-length yeast CTDs showed that the activity of the Ssu72 phosphatase is stimulated by the presence of Ess1 and PIN1.[67,68,110] Enzymes that specifically recognize the *cis* isomer of the serine–proline bond are perfect targets for regulation by isomerization. Thus, proline isomerization is a critical component of the CTD code.

Despite the established specificity of Ssu72 for the pS_5–P_6-containing CTD peptides, two recent genetic studies have suggested that Ssu72 is also a pS_7 phosphatase.[54,111] If verified *in vitro*, this would imply that Ssu72 is a less conformation-specific phosphatase (acting also upon the pS_7–Y_1 CTD peptide that is unlikely to exist in the *cis* conformation) than previously expected. Structural and mechanistic understanding for this activity will require further studies.

CONCLUSION

The concept of a CTD code that specifies the position of RNA pol II in the transcription cycle and thus recruits specific processing factors was suggested almost a decade ago. However, the attainment of knowledge regarding the structural and mechanistic bases of the CTD code have been hindered by the lack of structural data and the elements that are involved in the process. Extensive structural studies have revealed an enormous diversity of interactions within the CTD–protein complexes, suggesting that the simple rules of the CTD code are unlikely to exist. Currently available structures of the bound CTD peptides show that the specificity of interactions that spatially and temporally regulate the binding of processing factors originates in different modification isoforms, modification patterns, lengths of bound CTD peptides, and *cis* or *trans* conformations of the phosphoserine–proline peptidyl–prolyl bonds. Recent studies have demonstrated that these CTD modifications not only promote the binding of required factors but also block the binding of other factors that are incompatible with a given stage of the transcription cycle. Furthermore, the CTD interactions with effector molecules are accompanied by large structural rearrangements or they undergo conformational selection. The emerging importance of CTD plasticity requires future dynamic studies, including those at the single molecule level. By combining the structural and dynamic data of the CTD interactions, it may soon be possible to decipher how the CTD code is written, read, and erased. As a consequence, transcription and co-transcriptional processing could be understood at the atomic level.

ACKNOWLEDGMENTS

This work was supported by the project 'CEITEC - Central European Institute of Technology' (CZ.1.05/1.1.00/02.0068) from the European Regional Development Fund and the Czech Science Foundation (P305/12/G034 and P305/10/1490).

REFERENCES

1. van Dijk EL, Chen CL, d' Aubenton-Carafa Y, Gourvennec S, Kwapisz M, Roche V, Bertrand C, Silvain M, Legoix-Né P, Loeillet S, et al. XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature* 2011, 475:114–117.
2. Neil H, Malabat C, d' Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* 2009, 457:1038–1042.
3. Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Munster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM. Bidirectional promoters generate pervasive transcription in yeast. *Nature* 2009, 457:1033–1037.
4. Tisseur M, Kwapisz M, Morillon A. Pervasive transcription - lessons from yeast. *Biochimie* 2011, 93:1889–1896.
5. Werner M, Thuriaux P, Soutourina J. Structure-function analysis of RNA polymerases I and III. *Curr Opin Struct Biol* 2009, 19:740–745.
6. Meinhart A, Kamenski T, Hoeppner S, Baumli S, Cramer P. A structural perspective of CTD function. *Genes Dev* 2005, 19:1401–1415.
7. Phatnani HP, Greenleaf AL. Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev* 2006, 20:2922–2936.

8. Muñoz MJ, de la Mata M, Kornblihtt AR. The carboxy terminal domain of RNA polymerase II and alternative splicing. *Trends Biochem Sci* 2010, 35:497–504.
9. Egloff S, Murphy S. Role of the C-terminal domain of RNA polymerase II in expression of small nuclear RNA genes. *Biochem Soc Trans* 2008, 36:537.
10. Hirose Y, Ohkuma Y. Phosphorylation of the C-terminal domain of RNA polymerase II plays central roles in the integrated events of eucaryotic gene expression. *J Biochem* 2007, 141:601–608.
11. Buratowski S. The CTD code. *Nat Struct Mol Biol* 2003, 10:679–680.
12. Chapman RD, Heidemann M, Hintermair C, Eick D. Molecular evolution of the RNA polymerase II CTD. *Trends Genet* 2008, 24:289–296.
13. Liu P, Kenney JM, Stiller JW, Greenleaf AL. Genetic organization, length conservation, and evolution of rna polymerase ii carboxyl-terminal domain. *Mol Biol Evol* 2010, 27:2628–2641.
14. Stiller JW, McConaughy BL, Hall BD. Evolutionary complementation for polymerase II CTD function. *Yeast* 2000, 16:57–64.
15. West ML, Corden JL. Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations. *Genetics* 1995, 140:1223–1233.
16. Schwer B, Shuman S. Deciphering the RNA polymerase II CTD code in fission yeast. *Mol Cell* 2011, 43:311–318.
17. Coudreuse D, van Bakel H, Dewez M, Soutourina J, Parnell T, Vandenhaute J, Cairns B, Werner M, Hermand D. A gene-specific requirement of RNA polymerase II CTD phosphorylation for sexual differentiation in S. pombe. *Curr Biol* 2010, 20:1053–1064.
18. Akhtar MS, Heidemann M, Tietjen J, Zhang D, Chapman RD, Eick D, Ansari AZ. TFIIH kinase places bivalent marks on the carboxyl-terminal domain of RNA polymerase II. *Mol Cell* 2009, 34:387–393.
19. Kim M, Suh H, Cho E-J, Buratowski S. Phosphorylation of the Yeast Rpb1 C-terminal Domain at Serines 2, 5, and 7. *J Biol Chem* 2009, 284:26421–26426.
20. Kim H, Erickson B, Luo W, Seward D, Graber JH, Pollock DD, Megee PC, Bentley DL. Gene-specific RNA polymerase II phosphorylation and the CTD code. *Nat Struct Mol Biol* 2010, 17:1279–1286.
21. Tietjen JR, Zhang DW, Rodriguez-Molina JB, White BE, Akhtar MS, Heidemann M, Li X, Chapman RD, Shokat K, Keles S, et al. Chemical-genomic dissection of the CTD code. *Nat Struct Mol Biol* 2010, 17:1154–1161.
22. Chapman RD, Heidemann M, Albert TK, Mailhammer R, Flatley A, Meisterernst M, Kremmer E, Eick D. Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. *Science* 2007, 318:1780–1782.
23. Egloff S, O'Reilly D, Chapman RD, Taylor A, Tanzhaus K, Pitts L, Eick D, Murphy S. Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. *Science* 2007, 318:1777–1779.
24. Egloff S, Szczepaniak SA, Dienstbier M, Taylor A, Knight S, Murphy S. The integrator complex recognizes a new double mark on the RNA polymerase II carboxyl-terminal domain. *J Biol Chem* 2010, 285:20564–20569.
25. Egloff S, Zaborowska J, Laitem C, Kiss T, Murphy S. Ser7 phosphorylation of the CTD recruits the RPAP2 Ser5 phosphatase to snRNA genes. *Mol Cell* 2012, 45:111–122.
26. Stiller JW, Cook MS. Functional unit of the RNA polymerase II C-terminal domain lies within Heptapeptide pairs. *Eukaryot Cell* 2004, 3:735–740.
27. Liu P, Greenleaf AL, Stiller JW. The essential sequence elements required for RNAP II carboxyl-terminal domain function in yeast and their evolutionary conservation. *Mol Biol Evol* 2008, 25:719–727.
28. Maniatis T, Reed R. An extensive network of coupling among gene expression machines. *Nature* 2002, 416:499–506.
29. Moore MJ, Proudfoot NJ. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* 2009, 136:688–700.
30. Hsin J-P, Sheth A, Manley JL. RNAP II CTD phosphorylated on threonine-4 is required for histone mRNA 3′ end processing. *Science* 2011, 334:683–686.
31. Hintermair C, Heidemann M, Koch F, Descostes N, Gut M, Gut I, Fenouil R, Ferrier P, Flatley A, Kremmer E, et al. Threonine-4 of mammalian RNA polymerase II CTD is targeted by Polo-like kinase 3 and required for transcriptional elongation. *EMBO J* 2012, 31:2784–2797.
32. Baskaran R, Dahmus ME, Wang JY. Tyrosine phosphorylation of mammalian RNA polymerase II carboxyl-terminal domain. *Proc Natl Acad Sci U S A* 1993, 90:11167–11171.
33. Mayer A, Heidemann M, Lidschreiber M, Schreieck A, Sun M, Hintermair C, Kremmer E, Eick D, Cramer P. CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. *Science* 2012, 336:1723–1725.
34. Kelly WG, Dahmus ME, Hart GW. RNA polymerase II is a glycoprotein. Modification of the COOH-terminal domain by O-GlcNAc. *J Biol Chem* 1993, 268:10416–10424.
35. Comer FI, Hart GW. Reciprocity between O-GlcNAc and O-phosphate on the carboxyl terminal domain of RNA polymerase II. *Biochemistry* 2001, 40:7845–7852.

36. Ranuncolo SM, Ghosh S, Hanover JA, Hart GW, Lewis BA. Evidence of the involvement of O-GlcNAc-modified human RNA polymerase II CTD in transcription in vitro and in vivo. *J Biol Chem* 2012, 287:23549–23561.
37. Sims RJ, Rojas LA, Beck D, Bonasio R, Schüller R, Drury WJ, Eick D, Reinberg D. The C-Terminal domain of RNA polymerase II is modified by site-specific methylation. *Science* 2011, 332:99–103.
38. Li H, Zhang Z, Wang B, Zhang J, Zhao Y, Jin Y. Wwp2-mediated ubiquitination of the RNA polymerase II large subunit in mouse embryonic pluripotent stem cells. *Mol Cell Biol* 2007, 27:5296–5305.
39. Lu KP, Finn G, Lee TH, Nicholson LK. Prolyl cis-trans isomerization as a molecular timer. *Nat Chem Biol* 2007, 3:619–629.
40. Shaw PE. Peptidyl-prolyl isomerases: a new twist to transcription. *EMBO Rep* 2002, 3:521–526.
41. Shaw PE. Peptidyl-prolyl cis/trans isomerases and transcription: is there a twist in the tail? *EMBO Rep* 2007, 8:40–45.
42. Fanghänel J, Fischer G. Insights into the catalytic mechanism of peptidyl prolyl cis/trans isomerases. *Front Biosci* 2004, 9:3453–3478.
43. Egloff S, Dienstbier M, Murphy S. Updating the RNA polymerase CTD code: adding gene-specific layers. *Trends Genet* 2012, 7:333–341.
44. Buratowski S. Progression through the RNA polymerase II CTD cycle. *Mol Cell* 2009, 36:541–546.
45. Borggrefe T, Yue X. Interactions between subunits of the mediator complex with gene-specific transcription factors. *Sem Cell Dev Biol* 2011, 22:759–768.
46. Myers LC, Kornberg RD. Mediator of transcriptional regulation. *Annu Rev Biochem* 2000, 69:729–749.
47. Max T, Søgaard M, Svejstrup JQ. Hyperphosphorylation of the C-terminal repeat domain of RNA polymerase II facilitates dissociation of its complex with mediator. *J Biol Chem* 2007, 282:14113–14120.
48. Gu M, Lima CD. Processing the message: structural insights into capping and decapping mRNA. *Curr Opin Struct Biol* 2005, 15:99–106.
49. Hampsey M, Reinberg D. Tails of intrigue: phosphorylation of RNA polymerase II mediates histone methylation. *Cell* 2003, 113:429–432.
50. Brès V, Yoh SM, Jones KA. The multi-tasking P-TEFb complex. *Curr Opin Cell Biol* 2008, 20:334–340.
51. Ahn SH, Kim M, Buratowski S. Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3′ end processing. *Mol Cell* 2004, 13:67–76.
52. Licatalosi DD, Geiger G, Minet M, Schroeder S, Cilli K, McNeil JB, Bentley DL. Functional interaction of yeast pre-mRNA 3′ end processing factors with RNA polymerase II. *Mol Cell* 2002, 9:1101–1111.
53. Rosonina E, Kaneko S, Manley JL. Terminating the transcript: breaking up is hard to do. *Genes Dev* 2006, 20:1050–1056.
54. Bataille AR, Jeronimo C, Jacques P-É, Laramée L, Fortin M-È, Forest A, Bergeron M, Hanes SD, Robert F. A Universal RNA polymerase II CTD cycle is orchestrated by complex interplays between kinase, phosphatase, and isomerase enzymes along genes. *Mol Cell* 2012, 45:158–170.
55. Mayer A, Lidschreiber M, Siebert M, Leike K, Soding J, Cramer P. Uniform transitions of the general RNA polymerase II transcription complex. *Nat Struct Mol Biol* 2010, 17:1272–1278.
56. Cramer P, Srebrow A, Kadener S, Werbajh S, de la Mata M, Melen G, Nogués G, Kornblihtt AR. Coordination between transcription and pre-mRNA processing. *FEBS Letters* 2001, 498:179–182.
57. Cagas PM, Corden JL. Structural studies of a synthetic peptide derived from the carboxyl-terminal domain of RNA polymerase II. *Proteins* 1995, 21:149–160.
58. Kumaki Y, Matsushima N, Yoshida H, Nitta K, Hikichi K. Structure of the YSPTSPS repeat containing two SPXX motifs in the CTD of RNA polymerase II: NMR studies of cyclic model peptides reveal that the SPTS turn is more stable than SPSY in water. *Biochim Biophys Acta* 2001, 1548:81–93.
59. Bienkiewicz EA, Moon Woody A-Y, Woody RW. Conformation of the RNA polymerase II C-terminal domain: circular dichroism of long and short fragments. *J Mol Biol* 2000, 297:119–133.
60. Dobbins JR, Murali N, Long EC. Structural redesign and stabilization of the overlapping tandem β-turns of RNA polymerase II. *Int J Pept Protein Res* 1996, 47:260–268.
61. Gerber H-P, Hagmann M, Seipel K, Georgiev O, West MAL, Litingtung Y, Schaffner W, Corden JL. RNA polymerase II C-terminal domain required for enhancer-driven transcription. *Nature* 1995, 374:660–662.
62. Meinhart A, Cramer P. Recognition of RNA polymerase II carboxy-terminal domain by 3′-RNA-processing factors. *Nature* 2004, 430:223–226.
63. Fabrega C, Shen V, Shuman S, Lima CD. Structure of an mRNA capping enzyme bound to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Mol Cell* 2003, 11:1549–1561.
64. Becker R, Loll B, Meinhart A. Snapshots of the RNA processing factor SCAF8 bound to different phosphorylated forms of the carboxyl-terminal domain of RNA polymerase II. *J Biol Chem* 2008, 283:22659–22669.
65. Lunde BM, Reichow SL, Kim M, Suh H, Leeper TC, Yang F, Mutschler H, Buratowski S, Meinhart A, Varani G. Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain. *Nat Struct Mol Biol* 2010, 17:1195–1201.

66. Zhang Y, Kim Y, Genoud N, Gao J, Kelly JW, Pfaff SL, Gill GN, Dixon JE, Noel JP. Determinants for dephosphorylation of the RNA polymerase II C-terminal domain by Scp1. *Mol Cell* 2006, 24:759–770.

67. Werner-Allen JW, Lee C-J, Liu P, Nicely NI, Wang S, Greenleaf AL, Zhou P. cis-proline-mediated Ser(P)5 dephosphorylation by the RNA polymerase II C-terminal domain phosphatase Ssu72. *J Biol Chem* 2011, 286:5717–5726.

68. Xiang K, Nagaike T, Xiang S, Kilic T, Beh MM, Manley JL, Tong L. Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex. *Nature* 2010, 467:729–733.

69. Ghosh A, Shuman S, Lima CD. Structural insights to how mammalian capping enzyme reads the CTD code. *Mol Cell* 2011, 43:299–310.

70. Kubicek K, Cerna H, Holub P, Pasulka J, Hrossova D, Loehr F, Hofr C, Vanacova S, Stefl R. Serine phosphorylation and proline isomerization in RNAP II CTD control recruitment of Nrd1. *Genes Dev* 2012, 26:1891–1896.

71. Verdecia MA, Bowman ME, Lu KP, Hunter T, Noel JP. Structural basis for phosphoserine-proline recognition by group IV WW domains. *Nat Struct Mol Biol* 2000, 7:639–643.

72. Kim M, Krogan NJ, Vasiljeva L, Rando OJ, Nedea E, Greenblatt JF, Buratowski S. The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature* 2004, 432:517–522.

73. Patturajan M, Wei X, Berezney R, Corden JL. A nuclear matrix protein interacts with the phosphorylated C-terminal domain of RNA polymerase II. *Mol Cell Biol* 1998, 18:2406–2415.

74. Barillà D, Lee BA, Proudfoot NJ. Cleavage/polyadenylation factor IA associates with the carboxyl-terminal domain of RNA polymerase II in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* 2001, 98:445–450.

75. Sadowski M, Dichtl B, Hübner W, Keller W. Independent functions of yeast Pcf11p in pre-mRNA 3′ end processing and in transcription termination. *EMBO J* 2003, 22:2167–2177.

76. Steinmetz EJ, Brow DA. Control of pre-mRNA accumulation by the essential yeast protein Nrd1 requires high-affinity transcript binding and a domain implicated in RNA polymerase II association. *Proc Nat Acad Sci* 1998, 95:6699–6704.

77. Vasiljeva L, Kim M, Mutschler H, Buratowski S, Meinhart A. The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. *Nat Struct Mol Biol* 2008, 15:795–804.

78. Noble CG, Hollingworth D, Martin SR, Ennis-Adeniran V, Smerdon SJ, Kelly G, Taylor IA, Ramos A. Key features of the interaction between Pcf11 CID and RNA polymerase II CTD. *Nat Struct Mol Biol* 2005, 12:144–151.

79. Hausmann S, Shuman S. Characterization of the CTD phosphatase Fcp1 from fission yeast. *J Biol Chem* 2002, 277:21213–21220.

80. Suh M-H, Ye P, Zhang M, Hausmann S, Shuman S, Gnatt AL, Fu J. Fcp1 directly recognizes the C-terminal domain (CTD) and interacts with a site on RNA polymerase II distinct from the CTD. *Proc Natl Acad Sci U S A* 2005, 102:17314–17319.

81. Yeo M, Lin PS, Dahmus ME, Gill GN. A novel RNA polymerase II C-terminal domain phosphatase that preferentially dephosphorylates serine 5. *J Biol Chem* 2003, 278:26078–26085.

82. Wrighton KH, Willis D, Long J, Liu F, Lin X, Feng X-H. Small C-terminal domain phosphatases dephosphorylate the regulatory linker regions of Smad2 and Smad3 to enhance transforming growth factor-β signaling. *J Biol Chem* 2006, 281:38365–38375.

83. Ghosh A, Shuman S, Lima CD. The structure of Fcp1, an essential RNA polymerase II CTD phosphatase. *Mol Cell* 2008, 32:478–490.

84. Hausmann S, Erdjument-Bromage H, Shuman S. Schizosaccharomyces pombe carboxyl-terminal domain (CTD) phosphatase Fcp1: distributive mechanism, minimal CTD substrate, and active site mapping. *J Biol Chem* 2004, 279:10892–10900.

85. Zorio DAR, Bentley DL. The link between mRNA processing and transcription: communication works both ways. *Exp Cell Res* 2004, 296:91–97.

86. Proudfoot NJ, Furger A, Dye MJ. Integrating mRNA processing with transcription. *Cell* 2002, 108:501–512.

87. Pei Y, Hausmann S, Ho CK, Schwer B, Shuman S. The length, phosphorylation state, and primary structure of the RNA polymerase II carboxyl-terminal domain dictate interactions with mRNA capping enzymes. *J Biol Chem* 2001, 276:28075–28082.

88. Shuman S. Structure, mechanism, and evolution of the mRNA capping apparatus. *Prog Nucleic Acid Res Mol Biol* 2001, 66:1–40.

89. Ho CK, Shuman S. Distinct roles for CTD Ser-2 and Ser-5 phosphorylation in the recruitment and allosteric activation of mammalian mRNA capping enzyme. *Mol Cell* 1999, 3:405–411.

90. Wen Y, Shatkin AJ. Transcription elongation factor hSPT5 stimulates mRNA capping. *Genes Dev* 1999, 13:1774–1779.

91. Lippens G, Landrieu I, Smet C. Molecular mechanisms of the phospho-dependent prolyl cis/trans isomerase Pin1. *FEBS J* 2007, 274:5211–5222.

92. Joseph JD, Yeh ES, Swenson KI, Winkler MAR. The peptidyl-prolyl isomerase Pin1. *Prog Cell Cycle Res* 2003, 5:477–487.

93. Wulf G, Finn G, Suizu F, Lu KP. Phosphorylation-specific prolyl isomerization: is there an underlying theme? *Nat Cell Biol* 2005, 7:435–441.
94. Xu Y-X, Manley JL. Pin1 modulates RNA polymerase II activity during the transcription cycle. *Genes Dev* 2007, 21:2950–2962.
95. Wulf G, Ryo A, Liou Y-C, Lu KP. The prolyl isomerase Pin1 in breast development and cancer. *Breast Cancer Res* 2003, 5:76–82.
96. Lu KP. Pinning down cell signaling, cancer and Alzheimer's disease. *Trend Biochem Sci* 2004, 29:200–209.
97. Butterfield DA, Abdul HM, Opii W, Newman SF, Joshi G, Ansari MA, Sultana R. Review: Pin1 in Alzheimer's disease. *J Neurochem* 2006, 98:1697–1706.
98. Hutchins JRA, Clarke PR. Many fingers on the mitotic trigger: post-translational regulation of the Cdc25C phosphatase. *Cell Cycle* 2004, 3:41–45.
99. Zhou XZ, Kops O, Werner A, Lu P-J, Shen M, Stoller G, Küllertz G, Stark M, Fischer G, Lu KP. Pin1-dependent prolyl isomerization regulates dephosphorylation of Cdc25C and Tau proteins. *Mol Cell* 2000, 6:873–883.
100. Koren J, Jinwal UK, Davey Z, Kiray J, Arulselvam K, Dickey CA. Bending tau into shape: the emerging role of peptidyl-prolyl isomerases in tauopathies. *Mol Neurobiol* 2011, 44:65–70.
101. Morris DP, Phatnani HP, Greenleaf AL. Phospho-carboxyl-terminal domain binding and the role of a prolyl isomerase in Pre-mRNA 3′-end formation. *J Biol Chem* 1999, 274:31583–31587.
102. Macias MJ, Wiesner S, Sudol M. WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. *FEBS Lett* 2002, 513:30–37.
103. Ilsley JL, Sudol M, Winder SJ. The WW domain: linking cell signalling to the membrane cytoskeleton. *Cell Signal* 2002, 14:183–189.
104. Sudol M, Hunter T. NeW wrinkles for an old domain. *Cell* 2000, 103:1001–1004.
105. Meinhart A, Silberzahn T, Cramer P. The mRNA transcription/processing factor Ssu72 is a potential tyrosine phosphatase. *J Biol Chem* 2003, 278:15917–15921.
106. Dichtl B, Blank D, Ohnacker M, Friedlein A, Roeder D, Langen H, Keller W. A role for SSU72 in balancing RNA polymerase II transcription elongation and termination. *Mol Cell* 2002, 10:1139–1150.
107. Krishnamurthy S, He X, Reyes-Reyes M, Moore C, Hampsey M. Ssu72 Is an RNA polymerase II CTD phosphatase. *Mol Cell* 2004, 14:387–394.
108. Steinmetz EJ, Brow DA. Ssu72 protein mediates both poly(A)-coupled and poly(a)-independent termination of RNA polymerase II transcription. *Mol Cell Biol* 2003, 23:6339–6349.
109. Hausmann S, Koiwa H, Krishnamurthy S, Hampsey M, Shuman S. Different strategies for carboxyl-terminal domain (CTD) recognition by Serine 5-specific CTD phosphatases. *J Biol Chem* 2005, 280:37681–37688.
110. Singh N, Ma Z, Gemmill T, Wu X, DeFiglio H, Rossettini A, Rabeler C, Beane O, Morse R, Palumbo MJ, et al. The Ess1 prolyl isomerase is required for transcription termination of small non-coding RNAs via the Nrd1 pathway. *Mol Cell* 2009, 36:255–266.
111. Zhang DW, Rodríguez-Molina JB, Tietjen JR, Nemec CM, Ansari AZ. Emerging views on the CTD code. *Genet Res Int* 2012, 2012:347214.

Serine phosphorylation and proline isomerization in RNAP II CTD control recruitment of Nrd1

Karel Kubicek, Hana Cerna, Peter Holub, et al.

Genes Dev. 2012 26: 1891-1896 originally published online August 14, 2012
Access the most recent version at doi:10.1101/gad.192781.112

RESEARCH COMMUNICATION

Serine phosphorylation and proline isomerization in RNAP II CTD control recruitment of Nrd1

Karel Kubicek,[1,3] Hana Cerna,[1,3] Peter Holub,[1] Josef Pasulka,[1] Dominika Hrossova,[1] Frank Loehr,[2] Ctirad Hofr,[1] Stepanka Vanacova,[1,4] and Richard Stefl[1,4]

[1]CEITEC-Central European Institute of Technology, Masaryk University, Brno, 62500, Czech Republic; [2]Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, Goethe-University Frankfurt, 60438 Frankfurt am Main, Germany

Recruitment of appropriate RNA processing factors to the site of transcription is controlled by post-translational modifications of the C-terminal domain (CTD) of RNA polymerase II (RNAP II). Here, we report the solution structure of the Ser5 phosphorylated (pSer5) CTD bound to Nrd1. The structure reveals a direct recognition of pSer5 by Nrd1 that requires the *cis* conformation of the upstream pSer5–Pro6 peptidyl-prolyl bond of the CTD. Mutations at the complex interface diminish binding affinity and impair processing or degradation of noncoding RNAs. These findings underpin the interplay between covalent and noncovalent changes in the CTD structure that constitute the CTD code.

Supplemental material is available for this article.

Received March 26, 2012; revised version accepted July 12, 2012.

The C-terminal domain (CTD) of the largest subunit of RNA polymerase II (RNAP II) consists of multiple tandem repeats of the heptapeptide consensus Tyr1–Ser2–Pro3–Thr4–Ser5–Pro6–Ser7 that is conserved from yeast to humans. The CTD is essential and forms a flexible tail of RNAP II. It serves as a binding platform for various cotranscriptional processing factors (Hirose and Manley 2000; Maniatis and Reed 2002; Meinhart et al. 2005). Phosphorylation and dephosphorylation of Ser2, Ser5, and Ser7 create a unique pattern in coordination with the transcription cycle (Komarnitsky et al. 2000; Meinhart et al. 2005; Phatnani and Greenleaf 2006; Kim et al. 2010; Mayer et al. 2010; Tietjen et al. 2010; Bataille et al. 2012). This phosphorylation pattern, often called the CTD code (Buratowski 2003, 2009; Chapman et al. 2008; Egloff and Murphy 2008), controls the recruitment, activation, and displacement of various factors involved in transcription and RNAP II transcript processing (Meinhart et al. 2005). It has been proposed that the CTD code is also affected by noncovalent changes in the CTD structure, such as peptidyl-prolyl bond isomerization (Buratowski 2003); however, the structural basis for such tuning of the CTD code and its role in recruitment of RNA processing factors are not yet fully understood. The importance of the CTD isomerization for Ser5 dephosphorylation has recently been demonstrated for the human and insect Ssu72 phosphatases (Xiang et al. 2010; Werner-Allen et al. 2011).

The pSer5 marks occur predominantly in the early elongation phase and are essential in 3′ end processing of short noncoding genes (Komarnitsky et al. 2000; Gudipati et al. 2008; Vasiljeva et al. 2008a). These marks are specifically recognized by Nrd1 and are required for the Nrd1-dependent termination pathway, used at small nuclear/nucleolar RNAs (sn/snoRNAs), cryptic unstable transcripts (CUTs), and other short RNAP II transcripts (Steinmetz et al. 2001; Arigo et al. 2006; Thiebaut et al. 2006; Vasiljeva et al. 2008a). Recent genome-wide studies in yeast demonstrated the co-occurrence of pSer5 and pSer7 marks at some genes in the early elongation phase (Kim et al. 2010; Mayer et al. 2010; Tietjen et al. 2010). In addition to specific phosphorylation, the Nrd1 termination pathway requires the Ess1 (Pin1 in humans) peptidyl-prolyl isomerase (Singh et al. 2009) that specifically isomerizes the pSer5–Pro6 peptidyl-prolyl bond in the CTD (Gemmill et al. 2005). In yeast, Ess1 stimulates dephosphorylation of pSer5–Pro6 in vivo (Singh et al. 2009), and therefore it has been hypothesized that it regulates the Nrd1 association with the CTD (Singh et al. 2009). To gain insights into the recruitment process of Nrd1 to the 5′ regions of genes, we determined the solution structure of the *Saccharomyces cerevisiae* Nrd1 CTD-interacting domain (CID) in complex with a CTD peptide phosphorylated at Ser5. The Nrd1 CID structure reveals a conserved CTD-binding site that engages the β-turn motif of the CTD formed by $Ser2_b$–$Pro3_b$–$Thr4_b$–$pSer5_b$ and a site recognizing selectively the upstream $pSer5_a$ and the *cis* conformation of the $pSer5_a$–$Pro6_a$ peptidyl-prolyl bond of the CTD. Furthermore, we show that the specific recognition of pSer5 CTD by Nrd1 CID is important for the processing and degradation of noncoding RNAs (ncRNAs) in vivo.

Results and Discussion

Affinity of Nrd1 to the CTD with 'early' phosphorylation marks

A previous study demonstrated that Nrd1 favors binding to the CTD with "early" pSer5 marks over the CTD with "late" pSer2 marks (Vasiljeva et al. 2008a). To test the effect of the unphosphorylated CTD and the CTD with the "early" pSer7 and pSer5 CTD marks on the affinity to Nrd1, we performed a quantitative solution-binding assay using fluorescence anisotropy (FA) experiments. We found that Nrd1 binds the pSer5 CTD with a significantly stronger affinity compared with the pSer7 CTD or unphosphorylated CTD (Fig. 1A). Nrd1 shows also only a slightly weaker binding to the doubly phosphorylated pSer5/7 CTD than to the pSer5 CTD (Fig. 1A),

[*Keywords*: RNA polymerase II; CTD code; phosphorylation; proline isomerization; RNA processing and degradation; NMR spectroscopy; structure]
[3]**These authors contributed equally to this work.**
[4]**Corresponding author**
E-mail richard.stefl@ceitec.muni.cz
E-mail vanacova@chemi.muni.cz
Article published online ahead of print. Article and publication date are online at http://www.genesdev.org/cgi/doi/10.1101/gad.192781.112.

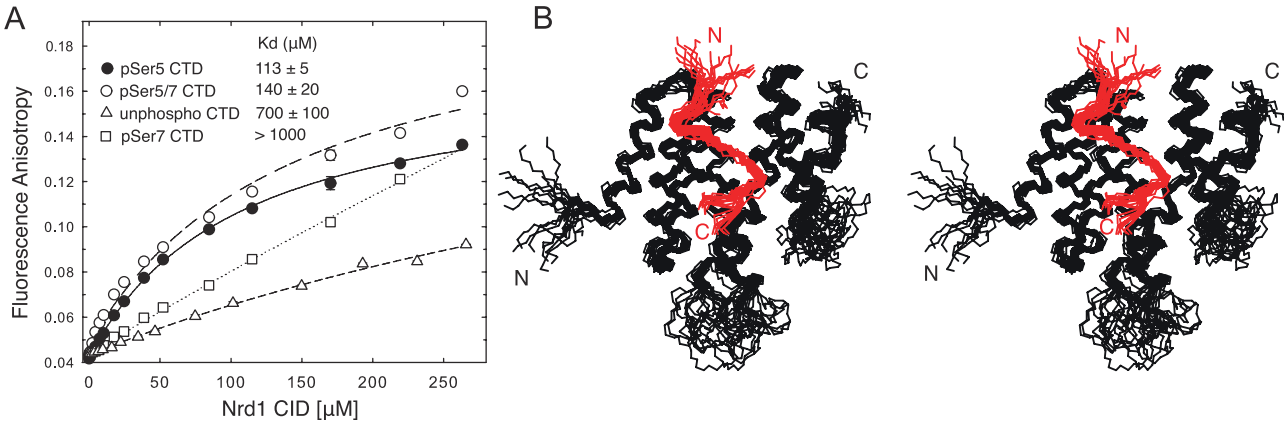


Figure 1. Structure of the Nrd1 CID–pSer5 CTD complex. (*A*) Equilibrium binding of the Nrd1 CID with differently phosphorylated CTD peptides monitored by FA. Binding isotherms and dissociation constants (K_d) are shown. (*B*) Overlay of the 20 lowest-energy structures of the Nrd1 CID–pSer5 CTD complex shown in stereo view. The backbone of the Nrd1 CID and pSer5 CTD is shown in black and red, respectively.

suggesting that the co-occurrence of the pSer5 and pSer7 marks may not impair the recruitment of the Nrd1 complex in the early elongation phase.

Structure of Nrd1 CID bound to the phosphorylated CTD

To understand how the pSer5 CTD is recognized by Nrd1, we determined the solution structure of a reconstituted complex consisting of the CID (residues 1–153) of Nrd1 and a 14-amino-acid peptide, the pSer5 CTD (two repeats of the heptapeptide CTD consensus phosphorylated at Ser5; $Tyr1_a$–$Ser2_a$–$Pro3_a$–$Thr4_a$–$pSer5_a$–$Pro6_a$–$Ser7_a$–$Tyr1_b$–$Ser2_b$–$Pro3_b$–$Thr4_b$–$pSer5_b$–$Pro6_b$–$Ser7_b$) (Fig. 1B; Supplemental Table S1; Supplemental Fig. S1). The ^1H, ^{13}C, and ^{15}N chemical shift assignments for the bound Nrd1 CID were obtained as described previously (Kubicek et al. 2011). The structure of the Nrd1 CID is formed by eight α helices in a right-handed superhelical arrangement (Fig. 1B) and is virtually identical to the structure of the Nrd1 CID in the free form (Vasiljeva et al. 2008a). The pSer5 CTD peptide contacts helices α2, α4, and α7 of the Nrd1 CID (Fig. 2A).

Recognition of the phosphorylated CTD by Nrd1

The CTD peptide adopts a β-turn conformation at $Ser2_b$–$Pro3_b$–$Thr4_b$–$pSer5_b$ and docks into a hydrophobic pocket of the Nrd1 CID that is formed by Ile29, Tyr67, Leu127, Ile130, and Met126 using $Tyr1_b$ and $Pro3_b$ residues (Fig. 2A,B). The hydroxyl group of $Tyr1_b$ forms a hydrogen bond with a conserved aspartate (Asp70) of Nrd1 (Fig. 2A,B). The binding mode of the peptide at the β-turn conformation resembles other previously determined structures of the CTD bound to CIDs of Pcf11, SCAF8, and Rtt103 (Supplemental Fig. S2; Meinhart and Cramer 2004; Becker et al. 2008; Lunde et al. 2010). However, in contrast to these CID–CTD complexes, Nrd1 binds more residues upstream of the pSer5 CTD via a conserved region at the N-terminal tip of helix α2 (Fig. 2A; Supplemental Fig. S3A). This unique region of Nrd1 is used to specifically recognize $pSer5_a$ via hydrogen bonding of Ser25 and Arg28 to the phosphate group of $pSer5_a$ (Fig. 2A,B). Another region that is more upstream in the α1–α2 loop has been previously suggested as the phosphoserine-binding site of Nrd1 based on a sulfate ion that was found in the crystal structure of the free Nrd1 CID (Vasiljeva et al. 2008a). The sulfate ion located ~8 Å away from

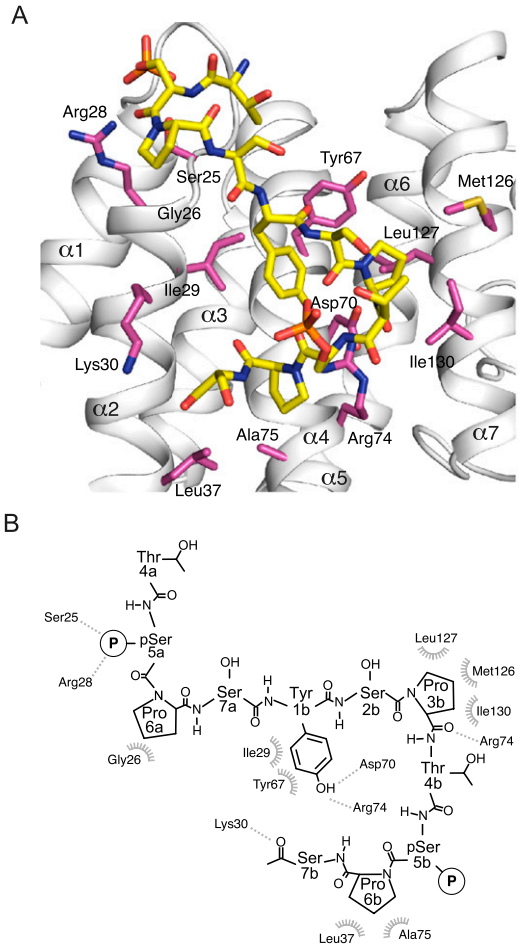


Figure 2. Recognition of the pSer5 CTD by Nrd1. (*A*) Scheme showing contacts between the Nrd1 CID and the CTD peptide. Protein residues that form hydrogen bonds and hydrophobic contacts to the CTD peptide are shown in white sticks. (*B*) Scheme showing contacts between the Nrd1 CID and the CTD peptide.

the phosphoserine location identified in our structure is coordinated using nonconserved amino acids. The structure of the Nrd1 CID–pSer5 CTD complex shows that the specific recognition of $pSer5_a$ is facilitated by the *cis* conformation of the $pSer5_a$–$Pro6_a$ peptidyl-prolyl bond. This *cis* conformation maximizes the intermolecular contacts and prevents the peptide from clashing with the α1–α2 loop of Nrd1. A conserved G26 in Nrd1 (all other CTD-containing proteins have a bulky and charged residues in this position) (Supplemental Fig. S3B) allows for loading of $pSer5_a$ into a highly electropositive pocket (Supplemental Fig. S3A). This is the first CID–CTD structure in which a phosphoserine–proline bond is observed in the *cis* conformation and in which a direct recognition of pSer5 is found. Importantly, both features are interconnected and thus required for the efficient binding. A similar conformation of the pSer5 CTD peptide was found in the crystal structure of the Ssu72–pSer5 CTD complex (Supplemental Fig. S4; Xiang et al. 2010; Werner-Allen et al. 2011).

Interaction between Nrd1 and the CTD is important for cell viability and the processing or degradation of ncRNAs

Specific association of the pSer5 CTD with the Nrd1 CID was further tested in a quantitative in vitro binding assay using FA. We titrated the wild-type and mutant Nrd1 CID against the fluorescently labeled pSer5 CTD. Alanine or aspartate (charge-swapping) substitutions at positions Ser25, Gly26, Arg28, Ile29, and Lys30 significantly decreased the binding affinity with the pSer5 CTD (Fig. 3A; Supplemental Fig. S6, control mutations of nonessential residues). In comparison, the effect of mutants at nonconserved positions Leu20, Lys21, and Ser22 in the region that was previously suggested to bind the phosphoserine (see above) is much smaller (Vasiljeva et al. 2008a). D70R and R74D variants of Nrd1 could not be assayed due to their instability at the high concentrations required for FA measurements. Next, we tested the effect of removal of phosphorylation in the downstream CTD repeat (at $Ser5_b$). We found that the CTD peptide with a single phosphorylation [pSer5(1P)] has the same affinity for Nrd1 as the CTD peptide phosphorylated at both Ser5s (Supplemental Fig. S5). Furthermore, mutations at $Pro6_a$ in the CTD peptide ($P6_aA$ CTD and $P6_aR$ CTD) have a larger negative effect on affinity to Nrd1 than mutations at $Pro6_b$ ($P6_bA$ CTD and $P6_bR$ CTD), confirming the requirement of the *cis* conformation at $pSer5_a$–$Pro6_a$ for the binding to Nrd1 (Supplemental Fig. S5). The effect of the $P6_aR$ mutation is larger than for $P6_aA$, as the bulkier side chain (the side chain at position 6_a is solvent-exposed) creates more unfavorable interactions with the adjacent side chain of pSer5, decreasing the stability of the *cis* conformation. This corroborates previous studies that showed that proline-to-alanine mutations do not necessarily alter the *cis* conformation if it is enforced by the structural context (Mayr et al. 1994; Xiong et al. 2000). Altogether, the FA data strongly support the phosphorylation-specific recognition ob-

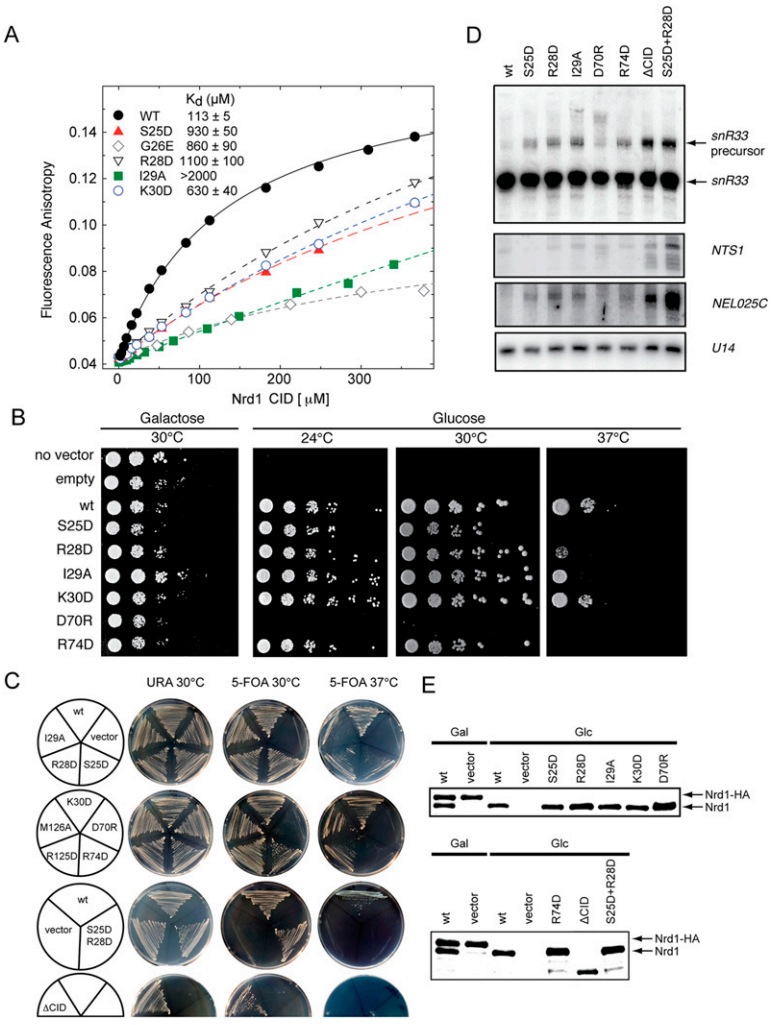


Figure 3. Critical residues of the Nrd1 CID that are required for CTD binding, cell viability, and RNA processing and degradation. (*A*) Equilibrium binding of the Nrd1 CID mutants with the pSer5 CTD peptide monitored by FA. Binding isotherms and dissociation constants (K_d) are shown for individual mutants. (*B*) Phenotypic analysis of the Nrd1 CID mutants. The mutants were expressed from *pRS415* plasmids in the *GAL1::NRD1* strain background. Growth on glucose-containing plates leads to the repression of *GAL1*-driven wild-type Nrd1 and thus shows the functionality of the different Nrd1 CID mutants. (*C*) Phenotypic analysis of the Nrd1 CID mutants. The mutants were expressed episomally from *pRS415* plasmids in the yeast strain where the endogenous *NRD1* was deleted and growth was supplemented with *NRD1* on the *URA3* plasmid (*pRS316*). To test the functionality of the different Nrd1 CID mutants, cells were grown on 5-FOA-containing plates for 3 d at the indicated temperatures. Wild-type *NRD1* was used as positive control, empty pRS415 plasmid was used as negative control, and $Nrd1_{\Delta 1-150}$ was the ΔCID. (*D*) *snR33* snoRNA processing efficiency and stability of *NTS1* and *NEL025C* CUTs analyzed by Northern blot analysis. Mature *U14* snoRNA represents a loading control. Total RNA was purified from cells expressing wild-type *NRD1* or the indicated mutants grown in glucose-containing medium. (*E*) Western blot analysis of expression levels of the wild-type and mutant Nrd1 originating from *pRS415* (faster-migrating band) in the yeast strain where the endogenous *NRD1* is under the galactose promoter. Protein extracts were prepared from the original *GAL1::NRD1* strain transformed with plasmids carrying wild-type and mutant *NRD1* grown in either galactose-containing medium (Gal) or glucose-containing medium (Glc). (*Top* band) The genomic *NRD1* copy contains a fusion HA tag, resulting in slower gel migration. Proteins were detected using specific antibodies against Nrd1p. Vectors with no insertion (vector) or containing the wild-type *NRD1* were used as negative and positive controls, respectively.

served in the structure of the Nrd1–pSer5 CTD complex and that the interaction relies on the presence of the invariant basic residues in the CID domain (Supplemental Fig. S3A).

To determine the importance of these individual residues for Nrd1 function in vivo, we monitored cell viability and ncRNA processing/stability in single-amino-acid mutants (Fig. 3B–D). The Nrd1 strain lacking the CID (Nrd1Δ1–150) was used as a reference for the CID-related function. Deletion of the CID is not lethal (Vasiljeva et al. 2008a); however, we observed that deletion of the CID led to inviability at 37°C (Fig. 3C). Similarly, to a lesser extent, a temperature-sensitive (ts) growth defect was observed in mutants of Ser25, Arg28, Ile29, and Arg74 (Fig. 3B,C). The double mutant in the residues contacting the phosphorylated serine (Ser25+Arg28) exhibited the same growth phenotype as ΔCID Nrd1 (Fig. 3C). Nrd1 CID deletion causes an accumulation of *snR33* precursors in vivo (Vasiljeva et al. 2008a). We observed that point mutants with the *ts* growth phenotype showed snoRNA processing and CUT degradation defects demonstrated by an accumulation of *pre-snR33* snoRNA and *NEL025c* and *NTS1* CUTs, respectively (Fig. 3D,E; Supplemental Fig. S7). None of the Nrd1 CID mutants tested displayed transcription termination defects that would be represented by readthrough product accumulation. Importantly, the Ser25+Arg28 double mutant showed processing and degradation defects comparable with those of the mutant lacking the entire Nrd1 CID. Taken together, these data demonstrate that specific recognition of the pSer5 CTD by the Nrd1 CID is important for the processing and degradation of ncRNAs in vivo.

The Nrd1 complex associates with the exosome and the TRAMP complex (Vasiljeva and Buratowski 2006). Based on our data, it is tempting to speculate that the Nrd1 CID mediates interaction with this processing and degradation apparatus. As the Nrd1 truncation lacking CID (Δ39–169) has no effect on the exosome copurification (demonstrated for Rrp6) (Vasiljeva and Buratowski 2006), it is likely that other RNA processing and degradation auxiliary factors, such as the TRAMP complex (LaCava et al. 2005; Vanacova et al. 2005), are recruited through the Nrd1 CID.

Previous chromatin immunoprecipitation (ChIP) experiments showed that Ess1 promotes the release of Nrd1 from terminator regions (Singh et al. 2009). This observation has been attributed to the indirect effect of Ess1 in which it stimulates dephosphorylation by Ssu72 (Singh et al. 2009). Ssu72 targets specifically the *cis* conformation of the peptidyl-prolyl bond of the pSer-Pro-containing peptides (Xiang et al. 2010; Werner-Allen et al. 2011). Here we suggest that Ess1 can also directly regulate the association of Nrd1 with the RNA Pol II CTD, as it specifically recognizes the pSer5–Pro6 CTD in the *cis* conformation, indicating that Ess1 may play a dual role in regulating the Nrd1 pathway. A detailed understanding of this mechanism will require further studies, but it is an exciting possibility, particularly in light of recent observations that suggested that Ssu72 may be a less "conformation-specific" phosphatase (acting also on the pSer7–Tyr1 CTD peptide that is unlikely to exist in the *cis* conformation) than previously expected (Bataille et al. 2012; Zhang et al. 2012).

Conclusions

The structure of Nrd1 CID–pSer5 CTD presented here reveals that the CTD recognition by Nrd1 requires both phosphorylation and isomerization of the RNAPII CTD. This suggests that the coupling of covalent and noncovalent changes in the CTD structure regulated by kinases/phosphatases and isomerases is crucial for the dynamical process of recruitment and displacement of appropriate processing factors during the transcriptional cycle. In addition, we show that specific recognition of the pSer5 CTD by the Nrd1 CID is important for the processing and degradation of ncRNAs in vivo, suggesting that these events occur cotranscriptionally.

Materials and methods

Protein expression and mutagenesis

The DNA encoding the *Saccharomyces cerevisiae* Nrd1 CID domain (residues 1–153) was amplified and cloned into a pET22b expression vector (Novagen) via NdeI and XhoI restriction sites. Details on cloning, expression, and purification of the Nrd1 CID construct have been described previously (Kubicek et al. 2011). Protein mutants were designed on the basis of the NMR structure of the Nrd1–pSer5 CTD complex and were prepared using the QuikChange site-directed mutagenesis kit (Stratagene).

The CTD of RNAP II

It has been established previously that the CTD mimic consisting of two repeats of the CTD canonical heptad yields the same binding affinity to Nrd1 CID as the CTD mimic of four repeats (Vasiljeva et al. 2008a). Thus, we used a 14-amino-acid peptide, the pSer5 CTD (two repeats of the heptapeptide CTD consensus phosphorylated at two Ser5s; $Tyr1_a$–$Ser2_a$–$Pro3_a$–$Thr4_a$–$pSer5_a$–$Pro6_a$–$Ser7_a$–$Tyr1_b$–$Ser2_b$–$Pro3_b$–$Thr4_b$–$pSer5_b$–$Pro6_b$–$Ser7_b$), in our study to mimic the CTD phosphorylated at Ser5. Similarly, the unphosphorylated CTD, the pSer7 CTD, and the doubly phosphorylated pSer5/7 CTD were used. The peptides were purchased from Clonestar Peptide Services.

NMR

All NMR spectra for the backbone and side chain assignments of 2.0 mM uniformly ^{15}N,^{13}C-labeled Nrd1 CID in 50 mM sodium phosphate buffer (pH 8.0), 100 mM NaCl, and 10 mM β-mercaptoethanol (90% H_2O/10% D_2O) were recorded on Bruker AVANCE 600- and 950-MHz spectrometers equipped with a cryoprobe at a sample temperature of 20°C. The spectra were processed using an NMRPipe package (Delaglio et al. 1995), and the protein resonances were assigned manually using Sparky software (T.G. Goddard and D.G. Kellner, University of California at San Francisco). The ^1H, ^{13}C, and ^{15}N chemical shifts of the bound form of the Nrd1 CID were assigned as described elsewhere (Kubicek et al. 2011). All distance constraints were derived from the three-dimensional (3D) ^{15}N- and ^{13}C-separated NOESYs and two-dimensional (2D) ^1H-^1H NOESY (with a mixing time of 80 msec) collected on a 950-MHz spectrometer. Intermolecular distance constraints were obtained from the 3D F_1-^{13}C/^{15}N-filtered NOESY-[^{13}C,^1H]-HSQC experiment (Zwahlen et al. 1997; Peterson et al. 2004), with a mixing time of 150 msec on a 950-MHz spectrometer. Intramolecular distance constraints of the bound CTD peptide (unlabeled) were derived from a 2D F_1,F_2-^{13}C/^{15}N-filtered [^1H,^1H]-NOESY (τ_m = 150 msec) (Zwahlen et al. 1997; Peterson et al. 2004). The NOEs were semi-quantitatively classified based on their intensities in the 2D and 3D NOESY spectra.

Structure calculations

The preliminary structure determinations of the Nrd1–pSer5 CTD complex were performed with the automated NOE assignment module implemented in the CYANA program (Guntert 2004). In the next step, CYANA-generated restraints along with manually assigned protein–CTD intermolecular restraints were used for further refinement of the preliminary structures with AMBER 10.0 software (Case et al. 2005). These calculations used a modified version (AMBER ff99SB) of the force field described by Cornell et al. (1995) using a protocol described previously (Stefl et al. 2010; Hobor et al. 2011). From 40 refined structures, the 20 conformers with the lowest AMBER energy were selected to form the final ensemble of structures. Molecular graphics were generated using MOLMOL (Koradi et al. 1996) and PyMOL (http://www.pymol.org). The

atomic coordinates and restraints for the Nrd1 CID–pSer5 CTD complex have been deposited in the Protein Data Bank under ID code 2lo6.

FA

The equilibrium binding of the Nrd1 CID to the differently phosphorylated CTD was analyzed by FA. The CTD peptides were N-terminally labeled with the 5,6-carboxyfluorescein (FAM). The measurements were conducted on a FluoroMax-4 spectrofluorometer (Horiba Jobin-Yvon). The instrument was equipped with a thermostatted cell holder with a Neslab RTE7 water bath (Thermo Scientific). Samples were excited with vertically polarized light at 477 nm, and both vertical and horizontal emissions were recorded at 525 nm. All measurements were conducted at 10°C in 50 mM phosphate buffer (pH 8.0) containing 100 mM NaCl and 10 mM β-mercaptoethanol. Each data point is an average of five measurements. The experimental binding isotherms were analyzed by nonlinear least-squares regression in SigmaPlot 11 software (Systat Software) using a single-site binding model according to Heyduk and Lee (1990).

Construction of yeast plasmids

The *pRS415* plasmid (*CEN*, *LEU2*) with insertion of the wild-type *NRD1* gene surrounded by the *NRD1* promoter and terminator (Vasiljeva et al. 2008a) was used as a template for QuikChange site-directed mutagenesis (Stratagene). See Supplemental Table S2 for primer sequences and Supplemental Table S3 for constructs generated in this study. The *NRD1 ΔCID* region was amplified as follows: Fragment 1: −340 nt up to +6 nt of *NRD1* with SVO F71 and SVO F72; Fragment 2: +453 nt up to TAA +300 nt from the 3′ untranslated region (UTR) with SVO F73 and SVO F74 primers. The two resulting PCR products were ligated together and inserted into the above-mentioned *pRS415*-based construct.

Yeast cultures and manipulation

Yeast were cultured under standard conditions in media with selective markers corresponding to particular strains and vectors. Yeast transformations were performed by the lithium acetate method.

Yeast growth test analysis

W303 (*GAL1::NRD1*)-derived strains (Supplemental Table S4) carrying appropriate mutant *NRD1* plasmids were grown in SD-LEU-HIS + 2% galactose at 30°C to an O.D. of 1.0. The cultures were serially diluted by a factor of 10 and spotted onto SD-LEU-HIS medium containing 2% glucose to repress the expression of the endogenous *NRD1* or control medium (SD-HIS + 2% galactose). Plates were incubated at 25°C, 30°C, and 37°C. EJS101-9d-derived strains (Supplemental Table S4) were grown for 3 d on SC-LEU plates, then spread on SC-LEU plates with or without 5-FOA and incubated at 25°C, 30°C, and 37°C.

Whole-cell protein extract preparation for Western blot analysis

Protein extracts were prepared from cultures grown on either galactose- or glucose-containing medium to an O.D. of 1.0. Five milliliters of culture was harvested and lyzed by 1.85 M NaOH for 15 min on ice, and proteins were subsequently precipitated with ice-cold trichloroacetic acid. Pellets were resuspended in 5% SDS and 8 M urea buffer prior to SDS-PAGE analysis. Proteins were resolved on a 12% SDS-PAGE gel, transferred to a nitrocellulose membrane by a semidry electroblotter (Bio-Rad), and probed for the presence of Nrd1p with the anti-Nrd1 sera kindly provided by David Brow (Steinmetz and Brow 1998).

RNA isolation and analysis

For RNA analysis, cells were inoculated to an O.D. of 0.1 in glucose- or galactose-containing SD medium and grown for 16 h at 30°C. RNA was isolated by hot phenol extraction and stored at −80°C. Five micrograms of total RNA was denatured in 25% formamide, separated on an 8% denaturing (8 M urea) polyacrylamide gel, and transferred to nylon membrane using semidry electro-transfer. RNA was cross-linked to the membrane by UV light (120 mJ/cm^2) and hybridized with a probe in Ultra-Hyb buffer (Ambion) according to the manufacturer's instructions.

Preparation of DNA probes for Northern blot analysis

The DNA probe for *snR33* was amplified from *S. cerevisiae* S288C genomic DNA with primers Forward, 5′-CGGAACGGTACATAAGAA TAGAAGAG-3′, and Reverse, 5′-TAAAGAAAACGATAAGAACTAA CCTC-3′. The *NTS1* probe 1 was prepared according to Vasiljeva et al. (2008b), by using primers Forward, 5′-TGAGTGCTTGTATAAGTTTA GAGAATTGA-3′, and Reverse, 5′-TTAATACTTTCCTCTTCGTCTTTT TCTAC-3′. The *NEL025c* probe was amplified with primers Forward, 5′-CCTGTTGACATTGCAGACAA-3′, and Reverse, 5′-GCAAAGATCTG TATGAAAGG-3′. The resulting PCR products were used as templates for random primed labeling using [α-^{32}P]dATP and the commercial kit (Roche). To detect *U14* snoRNA, the oligonucleotide 5′-TCACTCAGACATCC TAGG-3′ was 5′-phosphate-labeled by T4 polynucleotide kinase (New England Biolabs) and [γ-^{32}P]ATP.

Acknowledgments

We thank Anton Meinhart for helpful advice. We also thank Domenico Libri and David Brow for yeast strains, constructs, and antibodies. This work was supported by the project "CEITEC-Central European Institute of Technology" (CZ.1.05/1.1.00/02.0068) from the European Regional Development Fund, Czech Science Foundation (P305/12/G034; K.K. was supported by P305/10/1490, D.H. was supported by P305/11/1095, and C.H. was supported by P205/12/0550), Wellcome Trust 084316/Z/07/Z, and EMBO Installation Grant 1642. The NOESY spectra were obtained at the BMRZ NMR facility supported by the EU-NMR program (RII3-026145). P.H. is in receipt of the Brno City Municipality Scholarship for Talented PhD Students.

References

Arigo JT, Eyler DE, Carroll KL, Corden JL. 2006. Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. *Mol Cell* **23:** 841–851.

Bataille AR, Jeronimo C, Jacques PE, Laramee L, Fortin ME, Forest A, Bergeron M, Hanes SD, Robert F. 2012. A universal RNA polymerase II CTD cycle is orchestrated by complex interplays between kinase, phosphatase, and isomerase enzymes along genes. *Mol Cell* **45:** 158–170.

Becker R, Loll B, Meinhart A. 2008. Snapshots of the RNA processing factor SCAF8 bound to different phosphorylated forms of the carboxyl-terminal domain of RNA polymerase II. *J Biol Chem* **283:** 22659–22669.

Buratowski S. 2003. The CTD code. *Nat Struct Biol* **10:** 679–680.

Buratowski S. 2009. Progression through the RNA polymerase II CTD cycle. *Mol Cell* **36:** 541–546.

Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ. 2005. The Amber biomolecular simulation programs. *J Comput Chem* **26:** 1668–1688.

Chapman RD, Heidemann M, Hintermair C, Eick D. 2008. Molecular evolution of the RNA polymerase II CTD. *Trends Genet* **24:** 289–296.

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 Biol NMR* **6:** 277–293.

Egloff S, Murphy S. 2008. Cracking the RNA polymerase II CTD code. *Trends Genet* **24:** 280–288.

Gemmill TR, Wu X, Hanes SD. 2005. Vanishingly low levels of Ess1 prolyl-isomerase activity are sufficient for growth in *Saccharomyces cerevisiae*. *J Biol Chem* **280:** 15510–15517.

Gudipati RK, Villa T, Boulay J, Libri D. 2008. Phosphorylation of the RNA polymerase II C-terminal domain dictates transcription termination choice. *Nat Struct Mol Biol* **15:** 786–794.

Guntert P. 2004. Automated NMR structure calculation with CYANA. *Methods Mol Biol* **278:** 353–378.
Heyduk T, Lee JC. 1990. Application of fluorescence energy transfer and polarization to monitor *Escherichia coli* cAMP receptor protein and *lac* promoter interaction. *Proc Natl Acad Sci* **87:** 1744–1748.
Hirose Y, Manley JL. 2000. RNA polymerase II and the integration of nuclear events. *Genes Dev* **14:** 1415–1429.
Hobor F, Pergoli R, Kubicek K, Hrossova D, Bacikova V, Zimmermann M, Pasulka J, Hofr C, Vanacova S, Stefl R. 2011. Recognition of transcription termination signal by the nuclear polyadenylated RNA-binding (NAB) 3 protein. *J Biol Chem* **286:** 3645–3657.
Kim H, Erickson B, Luo W, Seward D, Graber JH, Pollock DD, Megee PC, Bentley DL. 2010. Gene-specific RNA polymerase II phosphorylation and the CTD code. *Nat Struct Mol Biol* **17:** 1279–1286.
Komarnitsky P, Cho EJ, Buratowski S. 2000. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes Dev* **14:** 2452–2460.
Koradi R, Billeter M, Wuthrich K. 1996. MOLMOL: A program for display and analysis of macromolecular structures. *J Mol Graph* **14:** 51–55.
Kubicek K, Pasulka J, Cerna H, Lohr F, Stefl R. 2011. ^{1}H, ^{13}C, and ^{15}N resonance assignments for the CTD-interacting domain of Nrd1 bound to Ser5-phosphorylated CTD of RNA polymerase II. *Biomol NMR Assign* **5:** 203–205.
LaCava J, Houseley J, Saveanu C, Petfalski E, Thompson E, Jacquier A, Tollervey D. 2005. RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell* **121:** 713–724.
Lunde BM, Reichow SL, Kim M, Suh H, Leeper TC, Yang F, Mutschler H, Buratowski S, Meinhart A, Varani G. 2010. Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain. *Nat Struct Mol Biol* **17:** 1195–1201.
Maniatis T, Reed R. 2002. An extensive network of coupling among gene expression machines. *Nature* **416:** 499–506.
Mayer A, Lidschreiber M, Siebert M, Leike K, Soding J, Cramer P. 2010. Uniform transitions of the general RNA polymerase II transcription complex. *Nat Struct Mol Biol* **17:** 1272–1278.
Mayr LM, Willbold D, Rösch P, Schmid FX. 1994. Generation of a non-prolyl *cis* peptide bond in ribonuclease T1. *J Mol Biol* **240:** 288–293.
Meinhart A, Cramer P. 2004. Recognition of RNA polymerase II carboxy-terminal domain by 3′-RNA-processing factors. *Nature* **430:** 223–226.
Meinhart A, Kamenski T, Hoeppner S, Baumli S, Cramer P. 2005. A structural perspective of CTD function. *Genes Dev* **19:** 1401–1415.
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 Biol NMR* **28:** 59–67.
Phatnani HP, Greenleaf AL. 2006. Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev* **20:** 2922–2936.
Singh N, Ma Z, Gemmill T, Wu X, Defiglio H, Rossettini A, Rabeler C, Beane O, Morse RH, Palumbo MJ, et al. 2009. The Ess1 prolyl isomerase is required for transcription termination of small non-coding RNAs via the Nrd1 pathway. *Mol Cell* **36:** 255–266.
Stefl R, Oberstrass FC, Hood JL, Jourdan M, Zimmermann M, Skrisovska L, Maris C, Peng L, Hofr C, Emeson RB, et al. 2010. The solution structure of the ADAR2 dsRBM–RNA complex reveals a sequence-specific readout of the minor groove. *Cell* **143:** 225–237.
Steinmetz EJ, Brow DA. 1998. Control of pre-mRNA accumulation by the essential yeast protein Nrd1 requires high-affinity transcript binding and a domain implicated in RNA polymerase II association. *Proc Natl Acad Sci* **95:** 6699–6704.
Steinmetz EJ, Conrad NK, Brow DA, Corden JL. 2001. RNA-binding protein Nrd1 directs poly(A)-independent 3′-end formation of RNA polymerase II transcripts. *Nature* **413:** 327–331.
Thiebaut M, Kisseleva-Romanova E, Rougemaille M, Boulay J, Libri D. 2006. Transcription termination and nuclear degradation of cryptic unstable transcripts: A role for the nrd1–nab3 pathway in genome surveillance. *Mol Cell* **23:** 853–864.
Tietjen JR, Zhang DW, Rodriguez-Molina JB, White BE, Akhtar MS, Heidemann M, Li X, Chapman RD, Shokat K, Keles S, et al. 2010. Chemical-genomic dissection of the CTD code. *Nat Struct Mol Biol* **17:** 1154–1161.
Vanacova S, Wolf J, Martin G, Blank D, Dettwiler S, Friedlein A, Langen H, Keith G, Keller W. 2005. A new yeast poly(A) polymerase complex involved in RNA quality control. *PLoS Biol* **3:** e189. doi: 10.1371/journal.pbio.0030189.
Vasiljeva L, Buratowski S. 2006. Nrd1 interacts with the nuclear exosome for 3′ processing of RNA polymerase II transcripts. *Mol Cell* **21:** 239–248.
Vasiljeva L, Kim M, Mutschler H, Buratowski S, Meinhart A. 2008a. The Nrd1–Nab3–Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. *Nat Struct Mol Biol* **15:** 795–804.
Vasiljeva L, Kim M, Terzi N, Soares LM, Buratowski S. 2008b. Transcription termination and RNA degradation contribute to silencing of RNA polymerase II transcription within heterochromatin. *Mol Cell* **29:** 313–323.
Werner-Allen JW, Lee CJ, Liu P, Nicely NI, Wang S, Greenleaf AL, Zhou P. 2011. *Cis*-proline-mediated Ser(P)5 dephosphorylation by the RNA polymerase II C-terminal domain phosphatase Ssu72. *J Biol Chem* **286:** 5717–5726.
Xiang K, Nagaike T, Xiang S, Kilic T, Behav MM, Manley JL, Tong L. 2010. Crystal structure of the human symplekin–Ssu72–CTD phosphopeptide complex. *Nature* **467:** 729–733.
Xiong Y, Juminaga D, Swapna GV, Wedemeyer WJ, Scheraga HA, Montelione GT. 2000. Solution NMR evidence for a *cis* Tyr–Ala peptide group in the structure of [Pro93Ala] bovine pancreatic ribonuclease A. *Protein Sci* **9:** 421–426.
Zhang DW, Mosley AL, Ramisetty SR, Rodríguez-Molina JB, Washburn MP, Ansari AZ. 2012. Ssu72 phosphatase-dependent erasure of phospho-Ser7 marks on the RNA polymerase II C-terminal domain is essential for viability and transcription termination. *J Biol Chem* **287:** 8541–8551.
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 λ N-peptide/boxB RNA complex. *J Am Chem Soc* **119:** 6711–6721.

The EMBO Journal (2012) 31, 3935–3948 | © 2012 European Molecular Biology Organization | All Rights Reserved 0261-4189/12
www.embojournal.org

THE EMBO JOURNAL

In vivo SELEX reveals novel sequence and structural determinants of Nrd1-Nab3-Sen1-dependent transcription termination

Odil Porrua[1], Fruzsina Hobor[2], Jocelyne Boulay[1], Karel Kubicek[2], Yves D'Aubenton-Carafa[1], Rajani Kanth Gudipati[1,3], Richard Stefl[2] and Domenico Libri[1,*]

[1]Centre de Génétique Moléculaire, Gif sur Yvette, Paris, France and [2]CEITEC—Central European Institute of Technology, Masaryk University, Brno, Czech Republic

The Nrd1-Nab3-Sen1 (NNS) complex pathway is responsible for transcription termination of cryptic unstable transcripts and sn/snoRNAs. The NNS complex recognizes short motifs on the nascent RNA, but the presence of these sequences alone is not sufficient to define a functional terminator. We generated a homogeneous set of several hundreds of artificial, NNS-dependent terminators with an *in vivo* selection approach. Analysis of these terminators revealed novel and extended sequence determinants for transcription termination and NNS complex binding as well as supermotifs that are critical for termination. Biochemical and structural data revealed that affinity and specificity of RNA recognition by Nab3p relies on induced fit recognition implicating an α-helical extension of the RNA recognition motif. Interestingly, the same motifs can be recognized by the NNS or the mRNA termination complex depending on their position relative to the start of transcription, suggesting that they function as general transcriptional insulators to prevent interference between the non-coding and the coding yeast transcriptomes.

The EMBO Journal (2012) **31,** 3935–3948. doi:10.1038/emboj.2012.237; Published online 28 August 2012
Subject Categories: RNA
Keywords: cryptic unstable transcripts; hidden transcription; Nrd1p-Nab3p-Sen1p complex; transcriptional insulators; transcription termination

Introduction

The concept of pervasive and/or hidden transcription has emerged in the last few years from studies revealing that the transcribed fraction of eukaryotic genomes is considerably higher than expected from early annotations based on RNA steady-state abundance and phylogenetic conservation (Johnson *et al*, 2005; Jacquier, 2009). Whether pervasive transcription predominantly pollutes the expression of meaningful genetic information or increases the regulatory potential of the cell remains matter of debate. What is clear, however, is that it has to be controlled to maintain the stability and intelligibility of the coding transcriptome. RNA degradation and termination of transcription are crucial in this respect.

In *Saccharomyces cerevisiae*, termination of RNA polymerase II (RNAPII) transcription occurs *via* two major pathways (Kuehner *et al*, 2011). Transcription termination of mRNA coding genes depends on a multi-subunit complex, composed by the Cleavage and Polyadenylation Factor and the Cleavage Factors IA and IB (hereafter referred to as the CPF complex). The CPF complex is recruited to the nascent RNA when the latter contains signals that are recognized by RNA binding subunits, among which, Hrp1p and Rna15p. Termination occurs concomitantly or shortly after cleavage of the nascent transcript and polyadenylation by the poly(A) polymerase Pap1p, which is required for subsequent export to the cytoplasm and translation (Mandel *et al*, 2008; Kuehner *et al*, 2011; Millevoi and Vagner, 2011).

The second pathway plays a central role in the control of pervasive transcription as well as in the biogenesis of sn- and snoRNAs (Steinmetz *et al*, 2001; Thiebaut *et al*, 2006; Arigo *et al*, 2006b; Gudipati *et al*, 2008). It is dependent on an essential protein complex constituted by the RNA-binding proteins Nrd1p and Nab3p and the putative helicase Sen1p (hereafter referred to as the NNS complex). The targets of the NNS complex include transcription units producing short 200–600 nt unstable RNAs dubbed CUTs (Cryptic Unstable Transcripts) (Wyers *et al*, 2005; Thiebaut *et al*, 2006; Arigo *et al*, 2006b). Contrary to the CPF pathway, termination by the NNS pathway is coupled to degradation of the transcript produced or trimming of the precursor in case of sn- and snoRNAs. The RNAs are polyadenylated by the TRAMP complex, containing a different poly(A) polymerase encoded by the *TRF4* gene (LaCava *et al*, 2005; Vanacova *et al*, 2005; Wyers *et al*, 2005; Egecioglu *et al*, 2006), which stimulates degradation by the nuclear exosome, a complex with exo- and endonuclease activities borne by its catalytic subunits, Rrp6p and Dis3p (Lebreton and Seraphin, 2008; Schmid and Jensen, 2008; Chlebowski *et al*, 2011). One important specificity of the NNS pathway is that it functions almost exclusively within a window of <1000 bp after transcription initiation (Jenks and Reines, 2005; Steinmetz *et al*, 2006; Kopcewicz *et al*, 2007; Gudipati *et al*, 2008). This is thought to relate to the preferential interaction of Nrd1p with the RNAPII carboxy terminal domain (CTD) phosphorylated on serine 5 of its heptapeptide repeats (Ser5-P), which predominates early in transcription (Vasiljeva *et al*, 2008; Mayer *et al*, 2011; Tietjen *et al*, 2011).

CUTs constitute the largest share of hidden transcription in yeast, and are produced by at least as many transcription

*Corresponding author. Centre de Génétique Moléculaire, Centre National de la Recherche Scientifique, avenue de la Terrasse, Gif sur Yvette, Paris 91190, France. Tel.: +33 1 69823663; Fax: +33 1 69823877; E-mail: libri@cgm.cnrs-gif.fr
[3]Present address: Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, MA 02115, USA

Received: 30 April 2012; accepted: 26 July 2012; published online: 28 August 2012

units as mRNA coding genes (Wyers *et al*, 2005; Davis and Ares, 2006; Houalla *et al*, 2006; Neil *et al*, 2009; Xu *et al*, 2009). These RNAs are widespread, generally originating from bidirectional promoters associated with mRNA coding genes. They are often found in intergenic regions and in several cases they overlap mRNA coding genes, either in the sense or in antisense orientation. Termination by the NNS complex most often prevents full transcriptional overlap, which would be disruptive. At the same time, overlapping non-coding transcription has been clearly involved in the regulation of gene expression, exemplified by the nucleotide biogenesis and glycolysis pathways (Kuehner and Brow, 2008; Thiebaut *et al*, 2008; Neil *et al*, 2009). Thus, the NNS pathway plays a pivotal role in shaping the balance between regulation and protection of the coding transcriptome.

Understanding the sequence motifs that encode termination signals is a prerequisite to decrypt the mechanism of NNS-dependent termination pathway and its impact in controlling pervasive transcription. Transcription termination by the NNS pathway critically requires the interaction of Nrd1p and Nab3p with the nascent transcript containing GUAA/G and UCUU tetranucleotides, respectively (Carroll *et al*, 2004, 2007; Steinmetz *et al*, 2006; Hobor *et al*, 2011; Lunde *et al*, 2011). Although the importance of these motifs has been clearly established in several studies with model termination substrates, their presence is not sufficient to univocally define terminators. The abundance of these motifs is highly variable among the characterized terminators, ranging from one to more than ten (Thiebaut *et al*, 2006; Arigo *et al*, 2006a; Kuehner and Brow, 2008), strongly suggesting that additional sequences and/or a particular arrangement of motifs are required for defining *bona fide* NNS-dependent terminators.

Because a large number of natural cryptic transcripts are known, most of which are NNS-dependent, termination signals could theoretically be extracted from these sequences. However, the existence of strong sequence biases complicates the statistical analysis. For instance, it is not trivial to define a robust background model, that is, a set of neutral elements relative to which the test set can be judged to contain over- or under-represented words. Also, the co-existence of multiple, positive or negative selective pressures complicate the recognition of specific signals in the natural genomic environment. For instance, GUAA and GUAG are indeed underrepresented in coding regions, but whether this is due to the presence of stop codons (underlined) or to their role in NNS-dependent termination remains undetermined. Similarly, intergenic regions are enriched in termination motifs, but also in regulatory signals for the initiation of transcription.

The strategy we undertook in the present study allows circumventing both limitations. We adopted an *in vivo* SELEX strategy using an original genetic system and selected short terminators of uniform length among a pool of random sequences. This has provided a large winning set of sequences selected exclusively on their ability to induce termination, mostly by the NNS pathway. Importantly, this strategy also provided a very robust background model in the set of non-selected sequences, allowing a very reliable statistical evaluation of the overrepresentation of sequence motifs in terminators. We identified and validated extended binding sites for Nrd1p and Nab3p as well as novel, AU-rich motifs that are also bound by the complex. NMR titration experiments revealed that Nab3p recognizes its extended site via an induced fit mechanism and allowed identifying a novel region of the protein that critically contributes to the specificity of the interaction. Importantly, we show that the overall affinity of the NNS complex for the nascent RNA is not always limiting for termination. Rather, the arrangement of sites and their association in supermotifs is critical for function. Finally, we demonstrate that the same sequence motifs can be recognized by either the NNS or the CPF complex, depending merely on the distance from the transcriptional start site. Thus, both pathways have adapted to recognize largely overlapping signals, in spite of the different protein composition and fate of the transcripts produced. These results have important implications for the mechanism of NNS complex termination and its function in the control of pervasive transcription.

Results

In vivo selection of artificial CUTs from a naive pool of sequences

To isolate the elements dictating NNS-dependent termination from overlapping signals that might be present in natural CUTs, we devised a strategy to select for terminators from a naïve pool of sequences. We constructed a reporter system containing two strong promoters in tandem, P_{Tet} and P_{Gal} (Supplementary Figure 1A). The two promoters are separated by a test sequence and P_{Gal} drives expression of *CUP1*, which confers copper-resistant growth to yeast. Transcription from the upstream P_{Tet} promoter interferes with transcription from P_{Gal}, generating copper-sensitive yeasts unless the test sequence contains a terminator, which allows expression of *CUP1* and copper-resistant growth (Supplementary Figure 1B–D). A pool of naïve sequences containing a random region of 120 nt obtained by chemical synthesis was introduced in the reporter system by recombination and subjected to two rounds of selection on copper-containing medium (see Materials and methods; Figure 1A and Supplementary Figure 2A).

Sequencing of roughly 130 inserts from the selected pool allowed defining two classes of potential terminators. Approximately 70% of the selected sequences were enriched in the previously identified Nrd1p- and Nab3p-binding sites, suggesting that these are NNS-dependent terminators (Supplementary Figure 2B). The other class of sequences contained a different set of motifs that will be described elsewhere (Colin *et al*, in preparation). We will hereafter only refer to the sequences belonging to the first class.

Northern blot analyses largely validated the predominant occurrence of NNS complex-dependent termination in the selected clones. We observed expression of the short and functional *CUP1* transcript in all the constructs analysed but not in a negative control containing a copper-sensitive clone (data not shown). Short transcripts driven by P_{Tet} and terminating upstream of P_{Gal} were specifically observed in the selected clones. These RNAs were generally strongly stabilized in the absence of the nuclear exosome subunit Rrp6p (Figure 1B and data not shown), confirming the production of unstable transcripts, which is characteristic of the NNS pathway. Importantly, termination of these clones was affected by metabolic depletion of Nrd1p (Figure 1C) leading to the appearance of long read-through transcripts.

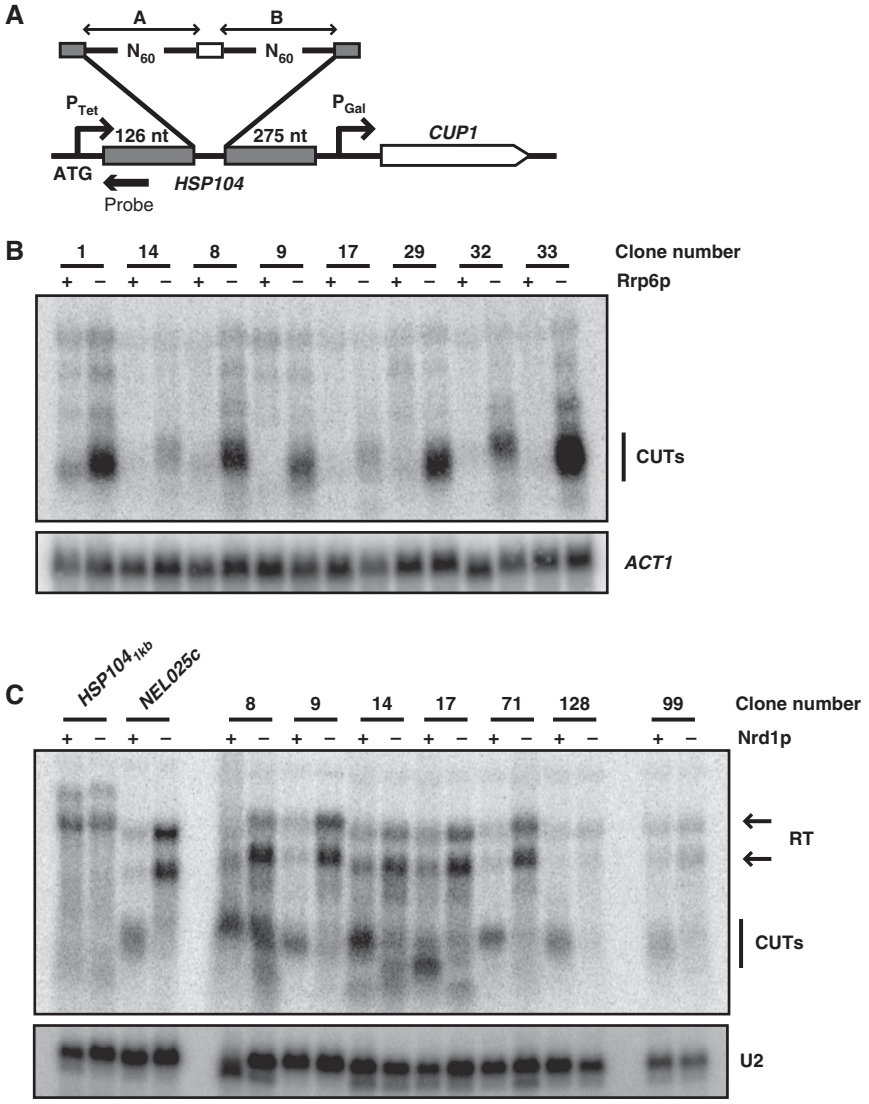


Figure 1 *In vivo* selection of transcriptional terminators. (**A**) Scheme of the reporter construct used for the selection. The structure of the random sequences is shown on the top. A white box indicates a short central constant region, introduced for cloning purposes. The position of the probe used for the northern analyses in (**B**) and (**C**) is marked by an arrowhead. (**B**) Northern blot analysis of selected clones in a wild-type or Δ*rrp6* strain as indicated. Short unstable transcripts resulting from transcription termination at the inserted sequences are labelled CUTs. *ACT1* mRNA is used as a loading control. (**C**) Northern blot analysis of selected clones to assess NNS complex dependency. Analysis was performed in a P_{Gal}-*NRD1*, Δ*rrp6* strain, grown either on galactose ('+' lanes) or on glucose for 6 h ('−' lanes) to deplete Nrd1p. Small unstable transcripts derived from termination at the selected region (CUTs) are observed in the presence of Nrd1p. $HSP104_{1\,kb}$ contains the first 1 kb of *HSP104* coding sequences as a negative control for termination. As a positive control, we used sequences from the *NEL025c* CUT that induces NNS complex-dependent termination. Transcriptional read-through (RT) at the selected terminators (or the control) produces transcripts that terminate at the downstream *CUP1* terminator or a cryptic terminator in the *GAL1* promoter (marked by two arrowheads). The U2 RNA is used as a loading control. Figure source data can be found with the Supplementary data.

Taken together, these results indicate that the pool of selected sequences is strongly enriched for artificial, fully functional CUT-like NNS-dependent terminators.

Statistical analyses of artificial CUTs lead to the identification of putative new motifs involved in NNS-dependent termination

We bulk sequenced the inserts with a paired ends protocol as described in Materials and methods. The presence of over-represented motifs was statistically evaluated relative to the robust background model provided by the pool of non-selected sequences using RSAT (van Helden, 2003). After filtering, ~700 selected clones were analysed against a neutral set of roughly 8000 non-selected sequences. Since the starting pool of random sequences contains two regions with a different sequence bias (Figure 1A, regions A and B; see Materials and methods) we evaluated the statistical significance of overabundant motifs of four and five nucleotides separately in the two regions. The selection nicely converged towards a common set of motifs found in the two regions. As expected, the known sites (UCUU, GUAA and GUAG) recognized by Nrd1p and Nab3p (Carroll *et al*, 2004) were all significantly overrepresented (Figure 2A; Supplementary Figure 3), although GUAG was less prominent, possibly suggesting a minor role for this motif. The highly significant abundance of the UCUU-overlapping tetranucleotide CUUG in both sets suggested that the two motifs are part of an extended motif (UCUUG), which was confirmed

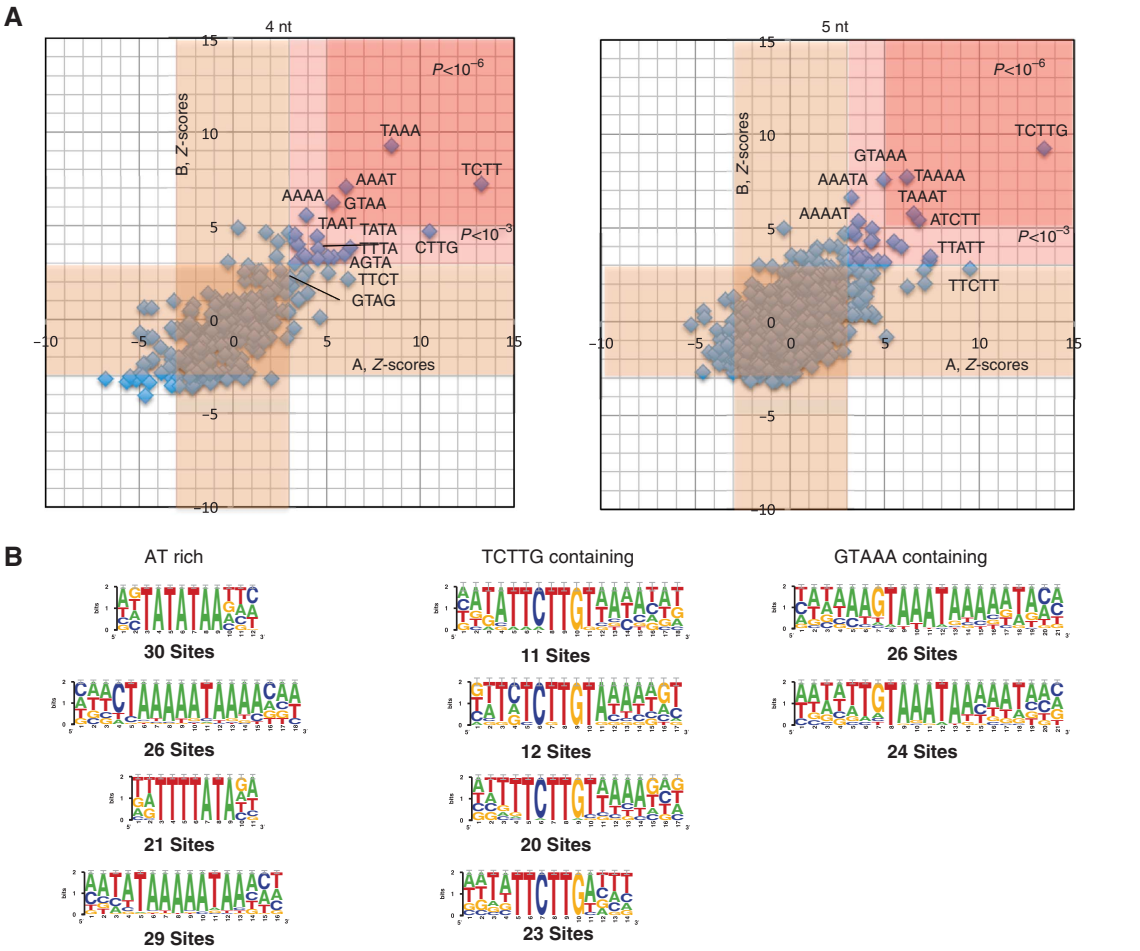


Figure 2 Statistical analysis of the motifs overrepresented in the selected terminators. (**A**) Dispersion plot of the *Z*-values for all possible tetra- and pentanucleotides in the two regions (A, x axis and B, y axis) of the selected sequences. *Z*-values are calculated by RSAT relative to the observed frequency distribution of tetra- and pentanucleotides in the non-selected sequences, determined separately for the A and B regions due to the different sequence bias in the starting pool. Two red zones indicate *P*-values associated to the Z-scores that are respectively lower than 10^{-3} (light red) or 10^{-6} (darker red). Exact values for the most represented oligonucleotides are indicated in Supplementary Figure 3. A light brown region indicates a Z-score range of 3 around the centre of the distribution. The sequence of the oligonucleotides with the highest Z-scores is indicated on the graphic. (**B**) Sequence logos of extended motifs determined by pattern assembly. RSAT uses the most frequent overlapping oligonucleotides to generate larger motifs. Logos are determined based on the occurrence of these supermotifs in the sequences analysed. The number of sequences used to generate the logos is indicated below each logo. Supermotifs were classed according to the presence or absence of Nrd1p- or Nab3p-binding sites or AU-rich motifs. (**C**) Examples of sequences containing supermotifs. Previously identified Nrd1p- and Nab3p-binding sites as well as their new extended versions are indicated in orange and green, respectively, while the AU-rich motifs are shown in grey. The first set of sequences (numbered) represents clones that have been validated by northern analyses. The second set of sequences includes natural known NNS-dependent terminators, including well-characterized snoRNAs (*SNR13*, *SNR33* and *SNR71*) and CUTs (*IMD2* and *URA2* CUTs) and other CUTs identified in a previous genome-wide analysis (Xu *et al*, 2009).

by pentanucleotide analysis ($P<3\times E^{-14}$). This extended motif is by far the most prominent feature detected in the set of artificial terminators but also in natural CUTs ($P=7.5\times E^{-44}$) and is markedly excluded in protein coding regions ($P=1\times E^{-46}$, see Materials and methods). No strong preference was observed for the nucleotide preceding UCUU, while GUAA sites followed by an A and preceded by an A or a U were significantly enriched over other GUAA-containing pentanucleotides (Supplementary Figure 3), suggesting that (A/U)GUAAA represents an extended binding site for Nrd1p.

Surprisingly, AU-rich motifs were strongly overrepresented in parts A and B of artificial CUTs to the same extent as Nrd1p and Nab3p canonical or extended sites (Figure 2A; Supplementary Figure 3). For instance, the sequences UAAA and AAAU were generally enriched to a higher extent ($P<E^{-8}$) than the *bona fide* Nrd1p-binding sites GUAA and GUAG. This finding strongly suggests that AU-rich motifs are major determinants for termination by the NNS pathway. To assess the possible functional association of these motifs, we performed a pattern assembly analysis using RSAT (Figure 2B and C). The program assembles overlapping motifs that are statistically overrepresented to obtain larger consensus regions and generates logos based on occurrences in the set of sequences analysed (van Helden, 2003). Besides extended AU-rich regions, we observed a significant association between the UCUUG or the GUAA sequences and AU motifs, suggesting that proximity between these sites is functionally relevant within larger termination supermotifs. Importantly, such motifs are frequently observed in natural sequences including the well-characterized *SNR13* and *SNR33* terminators or the *IMD2* and *URA2* CUTs (Figure 2C).

Overall, these analyses point to the sequences UCUUG and A/UGUAAA as the main Nrd1p–Nab3p binding sites. They also strongly point to the existence of supermotifs as major determinants for NNS-dependent termination.

Mutational analysis of an artificial CUT confirms the role of the new motifs in termination by the NNS pathway

The analyses presented above suggested that specificity for transcription termination by the NNS complex might rely on longer and more complex arrangements of sites. We therefore undertook a mutational analysis of one of our artificial CUTs (clone #78) containing some of the selected motifs clustered in a short region that is necessary and sufficient for NNS-dependent termination (Supplementary Figure 4). This region contains a Nrd1p-binding site (GUAA), two variants of the Nab3p-binding site (UCUUG, $Nab3_1$ and CUUG, $Nab3_2$) and an AU-rich motif (UAUAUAA) (Figure 3A). Importantly, the AU-rich motif is located immediately downstream of the $Nab3_1$-binding site, defining a putative supermotif.

Interestingly, among the three binding sites for the NNS complex, only the extended $Nab3_1$ site, UCUUG, was required for termination as assessed by copper resistance assays and northern blot analysis of mutated constructs (Figure 3B; Supplementary Figure 5, compare constructs μNrd1, $\mu Nab3_1$ and $\mu Nab3_2$). Importantly, the G nucleotide extending the $Nab3_1$ site was essential in this context as its mutation alone (μG) led to a termination defect similar to that observed for $\mu Nab3_1$ (Figure 3B, compare lanes 1, 4 and 7). Finally, mutation of the AU-rich motif alone (μAU) dramatically impaired termination (Figure 3B, compare lanes 1 and 2), indicating that this sequence plays an essential role in NNS-dependent termination. The double mutants tested did not worsen the termination defects observed in single mutants (Figure 3B, lanes 3 and 6), even when the latter were both defective (i.e., μGμAU), likely indicating a threshold effect. Together, these results indicate that the G extending the $Nab3_1$ site and the AU-rich motif are important elements for NNS-dependent termination. Also, they indicate that the binding sites for the NNS complex are not all functionally equivalent.

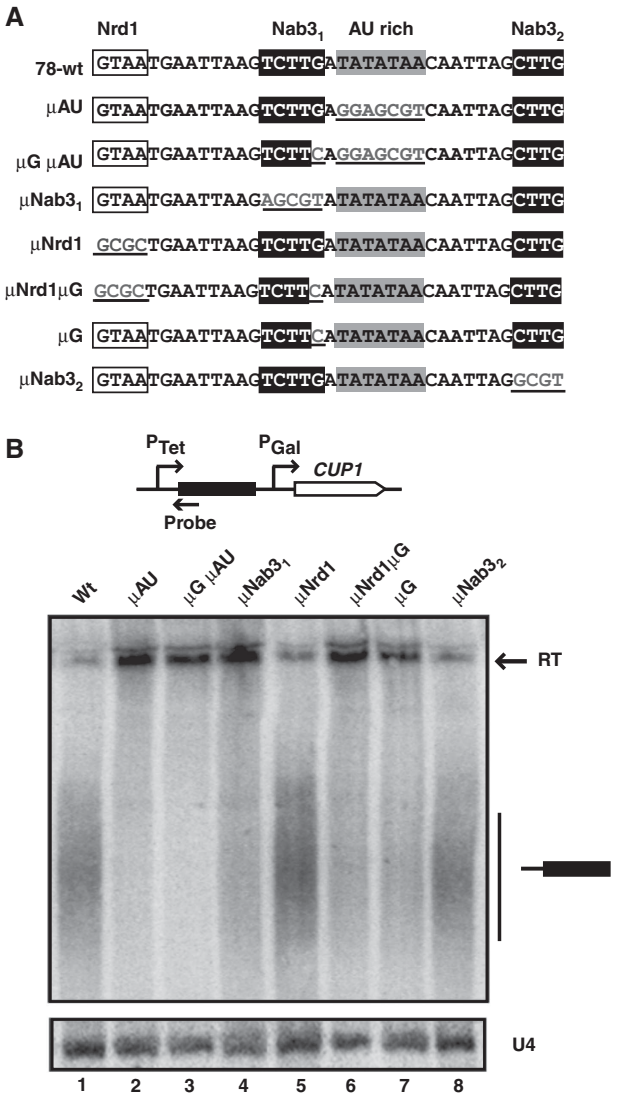


Figure 3 Mutational analysis of the sequence motifs in clone 78. (**A**) Position of the motifs and sequence of the mutations introduced. (**B**) Northern blot analysis to detect termination induced by the different constructs. Transcripts were separated by 6% PAGE for better resolution of the short species. Read-through transcripts (RT) are indicated by a black arrow. Note that the two read-through species shown in Figure 1 are not resolved in this experimental set-up. The small RNA U4 was used as a loading control. Figure source data can be found with the Supplementary data.

The arrangement of sequence motifs dictates the efficiency and specificity of a terminator

The latter results might suggest that the impact of a given motif on termination is context dependent and that, in the appropriate configuration, robust termination can be induced

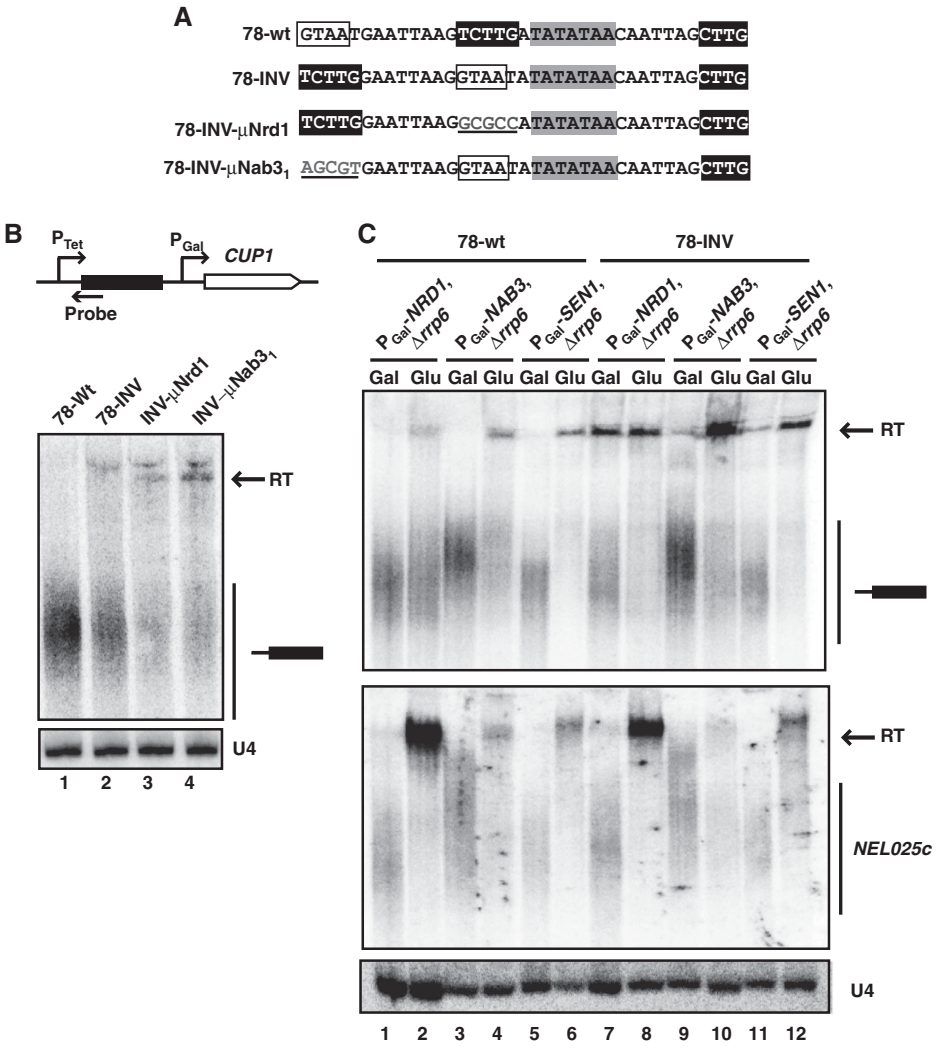


Figure 4 Impact of the arrangement of termination motifs on termination. (**A**) Sequence of the terminator variants analysed. (**B**) Northern blot analysis of RNAs derived from the constructs indicated. Read-through transcripts (RT) are indicated by a black arrow. (**C**) Northern blot analysis of constructs 78-wt and 78-INV upon metabolic depletion of the different components of the NNS complex. Cells were grown on glucose for 6 h (to deplete Nrd1p) or 14 h (to deplete Nab3p and Sen1p). Detection of the transcripts expressed from P_{Tet} is shown on the top. Expression of the *NEL025c* CUT (bottom) was monitored to verify the correct depletion of the proteins. The small RNA U4 was used as a loading control.

by a few 'strong' elements. Because only the Nab3$_1$ site is indispensable for termination, and this site is associated to the AU-rich motif in one supermotif (Figure 2), we set out to investigate whether this proximity defines a termination-proficient context. To this end, we first inverted the positions of the Nrd1 and the Nab3$_1$ sites in construct 78-INV (Figure 4A) and asked whether the acquired proximity with the AU-rich sequence makes the Nrd1 site functionally important for termination. Note that in this novel configuration the two sites constitute a GUAA-AU, supermotif (Figure 2). The efficiency of termination in construct 78-INV was only slightly diminished relative to construct 78 (Figure 4B, lanes 1 and 2). Strikingly, however, when the Nrd1-binding site was mutated in this position (78-INV-μNrd1), termination was impaired (Figure 4B, lanes 2 and 3, Supplementary Figure 5), indicating that this site becomes significant when associated to the AU-rich motif. Consistent with this notion, mutation of the AU-rich motif in this context completely abolishes termination (Supplementary Figure 5). Interestingly, the Nab3$_1$ site set apart from the AU-rich region retained some functionality, as it was still necessary for efficient termination (Figure 4B, lanes 2 and 4).

The above results predict that substrates with different functional sites should be differentially sensitive to impairment of either Nrd1p or Nab3p function. For instance, clone 78 (that does not contain functional Nrd1 sites) is expected to be less dependent on Nrd1p while 78-INV (containing Nrd1 and Nab3 functional sites) should be dependent on both proteins. Indeed, metabolic depletion of Nrd1p impaired recognition of the 78-INV terminator or the endogenous *NEL025c* as a control but only affected to a limited extent termination of clone 78 (Figure 4C, lanes 1–2 and 7–8). Nrd1p depletion was effective as very low to undetectable levels of protein were observed after 2 h of metabolic depletion and longer depletion times did not further affect termination (Supplementary Figure 6 and data not shown). Conversely, termination induced by both 78 and 78-INV sequences was markedly dependent on Nab3p (Figure 4C,

lanes 3, 4, 9 and 10). Finally, metabolic depletion of Sen1p impinged similarly on termination of both constructs, consistent with the notion that this factor is required for termination but not for the recognition of termination signals (Figure 4C, lanes 5, 6, 11 and 12).

These findings confirm the importance of the supermotifs containing canonical NNS complex binding sites associated with AU-rich regions. This underscores the notion that the arrangement of termination motifs, and not their mere presence on the nascent transcript, determines the 'strength' and specificity of NNS-dependent terminators.

The AU-rich motif is recognized by the NNS complex in vitro

The prominent enrichment of AU-rich motifs in artificial CUTs (Figure 2), together with our mutational analysis (Figure 3) defines these elements as novel termination signals for the NNS pathway. The CPF-CF components Hrp1p and Rna15p recognize similar sequences (Kessler *et al*, 1997; Valentini *et al*, 1999; Gross and Moore, 2001) and have been involved, directly or indirectly, in termination of snoRNA (Fatica *et al*, 2000) and CUTs (Kuehner and Brow, 2008). It was important to assess whether *in vivo* these two proteins recognize the AU-rich element in the context of the NNS pathway. Therefore, we performed northern blot analysis of RNAs derived from clone 78 expressed in P_{Gal}-*HRP1*, Δ*rrp6* or *rna15-3*, Δ*rrp6* strain. As shown in Supplementary Figure 7, no significant effect on termination was observed upon metabolic depletion of Hrp1p or temperature inactivation of Rna15p, while termination of the *SUA7* gene was clearly impaired in the same conditions, indicating that inactivation of both proteins was effective. A similar experiment performed with a P_{Gal}-*PAB1*, Δ*rrp6* also showed that depletion of Pab1p, a poly(A) binding protein, did not impinge on the recognition of the terminator (data not shown). Therefore, the impact of AU-rich elements in NNS complex-dependent termination does not relate to a possible role of Hrp1p, Rna15p or Pab1p in this pathway.

To assess whether the NNS complex recognizes this element directly, we performed fluorescence anisotropy (FA) binding assays using recombinant Nrd1p–Nab3p heterodimer (Carroll *et al*, 2007) and synthetic RNA versions of clone 78 or its mutant derivatives (Figure 5A and B). As expected, mutation of the Nrd1 and $Nab3_1$ sites significantly decreased the affinity of the NNS complex for the RNA. Importantly, mutation of the AU-rich motif to a randomly chosen sequence (clone μAU) or an $(AC)_4$ motif (data not shown) affected binding similarly to mutation of the Nrd1 site, indicating that this sequence is an important determinant of the interaction, at least *in vitro*. Replacing the AU-rich motif in clone 78 with a stretch of As (78-A_8) or Us (78-U_8) restored both termination and wild-type interaction with the complex relative to the μAU construct (Supplementary Figure 8; Figure 5).

Combining several mutations strongly decreased binding, indicating that the different motifs all contribute independently to the interaction. However, the decrease in affinity upon mutation of the different sites correlates only to a limited extent with the efficiency of termination. For instance mutations μNrd1 and μAU affect similarly binding to Nrd1p–Nab3p (Figure 5), which does not reflect the radically different behaviour of these sequences in termination assays (Figure 3A and B; Supplementary Figure 5). Also, sequences 78-μNrd1 and 78-INV-μNrd1 bound the complex with virtually identical affinity (Figure 5A and B), yet only 78-μNrd1 is able to terminate transcription (Figures 3B and 4B).

These results indicate that the NNS complex recognizes directly the AU-rich termination motif. They also strongly suggest that above a given threshold, the overall affinity of the complex for the nascent RNA is not the limiting factor for termination.

Interaction with the 3′ guanine extension of the extended Nab3p-binding site remodels the surface of Nab3p RNA-recognition motif

The strong impact on termination of the 3′-end guanine extension of the Nab3-binding site (Figure 3) prompted us to evaluate its contribution to NNS complex binding.

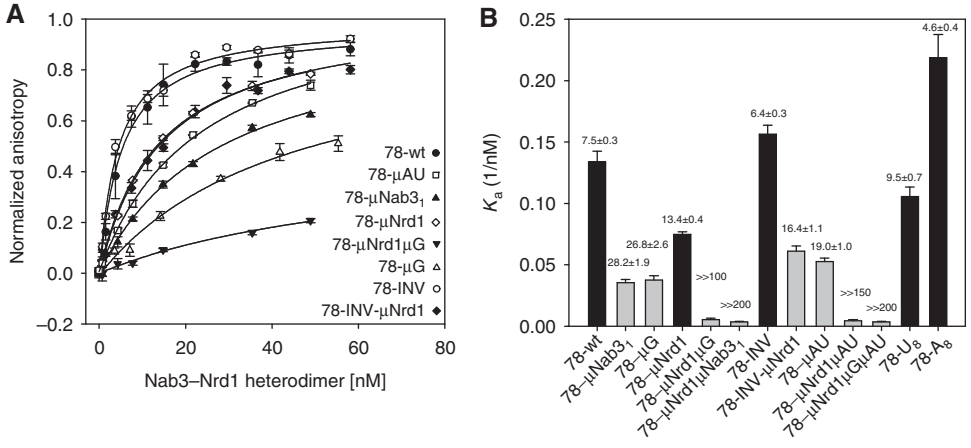


Figure 5 Fluorescence anisotropy assessment of Nrd1p–Nab3p heterodimer binding to wild-type and mutant terminators. (**A**) Binding isotherms for equilibrium binding of Nrd1p–Nab3p heterodimer to various mutants of clone 78, monitored by FA. In all, 10 nM fluorescently labelled RNA was titrated by Nrd1p–Nab3p heterodimer, the data were fitted using a single-site binding model, the data were normalized for visualization purposes. (**B**) Summary of association constants (K_a) derived from FA affinity measurements and termination proficiency for the wild-type and mutant constructs. The ionic strength and pH of the binding buffer were the same for all the measurements. The K_a values are derived from the equilibrium dissociation constants (K_d, indicated on top of every histogram, nM) calculated from the best fit to the data using a single-site binding isotherm (Heyduk and Lee, 1990). Black bars: termination proficient; grey bars: termination-deficient sequences. Each data point represents triplicate assays (error bars represent standard deviation).

Interestingly, mutation of the G alone (78-μG) caused the same decrease in affinity as the μNab3$_1$ mutation (Figure 5), suggesting that this nucleotide plays a critical role in RNA recognition by Nab3p. However, this is surprising in the light of the previously determined structure of the RNA-recognition motif (RRM) of Nab3 in complex with 5′-UCUU-3′ RNA (Hobor *et al*, 2011; Lunde *et al*, 2011), which revealed that the Nab3 RRM (321–415) specifically recognizes only the first three nucleotides of the UCUU substrate. Therefore, we undertook a more detailed structural analysis of the recognition of this nucleotide by Nab3p employing NMR spectroscopy.

The Nab3 RRM (321–415) construct previously used in structural studies showed virtually the same binding affinity to UCUU and UCUUG (data not shown) and displayed the same ^{1}H-^{15}N HSQC spectra of Nab3p RRM when bound to UCUU or UCUUG (Supplementary Figure 9). This indicates that the last G of UCUUG is not recognized by this Nab3p domain, yet it contrasts with the FA measurements obtained using the Nrd1-Nab3 heterodimer reported above (Figure 5). Since the latter experiments were performed with a larger Nab3p fragment, new constructs were designed with N- and C-terminal extensions to the RRM. We found that a construct containing a 40 amino-acids N-terminal extension of the RRM, encompassing a long α-helix (amino acids 283–415, referred to as αhxRRM throughout the text) binds to UCUUG with significantly higher affinity (37 ± 2 μM versus 170 ± 8 μM) than the original RRM construct (amino acids 321–415, data not shown). The UCUUc mutant displayed a significant drop in the affinity, confirming the importance of 3′-end guanine of the Nab3-binding site for recognition by αhxRRM (Figure 6A). Akin to binding experiments with short RNA substrates, identical results were obtained in the context of clone 78 (Figure 6A and B). Importantly, NMR titration experiments showed a very specific interaction between UCUUG and αhxRRM as evidenced by large perturbations of chemical shifts in the ^{1}H-^{15}N HSQC spectra of αhxRRM (Figure 6C and D). The largest perturbations occur in the distal N-terminal α-helix (α_D) and the β2-strand of the αhxRRM (Figure 6D). Mapping these perturbations on the previously determined structure of the Nab3 RRM (Hobor *et al*, 2011) show that the binding of UCUUG provides additional changes at the β2-strand and α-helices when compared to the binding of UCUUC (Figure 6E). This strongly indicates that the recognition of the G-containing Nab3 termination site involves an induced fit mechanism, in which α_D helix and the flanking regions are rearranged on the canonical RRM upon the RNA binding.

These results provide a mechanistic explanation for the importance of the 3′-end guanine extension of the Nab3-binding site. They also allow defining an additional motif in the Nab3p RRM that mediates a conformational change occurring upon the specific recognition of the conserved G of the Nab3-binding site.

Overlapping recognition of termination signals by the NNS- and the CPF-dependent pathways

It has been previously shown that natural CUTs can be used as terminators by the CPF-dependent pathway when mislocalized towards the 3′end of longer transcriptional units (Kopcewicz *et al*, 2007; Gudipati *et al*, 2008; Kuehner and Brow, 2008). This either suggests that the two termination pathways recognize identical or overlapping signals, or that the natural NNS terminators analysed in the previous studies are 'polluted' by independent CPF/CF non-overlapping termination signals (e.g., derived from other transcription units). Artificial NNS-dependent terminators provide a unique opportunity to distinguish between these possibilities. Clone 78 is particularly informative in this respect since it is insensitive to mutation or depletion of critical CPF/CF components (Supplementary Figure 7). This terminator is therefore unlikely to contain signals selected by virtue of a CPF-dependent selective pressure. Therefore, we set out to test whether termination signals selected based on NNS dependency could be used by the CPF pathway when moved away from the transcriptional start site. For this purpose, we modified the reporter system by inserting a 1.1-kb ORF (*LEU2*) between the P_{Tet} and P_{Gal} promoter and cloned the sequence of clone 78 and two additional artificial, NNS-dependent terminators immediately downstream of this ORF (Figure 7A). These terminators that are NNS dependent in a promoter proximal position also induced termination when localized at the 3′-end of the 1.1 kb *LEU2* gene (Figure 7A and data not shown). However, in this position stable RNAs were produced and termination was no longer dependent on the NNS pathway (Figure 7A; Supplementary Figure 10). Rather, termination was impaired by heat inactivation of the thermosensitive mutant Rna14-3p, which is an essential component of the CPF complex (Figure 7A). These results indicate that signals selected for NNS-dependent termination can be also recognized by the CPF complex in the appropriate context. Importantly, mutation of most signals that impaired NNS-dependent termination also markedly affected termination induced by the CPF pathway, indicating considerable overlap for substrate recognition by the two pathways. Interestingly, mutation of the G extending the Nab3$_1$ site was ineffective in a promoter distal location (Figure 7C, 78-μG 3′), further supporting the notion that this particular nucleotide is a major determinant of the specific recognition of termination substrates by the NNS complex. Together, these results strongly suggest that the two major yeast termination pathways recognize very similar motifs in their substrates. Recognition of these motifs is most likely operated by different factors depending on the distance from the transcriptional start site.

Discussion

The emerging concept that transcription occurs pervasively irrespective of canonical gene borders (Jacquier, 2009) raises the important question of how the cell 'protects' regulatory regions from invasive polymerases. 'Wild' polymerases have to be controlled because potentially disruptive for the expression of neighbouring genes and genome stability. Transcription termination operated by the NNS complex plays a role of utmost importance in this respect.

This work has been inspired by the consideration that the sequence motifs recognized by the NNS complex are seemingly insufficient to provide the required specificity and efficiency to the process. From our very sensitive *in vivo SELEX* approach we revealed extended Nrd1p- and Nab3p-binding sites together with novel, AU-rich motifs. We show that transcription termination only to a limited extent

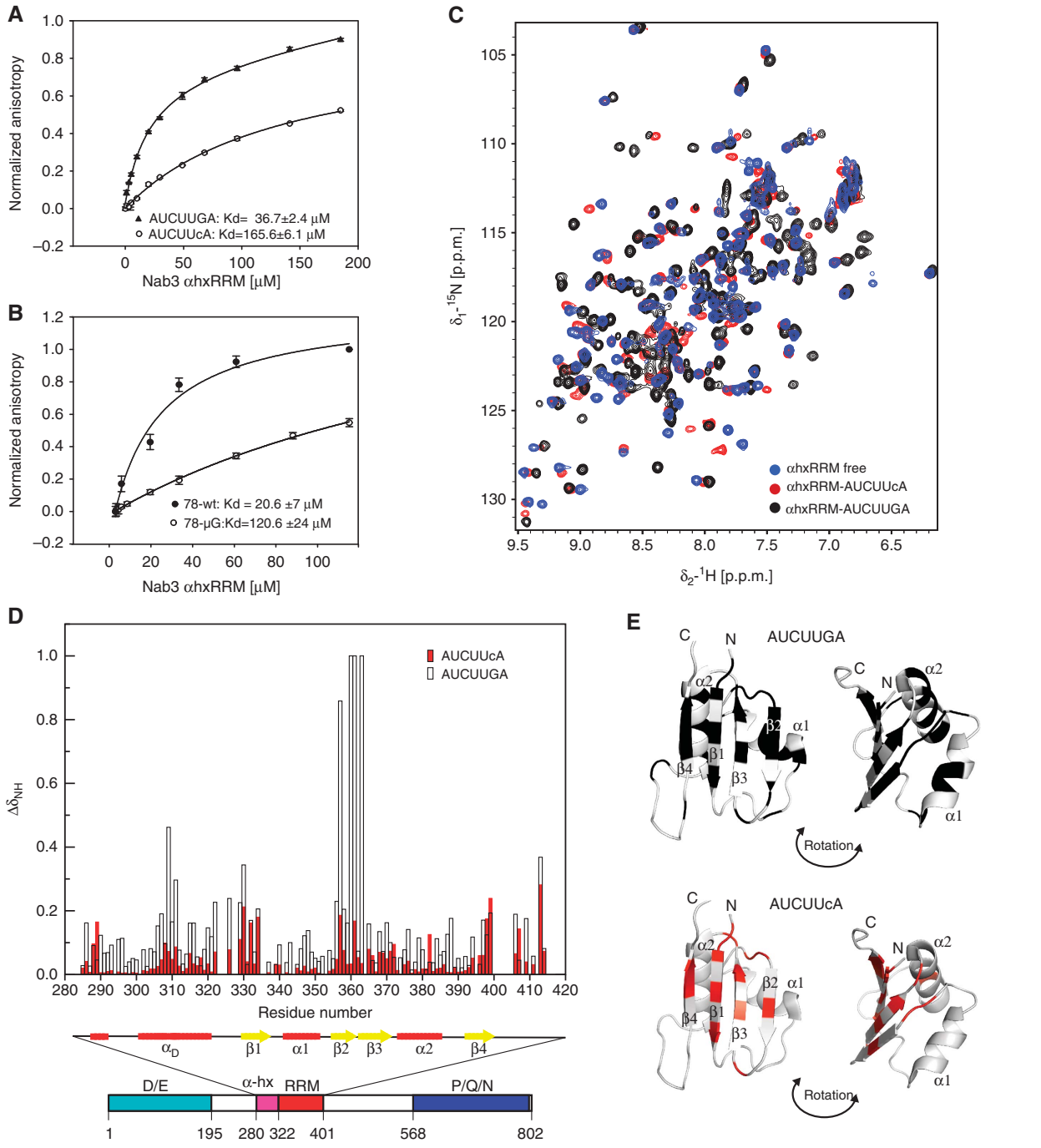


Figure 6 Specific recognition of the 3′-terminal guanine extension by an extended Nab3p RNA recognition domain. Equilibrium binding of αhxRRM with: (**A**) AUCUUGA and AUCUUcA and (**B**) clone 78, wild-type and μG mutant, monitored by FA measurements (the data were normalized for visualization purposes). Values correspond to the average of three independent experiments (error bars represent standard deviation). (**C**) ^{1}H-^{15}N HSQC NMR spectra of Nab3 αhxRRM alone (in blue) and in the presence of 1 eq of 5′-AUCUUGA-3′ (in black) or 1 eq of 5′-AUCUUcA-3′ (in red) measured on 600 MHz spectrometer at 303 K. The conformation of αhxRRM undergoes significant changes upon binding to AUCUUGA as reflected in the chemical shift changes when titrated with AUCUUGA RNA. The chemical changes are less significant upon binding to AUCUUcA. (**D**) Quantification of chemical shift perturbations of Nab3 αhxRRM upon binding to AUCUUGA (black) or AUCUUcA (red) RNA. The combined chemical shift perturbations ($([w_{HN}\Delta\delta_{HN}]^2+[w_N\Delta\delta_N]^2)^{1/2}$), where $w_{HN}=1$ and $w_N=0.154$ are weight factors of the nucleus, are plotted versus the residue numbers and schematic figure of secondary structure elements. Significant changes occur in the β1 and β2 strands and in the distal α-helical element (α_D). (**E**) Chemical shift changes (Δδ) upon RNA binding mapped to the surface of the Nab3 RRM structure (Hobor *et al*, 2011). The upper part shows Δδ of αhxRRM with AUCUUGA RNA (black), and the lower part shows Δδ of αhxRRM with AUCUUCA RNA (red). Only $\Delta\delta>0.075$ are displayed.

correlates with the affinity of NNS complex binding and that the arrangement of motifs is a critical parameter. Importantly, Nab3p binds its extended site by an induced fit mechanism whereby a novel module distinct from the canonical RRM specifically recognizes a strongly conserved G in the Nab3 site. Finally, we show that the same sequence motifs selected

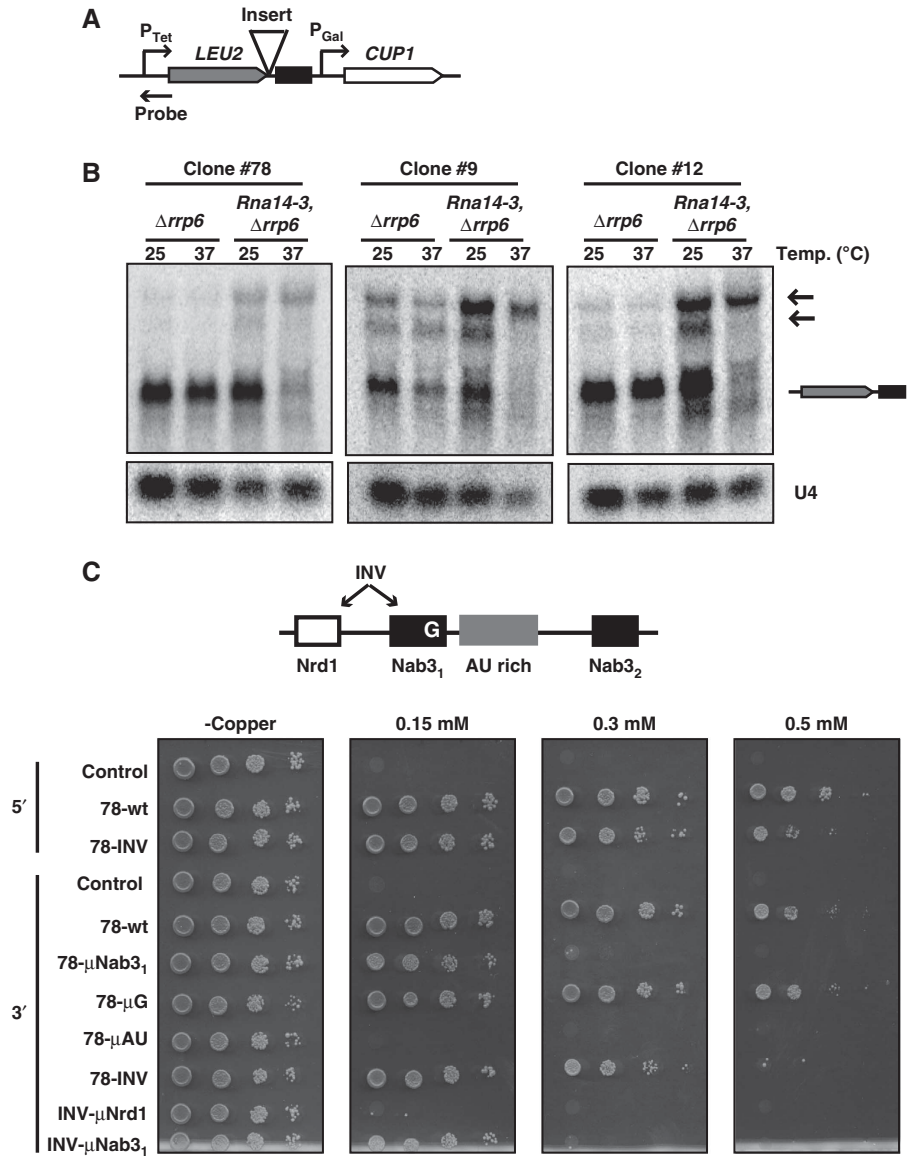


Figure 7 The NNS and the CPF pathways have largely overlapping sequence determinants. (**A**) A scheme of the reporter construct containing the 1.1-kb *LEU2* gene inserted between the P_{Tet} and the P_{Gal} promoters. Artificial CUT sequences were cloned by recombination downstream of *LEU2* as indicated. A stuffer fragment from the *HSP104* gene (486 nt) is present after the terminators insertion point as polymerases committed to CPF-dependent termination are known to transcribe downstream of the terminator. (**B**) Northern blot analysis of RNAs derived from clone 78 and two additional artificial terminators cloned downstream of *LEU2*. These constructs were expressed in an *rna14-3* strain to assess the dependency on the CPF pathway. A Δ*rrp6* strain was employed to stabilize read-through transcripts generated by mutation of Rna14p. (**C**) Copper growth assay of cells expressing the constructs containing the artificial terminators cloned downstream of *LEU2* as indicated (3′). A scheme indicates the positions of the relevant motifs. Sequences are reported in Figures 3A and 4A. For comparison, clones 78 and 78-INV inserted in a 5′ position (inducing termination by the NNS pathway) are also included in the test set as indicated.

for NNS-dependent termination can also be recognized by the CPF pathway indicating that the two termination pathways have adapted to recognize largely overlapping signals in spite of the very different factors involved and the different fate of the RNAs produced.

In vivo selection of artificial terminators

Our *in vivo* SELEX strategy overcomes many of the limitations inherent to the search for termination motifs in natural CUTs. First, we could use a robust neutral model by estimating background words frequencies in a non-selected pool of >8000 sequences. Second, sequences were specifically selected for their ability to induce transcription termination, thus limiting alternative selective pressures generating motifs that might pollute the statistical analyses. From our analysis of the winning pool we can roughly estimate the 'evolutionary cost' to generate a functional, NNS-dependent terminator. This could be a difficult task for the disperse nature of the information present in these terminators. However, we also selected another class of terminators that contains a single motif of 8 nt that is necessary and sufficient for termination (Colin *et al*, in preparation). Because the two classes of terminators have approximately equal abundance, the informational content of NNS-dependent terminators is equivalent

to 'fixing' 8 contiguous nucleotides, that is, roughly 16 bits of information (Schneider *et al*, 1986). Since NNS terminators contain split motifs, this is consistent with the notion that more than two sites of four nucleotides are required to induce termination.

Statistical analysis of overrepresented motifs indicates that the GUAG sequence, a known Nrd1p-binding site, was enriched to a lesser extent relative to the other binding sites of the complex, indicating that this particular motif is less important for binding or termination than GUAA (Nrd1p binding) or UCUU (Nab3p binding). We also found that these tetranucleotides are actually part of extended binding sites, U/AGUAAA and UCUUG, respectively, considerably increasing the informational content of individual termination motifs. These extended sites are similar but not identical to the consensus derived from *in vivo* crosslinking data (respectively UGUAG and GNUUCUGU for Nrd1p and Nab3; Creamer *et al*, 2011). The differences might pertain to crosslinking biases (which tend to favour Us) or to the different neutral model used to evaluate the statistical significance as discussed above. The presence of a G extending the Nab3 site was first noticed in experiments detecting RNAs associated with the NNS complex (Hogan *et al*, 2008) and confirmed in more recent crosslinking approaches (Creamer *et al*, 2011; Wlotzka *et al*, 2011). We show here that mutation of the G extending the Nab3 site is highly disruptive both for termination and for NNS complex binding. Importantly, we provide a mechanistic explanation for this effect by showing that this nucleotide is specifically recognized by long N-terminal α-helical extension of the Nab3p RRM and that binding provokes changes in the structure of the protein. These structural changes impact the strength and the geometry of the interaction and might explain why CUUG is strongly preferred as a Nab3p crosslinking site relative to UCUU (Wlotzka *et al*, 2011). It is certainly possible that the structural perturbations induced by RNA binding, besides affecting the strength and specificity of the interaction, are transmitted to other domains of the protein, allosterically altering the function of the complex in termination.

AU-rich motifs are prominent NNS complex-dependent termination signals

Remarkably, we found that AU-rich motifs are strongly overrepresented and functionally important in our set of synthetic terminators to similar or higher levels than 'canonical' Nrd1p-binding sites. We provide important clues on the mechanistic impact of this motif in termination. We show that it cannot be ascribed to recognition by Hrp1p or Rna15p, two proteins involved in the CPF pathway and known to bind AU-rich motifs (Kessler *et al*, 1997; Valentini *et al*, 1999; Gross and Moore, 2001; Mandel *et al*, 2008; Supplementary Figure 7). Rather, this motif contributes to NNS complex binding to a higher extent than a GUAA Nrd1p-binding site (Figure 5) and can be substituted, both for binding and for termination, by stretches of As or Us. This strongly suggests that recognition of AU-rich termination signals depends on the NNS complex, although we cannot exclude that another factor also recognizes these motifs concomitantly or sequentially *in vivo*.

These findings are relevant to the function of natural terminators. One of the CUTs generated by upstream transcription initiation at the *IMD2* locus is paradigmatic in this respect. This strong NNS pathway terminator (starting at position −67 relative to the *IMD2* AUG) (Kuehner and Brow, 2008) contains only one UCUU motif that is necessary but not sufficient for termination. However, it also contains an upstream CUUG motif and a downstream, prominent, AU-rich region, whose sequence is identical to that selected in clone 78. Importantly, random mutations that affect termination were identified within these motifs (Kuehner and Brow, 2008 and our unpublished results), which was previously unexplained and can now be rationalized in the light of our results.

The high AT richness of intergenic regions (roughly 66%) is thought to contribute to the establishment of nucleosome-free regions (NFRs), from where transcription generally originates. We suggest that the AT richness of NFRs also has an additional role in providing a favourable background for termination signals that could be generated at a low evolutionary cost. Thus, the same sequence background would favour the generation of transcription and protect it from interference due to 'invading' polymerases. The unexplained occurrence of 3′ NFRs between convergent genes (Neil *et al*, 2009; Xu *et al*, 2009) could even be due to a secondary effect of the AT richness of termination signals.

Significance of termination supermotifs

Our analysis of synthetic terminators reveals the statistically significant presence of supermotifs containing canonical NNS complex binding sites associated to AU-rich motifs and we show by mutational analysis that this association is functionally important. It is possible that the close proximity of the AU-rich motif favours cooperative binding of the complex to the RNA as previously suggested for closely positioned Nrd1 and Nab3 sites (Carroll *et al*, 2007). However, our experiments also indicate that the overall affinity of the complex for the terminator is not necessarily limiting for termination. Indeed, constructs 78-μNrd1 and 78-INV-μNrd1 bind the complex with very similar affinity (Figure 5), yet only the latter (in which the association Nrd1 site-AU-rich motif is lost) cannot induce transcription termination. This indicates that it is the local environment determined by a critical arrangement of motifs that favours termination, rather than the global load of NNS complex on the nascent RNA. 'Hot spots' of termination could therefore be determined by the local high affinity recognition of the supermotif, possibly in association with additional factors.

It is also possible that termination requires an additional or alternative read-out of the AU-rich motif than simple NNS complex binding, possibly by the elongating polymerase. For instance, it is possible that AT-rich regions induce RNAPII pausing and that the close juxtaposition of a pausing element with protein binding on the nascent RNA elicits termination. This would be reminiscent of the mechanism of intrinsic termination in bacteria whereby a T-stretch inducing the pause is immediately preceded by a GC-rich hairpin. Base pairing within the stem disrupts critical polymerase–nucleic acid contacts in the elongation complex and induces termination (Peters *et al*, 2011). Binding of the NNS complex might be analogous to hairpin formation in sequestering the sequence immediately upstream of the polymerase.

Termination supermotifs might be essential for short CUTs, as for the *IMD2* and *URA2* CUTs (Figure 2), and the small size

of our artificial CUTs might have favoured the prominent selection of these motifs. We suggest that, in the absence of termination supermotifs, multiple, less efficient termination event occur independently, the combination of which is required to fully prevent polymerase read through. This might explain the requirement for multiple NNS complex binding sites previously demonstrated for *NEL025c* and auto-regulation of the *NRD1* gene (Thiebaut *et al*, 2006; Arigo *et al*, 2006a).

Overlapping termination signals are recognized by the CPF and NRD1 pathway

Because most of our artificial terminators led to the production of Rrp6-sensitive transcripts and were sensitive to the NNS pathway, it is unlikely that CPF-dependent terminators are significantly present in our winning set. We suggest that these results from the predominance of the position effect that strongly favours NNS complex-dependent termination proximally to the transcription start site (Jenks and Reines, 2005; Steinmetz *et al*, 2006; Kopcewicz *et al*, 2007; Gudipati *et al*, 2008). However, in spite of a selective pressure favouring NNS complex-dependent termination (as verified for clone 78), these terminators also contain CPF signals as they are recognized by the CPF complex when re-localized at >1 kb from the transcription start site (Figure 7). Natural NNS-dependent terminators can be recognized by the CPF complex when mislocalized (Jenks and Reines, 2005; Steinmetz *et al*, 2006; Kopcewicz *et al*, 2007; Gudipati *et al*, 2008). However, whether these natural sequences contain independent and non-overlapping signals directing termination by either pathway is unclear, and even suggested in the case of the *SNR13* terminator (Steinmetz *et al*, 2006). Strikingly, we show that termination signals for the two pathways largely overlap. Indeed, not only the AU-rich region is required for CPF-dependent termination (which could have been expected) (Valentini *et al*, 1999; Gross and Moore, 2001; Proudfoot, 2011) but also the integrity of the Nrd1 and Nab3 sites. The only exception is mutation of the G residue mediating Nab3p-specific contacts, further underscoring the functional importance of this residue. This finding is remarkable in the light of the different machineries involved in the two pathways and the different fate of the RNAs produced. It is theoretically possible that a common 'recognition module' exists that is shared by the two termination complexes. However, this is not supported by experimental evidence since, with the exception of a few *pcf11* alleles, mutations in factors belonging to one given pathway do not affect significantly the other (Kim *et al*, 2006). We suggest that the two termination pathways have adapted independently to recognize highly similar signals, possibly converging on sequences that alter the processivity of the polymerase or favour its propensity to terminate. The fact that termination signals are bi-functional has important functional implications for the control of pervasive transcription. They would constitute efficient transcriptional insulators to halt both polymerases initiating in the immediate vicinity, for example, producing 5′ or 3′ ORF-overlapping CUTs, as well as polymerases deriving from remote initiation events, for example, reading through termination signals of neighbouring ORFs. Use of the same signals for both termination mechanisms would be highly economical in terms of evolutionary cost for a compact genome as that of *S. cerevisiae*.

Materials and methods

Construction of yeast strains, standard molecular biology analyses and proteins purification procedures are reported in Supplementary methods. Yeast strains, plasmids and oligonucleotides used in this work are listed in Supplementary Tables 1–3.

Generation of the pool of random sequences and in vivo selection

The library of random sequences for the *in vivo* selection was generated from two chemically synthesized oligonucleotides containing a 60-nt variable region flanked by two constant regions of 13 nt (5′-end) and 20 nt (3′-end) (oligonucleotides DL1698 and DL1665, Supplementary Table 3). In all, 500 pmol of each oligonucleotide was annealed over their 20 nt complementary 3′-ends and filled in using the klenow fragment of *E. coli* polymerase I. After purification, the mixture was PCR amplified with primers DL1702 and DL1666 (Supplementary Table 3) that anneal to the 5′ constant regions of DL1698 and DL1665 and extend the homology region for subsequent cloning of the random pool by recombination. The final pool contains two regions of random sequence (A and B, Figure 1A) separated by a constant segment of 20 nt. Transformation of a *Δcup1* yeast strain with the pool yielded roughly 10^5 colonies. These were directly replica plated on galactose medium containing 0.3 mM copper for selection based on expression of the *CUP1* gene under control of the P_{Gal} promoter on the reporter. Copper-resistant clones were pooled, the inserts were amplified by PCR with oligonucleotides directed against the constant regions and cloned by recombination in pDL367 for a second round of selection. This selection strategy effectively minimized the emergence of false positives due to rearrangements of the vector, generally leading to loss of the doxycycline repressible promoter. Roughly 100 of the copper-resistant clones were sequenced manually before large scale sequencing. Fifteen clones were subjected to northern blot analysis to verify the occurrence of termination between P_{tet} and P_{gal}.

Deep sequencing and statistical analyses

Inserts from the winning or the starting pools were amplified by PCR using primers containing a two nucleotides barcode. The primers included adaptors for flow cell amplification and annealing sites for sequencing primers. Paired ends sequencing was performed on an Illumina GAIIx platform. Sequences of parts A and B were coupled with a home-made algorithm. Roughly 30% of sequences of the winning pool were excluded from the analysis because containing NNS complex-independent terminators (Colin *et al*, in preparation). The remaining roughly 700 distinct sequences were submitted to subsequent statistical analyses with RSAT (van Helden, 2003). Motifs of four and five nucleotides that are significantly overabundant and constitute potential termination signals were identified by comparing frequencies observed in the winning pool relative to frequencies observed in the starting pool (8000 sequences). Since the two oligonucleotides used to generate parts A and B of the starting pool are on the opposite strand, the nucleotide bias due to the chemical synthesis is different in the two regions as assessed from sequencing (part A: 22.7% (A); 19.4% (C); 29.2% (G) and 28.7% (T); part B: 31.3% (A); 26.8% (C); 18.6% (G) and 23.1% (T)). Since this strongly influences the frequency of each motif in the starting pool, we adopted a separate background model for parts A and B of the pool. *P*-values were calculated by RSAT with a correction for overlapping occurrences. Extended motifs were identified from overrepresented hexanucleotides using the pattern assembly and convert-matrix tools of RSAT. Pattern assembly aligns overlapping overrepresented motifs to generate larger elements that can be converted to matrices. Logos are generated by RSAT based on these matrices. *P*-values for the overrepresentation of TCTTG in natural CUTs (Gudipati *et al*, submitted) and underrepresentation in ORFs have been calculated estimating expected frequencies from input sequences (CUTs or ORFs) with a Markov chain model of order 2.

FA assays

The equilibrium binding of Nab3 RRM, Nab3 αhxRRM and the $Nab3_{191-565}$–$Nrd1_{1-548}$ heterodimer to their specific substrates was analysed by FA. The RNA oligonucleotides were 5′ fluorescein labelled. The labelled RNA oligonucleotides were purchased from Sigma-Aldrich, the lyophilized samples were dissolved in water. The measurements were conducted on a FluoroMax-4 spectrofluorometer (Horiba Jobin-Yvon, USA). The instrument was equipped with a thermostatted cell holder with a Neslab RTE7 water bath (Thermo Scientific, USA). The whole system was operated using FluorEssence software (version 2.5.3.0, Horiba Jobin-Yvon). The fluorescein fluorophore was excited at 488 nm and its emission was collected at 520 nm. The width of both excitation and emission monochromatic slits was 14 nm for the 1 nM substrates, and 8 nm for the 10 nM substrates, the integration time was set to 3 s in both cases. All experiments were carried out at 25°C in a stirred 1.9 ml quartz cuvette. A fixed delay of 30 s was set between each aliquot addition and start of the measurement to allow the reaction to reach equilibrium. This delay was sufficient, as no further change in anisotropy was observed. Every data point is an average of three measurements.

The data were analysed in SigmaPlot 11 software (Systat Software, USA). The experimental isotherms were fit to a single-site binding model according to Heyduk and Lee (1990) using non-linear least squares regression or with single-site saturation binding model. The data were normalized for visualization purposes.

NMR analyses

All NMR spectra of 2.5 mM uniformly ^{15}N-labelled Nab3 ahxRRM and RRM in 50 mM sodium phosphate buffer (pH 8.0), 300 mM NaCl, 10 mM β-mercaptoethanol (90% H_2O/10% D_2O) were recorded on Bruker AVANCE 600 spectrometer equipped with a cryoprobe at a sample temperature of 30°C. All ^{1}H-^{15}N HSQC spectra were acquired with 8 scans, 1024 points in ^{1}H and 256 increments in ^{15}N dimension, processed standardly with TopSpin (Bruker BioSpin) and analysed in Sparky (Goddard, TD and Kneller, DG, SPARKY 3, University of California, San Francisco, USA). Spectra for NMR titration were measured on a ^{15}N isotopically enriched αhxRRM and unlabelled AUCUUG/CA, and on a ^{15}N isotopically enriched RRM and unlabelled UCUU or UCUUG. In all titrations, RNA was added stepwise (in 4–6 steps) in small volumes into protein solutions. For each addition of the RNAs, the ^{1}H-^{15}N HSQC spectra were acquired. In the course of the titrations, the resonances moved from their initial positions, which correspond to the free form, in a stepwise directional manner until they reached their final positions, which correspond to the fully bound state. These data indicate that, in all complexes, the proteins are in fast exchange between their free and bound forms relative to the NMR time scale.

Supplementary data

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

Acknowledgements

We would like to thank F Lacroute for help with yeast genetics, D Gautheret and T Barthel for help in analysing deep sequencing data and T Villa for critical reading of the manuscript. This work was supported by the Danish National Research Foundation (DL), the ANR (DL, ANR-08-Blan-0038-01), the CNRS (DL), the project 'CEITEC—Central European Institute of Technology' (RS, CZ.1.05/1.1.00/02.0068) from European Regional Development Fund, Czech Science Foundation (RS, P305/12/G034, P305/10/1490). OP and RKG received fellowships from EMBO and Région Ile de France, respectively. FH is in receipt of the Brno City Municipality Scholarship for Talented PhD Students. This research was carried out within the scope of the Associated European Laboratory LEA 'Laboratory of Nuclear RNA Metabolism'. This work has benefited from the facilities and expertise of the high throughput sequencing platform of IMAGIF (Centre de Recherche de Gif—www.imagif.cnrs.fr).

Author contributions: OP designed and performed experiments and wrote the paper. FH designed and performed experiments. JB performed experiments. KK performed experiments and analysed the data. YD-C analysed the data. RKG performed experiments. RS designed experiments, analysed the data and wrote the paper. DL designed experiments, analysed the data and wrote the paper.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Arigo JT, Carroll KL, Ames JM, Corden JL (2006a) Regulation of yeast NRD1 expression by premature transcription termination. *Mol Cell* **21:** 641–651

Arigo JT, Eyler DE, Carroll KL, Corden JL (2006b) Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. *Mol Cell* **23:** 841–851

Carroll KL, Ghirlando R, Ames JM, Corden JL (2007) Interaction of yeast RNA-binding proteins Nrd1 and Nab3 with RNA polymerase II terminator elements. *RNA* **13:** 361–373

Carroll KL, Pradhan DA, Granek JA, Clarke ND, Corden JL (2004) Identification of cis elements directing termination of yeast nonpolyadenylated snoRNA transcripts. *Mol Cell Biol* **24:** 6241–6252

Chlebowski A, Tomecki R, Lopez ME, Seraphin B, Dziembowski A (2011) Catalytic properties of the eukaryotic exosome. *Adv Exp Med Biol* **702:** 63–78

Creamer TJ, Darby MM, Jamonnak N, Schaughency P, Hao H, Wheelan SJ, Corden JL (2011) Transcriptome-wide binding sites for components of the Saccharomyces cerevisiae non-poly(A) termination pathway: Nrd1, Nab3, and Sen1. *PLoS Genet* **7:** e1002329

Davis CA, Ares Jr M (2006) Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in Saccharomyces cerevisiae. *Proc Natl Acad Sci USA* **103:** 3262–3267

Egecioglu DE, Henras AK, Chanfreau GF (2006) Contributions of Trf4p- and Trf5p-dependent polyadenylation to the processing and degradative functions of the yeast nuclear exosome. *RNA* **12:** 26–32

Fatica A, Morlando M, Bozzoni I (2000) Yeast snoRNA accumulation relies on a cleavage-dependent/polyadenylation-independent 3′-processing apparatus. *EMBO J* **19:** 6218–6229

Gross S, Moore CL (2001) Rna15 interaction with the A-rich yeast polyadenylation signal is an essential step in mRNA 3′-end formation. *Mol Cell Biol* **21:** 8045–8055

Gudipati RK, Villa T, Boulay J, Libri D (2008) Phosphorylation of RNA polymerase CTD dictates transcription termination choice. *Nat Struct Mol Biol* **15:** 786–794

Heyduk T, Lee JC (1990) Application of fluorescence energy transfer and polarization to monitor Escherichia coli cAMP receptor protein and lac promoter interaction. *Proc Natl Acad Sci USA* **87:** 1744–1748

Hobor F, Pergoli R, Kubicek K, Hrossova D, Bacikova V, Zimmermann M, Pasulka J, Hofr C, Vanacova S, Stefl R (2011) Recognition of transcription termination signal by the nuclear polyadenylated RNA-binding (NAB) 3 protein. *J Biol Chem* **286:** 3645–3657

Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO (2008) Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol* **6:** e255

Houalla R, Devaux F, Fatica A, Kufel J, Barrass D, Torchet C, Tollervey D (2006) Microarray detection of novel nuclear RNA substrates for the exosome. *Yeast* **23:** 439–454

Jacquier A (2009) The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet* **10:** 833–844

Jenks MH, Reines D (2005) Dissection of the molecular basis of mycophenolate resistance in Saccharomyces cerevisiae. *Yeast* **22:** 1181–1190

Johnson JM, Edwards S, Shoemaker D, Schadt EE (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet* **21:** 93–102

Kessler MM, Henry MF, Shen E, Zhao J, Gross S, Silver PA, Moore CL (1997) Hrp1, a sequence-specific RNA-binding protein that shuttles between the nucleus and the cytoplasm, is required for mRNA 3′-end formation in yeast. *Genes Dev* **11:** 2545–2556

Kim M, Vasiljeva L, Rando OJ, Zhelkovsky A, Moore C, Buratowski S (2006) Distinct pathways for snoRNA and mRNA termination. *Mol Cell* **24:** 723–734

Kopcewicz KA, O'Rourke TW, Reines D (2007) Metabolic regulation of IMD2 transcription and an unusual DNA element that generates short transcripts. *Mol Cell Biol* **27:** 2821–2829

Kuehner JN, Brow DA (2008) Regulation of a eukaryotic gene by GTP-dependent start site selection and transcription attenuation. *Mol Cell* **31:** 201–211

Kuehner JN, Pearson EL, Moore C (2011) Unravelling the means to an end: RNA polymerase II transcription termination. *Nat Rev Mol Cell Biol* **12:** 283–294

LaCava J, Houseley J, Saveanu C, Petfalski E, Thompson E, Jacquier A, Tollervey D (2005) RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell* **121:** 713–724

Lebreton A, Seraphin B (2008) Exosome-mediated quality control: substrate recruitment and molecular activity. *Biochim Biophys Acta* **1779:** 558–565

Lunde BM, Horner M, Meinhart A (2011) Structural insights into cis element recognition of non-polyadenylated RNAs by the Nab3-RRM. *Nucleic Acids Res* **39:** 337–346

Mandel CR, Bai Y, Tong L (2008) Protein factors in pre-mRNA 3′-end processing. *Cell Mol Life Sci* **65:** 1099–1122

Mayer A, Lidschreiber M, Siebert M, Leike K, Soding J, Cramer P (2011) Uniform transitions of the general RNA polymerase II transcription complex. *Nat Struct Mol Biol* **17:** 1272–1278

Millevoi S, Vagner S (2011) Molecular mechanisms of eukaryotic pre-mRNA 3′ end processing regulation. *Nucleic Acids Res* **38:** 2757–2774

Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A (2009) Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457:** 1038–1042

Peters JM, Vangeloff AD, Landick R (2011) Bacterial transcription terminators: the RNA 3′-end chronicles. *J Mol Biol* **412:** 793–813

Proudfoot NJ (2011) Ending the message: poly(A) signals then and now. *Genes Dev* **25:** 1770–1782

Schmid M, Jensen TH (2008) The exosome: a multipurpose RNA-decay machine. *Trends Biochem Sci* **33:** 501–510

Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. *J Mol Biol* **188:** 415–431

Steinmetz EJ, Conrad NK, Brow DA, Corden JL (2001) RNA-binding protein Nrd1 directs poly(A)-independent 3′-end formation of RNA polymerase II transcripts. *Nature* **413:** 327–331

Steinmetz EJ, Ng SB, Cloute JP, Brow DA (2006) cis- and trans-Acting determinants of transcription termination by yeast RNA polymerase II. *Mol Cell Biol* **26:** 2688–2696

Thiebaut M, Colin J, Neil H, Jacquier A, Seraphin B, Lacroute F, Libri D (2008) Futile cycle of transcription initiation and termination modulates the response to nucleotide shortage in S. cerevisiae. *Mol Cell* **31:** 671–682

Thiebaut M, Kisseleva-Romanova E, Rougemaille M, Boulay J, Libri D (2006) Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. *Mol Cell* **23:** 853–864

Tietjen JR, Zhang DW, Rodriguez-Molina JB, White BE, Akhtar MS, Heidemann M, Li X, Chapman RD, Shokat K, Keles S, Eick D, Ansari AZ (2011) Chemical-genomic dissection of the CTD code. *Nat Struct Mol Biol* **17:** 1154–1161

Valentini SR, Weiss VH, Silver PA (1999) Arginine methylation and binding of Hrp1p to the efficiency element for mRNA 3′-end formation. *RNA* **5:** 272–280

van Helden J (2003) Regulatory sequence analysis tools. *Nucleic Acids Res* **31:** 3593–3596

Vanacova S, Wolf J, Martin G, Blank D, Dettwiler S, Friedlein A, Langen H, Keith G, Keller W (2005) A new yeast poly(A) polymerase complex involved in RNA quality control. *PLoS Biol* **3:** e189

Vasiljeva L, Kim M, Mutschler H, Buratowski S, Meinhart A (2008) The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. *Nat Struct Mol Biol* **15:** 795–804

Wlotzka W, Kudla G, Granneman S, Tollervey D (2011) The nuclear RNA polymerase II surveillance system targets polymerase III transcripts. *EMBO J* **30:** 1790–1803

Wyers F, Rougemaille M, Badis G, Rousselle JC, Dufour ME, Boulay J, Regnault B, Devaux F, Namane A, Seraphin B, Libri D, Jacquier A (2005) Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121:** 725–737

Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Munster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457:** 1033–1037

Published online 8 March 2012

Nucleic Acids Research, 2012, Vol. 40, No. 12 ***5679–5693***
doi:10.1093/nar/gks223

Air2p is critical for the assembly and RNA-binding of the TRAMP complex and the KOW domain of Mtr4p is crucial for exosome activation

Peter Holub, Jana Lalakova, Hana Cerna, Josef Pasulka, Marie Sarazova, Kristyna Hrazdilova, Maria Sanudo Arce, Fruzsina Hobor, Richard Stefl* and Stepanka Vanacova*

CEITEC—Central European Institute of Technology, Masaryk University, Kamenice 5, CZ-62500 Brno, Czech Republic

Received November 16, 2011; Revised and Accepted February 21, 2012

ABSTRACT

Trf4/5p-Air1/2p-Mtr4p polyadenylation complex (TRAMP) is an essential component of nuclear RNA surveillance in yeast. It recognizes a variety of nuclear transcripts produced by all three RNA polymerases, adds short poly(A) tails to aberrant or unstable RNAs and activates the exosome for their degradation. Despite the advances in understanding the structural features of the isolated complex subunits or their fragments, the details of complex assembly, RNA recognition and exosome activation remain poorly understood. Here we provide the first understanding of the RNA binding mode of the complex. We show that Air2p is an RNA-binding subunit of TRAMP. We identify the zinc knuckles (ZnK) 2, 3 and 4 as the RNA-binding domains, and reveal the essentiality of ZnK4 for TRAMP4 polyadenylation activity. Furthermore, we identify Air2p as the key component of TRAMP4 assembly providing bridging between Mtr4p and Trf4p. The former is bound via the N-terminus of Air2p, while the latter is bound via ZnK5, the linker between ZnK4 and 5 and the C-terminus of the protein. Finally, we uncover the RNA binding part of the Mtr4p arch, the KOW domain, as the essential component for TRAMP-mediated exosome activation.

INTRODUCTION

Eukaryotic cells produce diverse types of protein-coding as well as noncoding RNAs, some of which have been recently shown to have direct regulatory role in the regulation of gene expression and chromatin structure [reviewed in (1)]. Most, if not all, RNAs are synthesized as a precursor molecule that needs to be posttranscriptionally processed and/or modified in order to form mature functional molecules. Several human genetic disorders relate to abnormalities in RNA processing and proper ribonucleoprotein assembly, such as cancer or neurodegenerative diseases (2). In yeast, nuclear RNA maturation and stability are under strict control from RNA surveillance carried out by the nuclear exosome and its cofactor the TRAMP4/5 polyadenylation complex (3–6).

The TRAMP recognizes a variety of nuclear transcripts produced by all three RNA polymerases, such as precursors of small nuclear and nucleolar RNAs (snRNAs and snoRNAs), ribosomal RNAs (rRNAs), tRNAs, as well as telomeric, and cryptic unstable transcripts (CUTs) (6–8). It adds short poly(A) tails to aberrant transcripts, creating a favorable substrate for the exosome (3–6,9), reviewed in (10). Among the complex subunits, Mtr4p provides helicase activity and Trf4p or Trf5p provide polyA-polymerase (PAP) activity, Air2p and Air1p are zinc-knuckle (ZnK) proteins with a predicted RNA-binding role within TRAMPs. The degradation presumably involves multiple polyadenylation cycles and the helicase activity of Mtr4p (5). In some instances, the RNA degradation does not require the polyadenylation activity of TRAMP (7,11,12). The mechanistic understanding of the substrate recognition and exosome activation is currently very limited.

The minimal TRAMP PAP consists of a heterodimer of one of the Trf4 or Trf5 proteins and either Air1p or Air2p (4–6). Trf4p and Trf5p are the catalytic subunits of the complex, whereas Air1p and Air2p are predicted to provide RNA-binding. Air proteins contain five tandemly arranged ZnK motifs of the CCHC-type ($CX_2CX_4HX_4C$) that have previously been shown to

*To whom correspondence should be addressed. Tel: +420 549495042; Fax: +420 549492556; Email: vanacova@chemi.muni.cz
Correspondence may also be addressed to Richard Stefl. Tel: +420 549492436; Fax: +420 549492556; Email: richard.stefl@ceitec.muni.cz

bind nucleic acids and proteins in unrelated viral proteins (13). The ZnK regions in these proteins recognize exposed guanosines of RNA loops or in single-stranded RNA (13). The RNA helicase Mtr4p is not required for the polyadenylation activity of TRAMP; however, it regulates the length of a poly(A) tail added by Trf4p-Air2p dimer and modulates its polyadenylation rate *in vitro* (14).

Recently, three works reported on the structure of the full-length Mtr4p and a heterodimer of Trf4p-Air2p fragments (15–17). The structures of Mtr4p revealed a five-domain organization in which four domains constitute the DExH helicase core [RecA-1, RecA-2, winged helix (WH) and a helical bundle] (16,17). Compared to known structures of other DNA and RNA helicases, Mtr4p DExH core has an N-terminal β-hairpin extension located on the surface of the globular core. The most unique feature however is an insertion of 265 amino acids between the WH helices protruding from the globular core to form the arch domain with an arm-like structure and a fist (also termed the KOW domain) at the end. The KOW domain shares similarities with rRNA-binding proteins and contributes to RNA binding *in vitro* (17). The arch is dispensable for Mtr4 helicase and ATPase activities and for TRAMP integrity *in vitro*. A strain expressing an archless Mtr4p displays a slow growth phenotype that seems to result from an Rrp6p-dependent incomplete processing of 5.8S rRNA (16).

The study of the structure of the minimal Trf4p-Air2p heterodimer, covering the catalytic and central domains of Trf4p and only the last two ZnKs of Air2p, revealed a phylogenetically conserved surface on the central domain of Trf4p that contacts the last linker and ZnK region of Air2p (15). Nevertheless, this study still left many open questions such as what is the role of the other Air2p ZnK domains and the N- and C- terminal regions of Air2p and Trf4p. Finally, the RNA-binding mode of Trf4p-Air2p has been addressed only indirectly. The structural determination of the entire complex has proven to be difficult so far due to the instability of recombinant subunits. Moreover, it still remains a mystery as to how TRAMP interacts with and activates the downstream acting exosome. Thus despite the advances in our understanding of TRAMP function *in vivo* and of the individual structures, there is still very limited knowledge on the assembly of the complex and of the mechanistic details of its action.

Here, we have focused on two aspects of the TRAMP4 biochemistry. On one hand, we aimed to identify the code of recognition of structurally complicated substrates that need to be polyadenylated by TRAMP4 before exosome degradation, such as the hypomodified $tRNA_i^{Met}$. Second, to tackle the TRAMP4 complex assembly and molecular mechanisms of its function, we have performed detail analyses of Trf4p, Air2p and Mtr4p domain organization *in vitro* and *in vivo*. Furthermore, we succeeded to obtain the solution structure of the five zinc-knuckle region of Air2p by NMR spectroscopy. Overall, we uncovered regions required for cell viability, protein–protein interactions, RNA binding domains and regions responsible for exosome activation.

MATERIALS AND METHODS

Yeast cultures and manipulation

Yeast were cultured under standard conditions in media with selective markers corresponding to particular strains and vectors. Yeast transformations were performed by lithium acetate method. For growth test analyses, yeast were cultured overnight in minimal SD medium at 30°C, diluted with fresh SD media for OD_{600} 0.2 and grown at 30°C to OD_{600} 1. The cells were then serially diluted in decimal dilutions and incubated at 17°C, 25°C, 30°C and 37°C in SD media (or as indicated in each experiment). The yeast strains used in this work are listed in Supplementary Table S2.

Purification of protein complexes from yeast

The yeast culture was grown at 30°C in YPD medium to OD_{600} ~2.5. Cells were harvested, frozen in liquid nitrogen and manually grinded. Lysate was melted in D100 buffer [100 mM KCl, 50 mM Tris, 10% (w/v) glycerol, 0.02% (v/v) Nonidet P–40 (NP–40), 0.2 mM EDTA, pH 7.9] with protease inhibitors, soluble fraction was recovered by centrifugation and bound to IgG-sepharose beads for 3 h at 4°C. Beads were washed thoroughly with D150 (same as D100 but 150 mM KCl) or D1000 (high-salt wash, same as D100 but 1 M KCl) buffer, and proteins were eluted by TEV protease cleavage to D150 buffer. Elutions were snap frozen and kept at −80°C. Alternatively, elutions were merged and bound o/n to Ni-NTA agarose beads. The beads were washed with D150 buffer with 10 mM imidazole and eluted by 250 mM imidazole to D50 buffer. Elutions were snap frozen and kept at −80°C.

All immunoprecipitated samples that were subsequently subjected to western blot analysis were prepared in the presence of 100 μg ml^{-1} of RNase A (added to the cell lysate). Samples that were used for enzymatic assays (polyadenylation, exosome degradation) were prepared in the absence of RNase in order to avoid interference of the residual RNase with the enzymatic analysis.

RNA isolation and analysis

For RNA analysis, cells were grown at 30°C o/n in liquid SD media, diluted to OD_{600} 0.25 and grown at 25°C to OD_{600} 1. RNA was isolated by the hot phenol extraction and stored at −80°C. Five micrograms of total RNA was denatured in 25% formamide, separated on 8% denaturing (8 M urea) polyacrylamide gel and transferred to nylon membrane using semi-dry electro-transfer. RNA was cross-linked to the membrane by UV light (120 mJ cm^{-2}) and hybridized with a probe in Ultra-Hyb buffer (Ambion) according to manufacturer's instructions.

Construction of yeast plasmids

Yeast genomic DNA or WT full-length *TRF4*, *AIR2* and *MTR4* inserted in pET22b (5) were used to PCR-amplify the untranslated and coding regions of *TRF4, AIR2* and *MTR4* genes.

TRF4 constructs. To obtain plasmids containing truncated forms of the *TRF4* gene, the desired region was amplified in a standard PCR reaction using primers listed in Supplementary Table S3 and inserted into pNOPPATA1L vector via restriction sites NdeI and SalI. To obtain point mutated forms, a standard PCR mediated site directed mutagenesis with pNOPPATA1L-*TRF4* (5) as a template was performed. All constructs contained N-terminal fusion proteinA tag and a TEV protease recognition sequence. With the exception of C-terminally truncated forms (that is *TRF4* CΔ544, CΔ499, CΔ463 and 182–463*)*, the constructs also contained fusion C-terminal hexahistidine anchor. The His_6 tag was not used during analyses and we observed no sign of its interference with the results. To express endogenous levels of the Trf4p, the genomic region spanning 500 nt upstream of the *TRF4* initiation codon was amplified using primers TRF4 506 5′-UTR XbaI For and TRF4 506 5′-UTR ATG SphI XhoI Rev and inserted into the pRS413 vector via XbaI and XhoI restriction sites. Into this construct, the *TRF4* full-length gene and/or alleles D425A, DXD and/or NΔ96 from pNOPPATTA1L constructs including all fusion tags were inserted through SphI site. For *Escherichia coli* expression, the *TRF4* coding sequence was inserted into the expression vector pET30a(+) allowing for the N-terminal fusion G-protein B1 tag (18) by using BamHI and XhoI restriction sites.

AIR2 constructs. Yeast genomic DNA or WT *AIR2* gene inserted in pET22b (5) were used to PCR-amplify the untranslated and coding regions of *AIR2*. To prepare *AIR2* constructs for phenotypic analyses *in vivo*, the coding sequence of *AIR2* surrounded by 425-bp upstream of initiation codon and 479-bp downstream of stop codon was PCR amplified from *Saccharomyces cerevisiae* genomic DNA (BY4741 background) with primers *AIR2* 5′-UTR_For1 and *AIR2* 3′-UTR_Rev1. The fragment was inserted into the pRS415 vector via PstI and NotI restriction sites resulting in the plasmid V379. To prepare tagged versions of *AIR2* for affinity purifications of TRAMP complexes, the full-length CDS of *AIR2* was subcloned to a yeast expression vector [pNOPPATA1L, (5)] via NdeI and SalI restriction enzymes, resulting in the plasmid V53. Single point mutations of *AIR2* were prepared with WT *AIR2* in pNOPPATA1L as a template by site-directed mutagenesis by using oligonucleotides listed in Supplementary Table S3. *AIR2* truncation mutants were constructed by PCR amplification with specific primers listed in Supplementary Table S2. The PCR product was inserted into pNOPPATA1L vector using NdeI and SalI restriction sites. To generate mutants pRS415 derived constructs, we first introduced an NdeI restriction site at the position 3-bp upstream of *AIR2* start codon resulting in the construct V394 WT (with primers *AIR2* NdeI ATG For and *AIR2* NdeI ATG Rev). Individual mutants prepared in pNOPPATA1L were then transferred via NdeI and BamHI sites to generate constructs listed in the Supplementary Table S4.

MTR4 constructs. Full-length *MTR4* and all deletion variants were cloned into pNOPPATA1L vector. C-terminal His_6 tag was added via PCR amplification with an extended reverse primer (MTR4 His SalI). The full-length *MTR4* was amplified from genomic DNA (primers *MTR4* NdeI and *MTR4* His SalI); *MTR4* archless was amplified from pAV674 vector (16), which was a kind gift from Dr Van Hoof. ΔKOW *MTR4* version was prepared by ligation of two PCR products into pNOPPATA1L vector (using NdeI/SalI restriction sites). N-terminal part (amplified with primers *MTR4* KOW N) of *MTR4* (aminoacids 1–665) was extended at the 3′-end by coding sequence for Gly-Ser-Gly-Ser-Ala-Ser and C-terminal part of *MTR4* (aminoacids 819–1073, obtained with oligos *MTR4* KOW C) was extended at 5′-end by coding sequence for Ala-Ser-Gly-Ser-Gly-Ser resulting in GSGSASGSGS bridge (connected via NheI restriction site in Ala-Ser coding sequence).

Preparation of unmodified $tRNA_i^{Met}$ for enzymatic and NMR binding experiments

The plasmid $ptRNA_i^{Met}$ for the *in vitro* transcription of yeast $tRNA_i^{Met}$ (5) was linearized with BstNI. Unlabeled hypomodified $tRNA_i^{Met}$ was produced by *in vitro* T7 polymerase run off transcription (19) using linearized plasmid as a template. The product was purified by anion exchange HPLC under denaturing conditions and desalted by a NAP column (Sephadex G-25). The hypomodified $tRNA_i^{Met}$ was denatured at 95°C, pH 6.2 for 5 min and snap cooled on ice, to favor a monomeric conformation. The concentration of RNA was determined by UV-spectroscopy.

Fluorescence anisotropy

The equilibrium binding of Air2p ZnK1-5 (aminoacids 57–180) to different oligonucleotides was analyzed by fluorescence anisotropy. The 15-mer poly-A RNA and DNA oligonucleotides were 5′-labeled with *N,N,N′,N′*-tetramethyl-6-carboxyrhodamine (TAMRA) that was attached via a hexyl linker. The measurements were conducted on a FluoroMax-4 spectrofluorometer (Horiba Jobin-Yvon, USA). The instrument was equipped with a thermostatted cell holder with a Neslab RTE7 water bath (Thermo Scientific, USA). The system was operated using FluorEssence software (version 2.5.3.0, Horiba Jobin-Yvon, USA). The TAMRA fluorophore was excited at 561 nm, and its emission was collected at 581 nm. The widths of both excitation and emission monochromatic slits were 8 nm and integration time was set to 3 s. During the measurement 10 nM labeled oligonucleotide (volume 1.4 ml) was titrated with increasing amounts of the protein in 50 mM Tris buffer (pH 7.9), supplemented with 250 mM NaCl, 20 mM β-mercap toethanol and 50 μM $ZnSO_4$. The titrations were carried out at 20°C in a stirred 1.5 ml quartz cuvette. A fixed delay of 30 s was set between each aliquot addition and start of the measurement to allow the reaction to reach equilibrium. This delay was sufficient, as no further change in anisotropy was observable. Every data point is an average

of three measurements. The data were analyzed in SigmaPlot 11 software (Systat Software, USA). The experimental isotherms were fit to a single-site binding model according to Heyduk and Lee (20) using nonlinear least squares regression. The data were normalized for visualization purposes.

Preparation of DNA probes for northern blot analysis

DNA probes for NEL025c and snR33 were amplified from *S. cerevisiae* BY4741 genomic DNA with the primers combinations (NEL025c For and Rev, and snR33 − mature + extension For and Rev, see Supplementary Table S5) and used as templates for Random Primed DNA Labeling Kit (Roche) according to manufacturer's instructions. To detect U14 and snR13 3′-extended snoRNAs and $tRNA_i^{Met}$, the oligonucleotides were 5′-phosphate labeled using T4 polynucleotide kinase (NEB) and [γ-^{32}P]-ATP.

Whole-cell protein extract preparation for western blot analysis

The yeast culture was grown at 30°C in selective medium to OD_{600} ~1.0. The cells were resuspended in the D150 buffer with protease inhibitors (0.5 μg ml^{-1} of leupeptin, 0.8 μg ml^{-1} of pepstatin A, 0.6 mM PMSF and 0.6 mM DTT). The cell suspension was vortexed with glass beads at 4°C for 10 min. The beads and the cell debris were pelleted by centrifugation, a part of the cell-free extract was mixed with 1×SDS loading buffer for western blot analysis. The rest of the extract was used for protein concentration measurement by the method of Bradford.

Antibodies

The generation of specific antibodies against Trf4p, Air2p, Air1p have been described in (5,7). Mtr4p antibodies were kindly provided by Dr Patrick Linder (University of Geneva). The anti-proteinA antibodies and the anti-His_6 antibodies were from Sigma. All antibodies were used at a 1:5000 dilution.

Expression and purification of recombinant proteins

Full-length Trf4p from *S. cerevisiae* was recombinantly expressed from pET30a vector with N-terminally fused G-proteinB1 solubility tag (18) in *E. coli* strain BL21-CodonPlus (DE3)-RIPL (Stratagene). The expression was induced by 0.5 mM IPTG at 37°C for 2 h in the presence of 1 mM $MgCl_2$. Cells were lysed in Lysis Buffer (50 mM carbonate buffer, 300 mM NaCl, 5% glycerol, 5 mM imidazole, 0.1% NP-40, 1 mM $MgCl_2$, 2 mM β-mercaptoethanol and protease inhibitors, pH 10.5) by sonication, and protein was purified in 50 mM Tris buffer, pH 8.0 containing 300 mM NaCl, 5% glycerol, 0.02% NP-40, 1 mM $MgCl_2$, 1 mM β-mercaptoethanol using affinity Ni-NTA chromatography. For NMR studies, the GB1-Trf4 fusion protein was dialyzed into a buffer containing 50 mM Tris, pH 8.0, 300 mM NaCl, 3% glycerol and 1 mM $MgCl_2$ and subsequently concentrated up to 230 μM by VIVASPIN column (Sartorius).

Wild-type and mutant forms of *AIR2* were recombinantly expressed from the pET22b vector in *E. coli* strain BL21-CodonPlus (DE3)-RIPL (Stratagene). The expression was induced by 0.4 mM IPTG at 37°C for 4 h in the presence of 0.5 mM $ZnSO_4$. Cells were lysed in the lysis buffer (50 mM carbonate buffer, 400 mM NaCl, 25% sacharose, 2 mM $MgCl_2$ 10% glycerol, 400 μM $ZnSO_4$, pH 10.5), and all proteins were purified in buffer containing 50 mM carbonate buffer, pH 10.5, 400 mM NaCl, 10% glycerol, 400 μM $ZnSO_4$ using affinity Ni-NTA chromatography followed by gel filtration.

Expression and preparation of Air2 ZnK protein for NMR studies

The coding sequence of Air2 protein comprising residues 57–180 from *S. cerevisiae* was expressed with the N-terminal fusion Smt3 protein (21). The expression was induced by 100 μM IPTG in *E. coli* BL21-Codon Plus (DE3)-RIPL (Stratagene) strain overnight at 16°C in M9 minimal medium, supplemented with 50 μM $ZnSO_4$. For isotope labeling, the medium was supplemented with $^{15}NH_4Cl$ and [U-$^{13}C_6$] glucose. Cells were harvested, resuspended in denaturing buffer (50 mM Tris, 500 mM NaCl, 8 M Urea, pH 8.0) and lysed by sonication, and proteins were purified using affinity Ni-NTA chromatography. The protein immobilized on beads was extensively washed and subsequently refolded in buffer containing 50 mM Tris pH 7.9, 500 mM NaCl, 20 mM β-mercaptoethanol, 5 mM imidazole, 50 μM $ZnSO_4$, and protease inhibitors. Refolded protein was eluted and dialyzed against buffer containing 50 mM Tris pH 7.9, 500 mM NaCl, 20 mM β-mercap toethanol and 50 μM $ZnSO_4$. The Smt3 fusion protein was cleaved by Ulp1p protease and removed on Ni-NTA column. For NMR measurements the protein was concentrated using a Vivaspin 20 concentrator (Sartorius).

NMR spectroscopy and structure determination

The protein used in NMR spectroscopy was concentrated to ~150 μM. Due to a high salt buffer [50 mM sodium phosphate buffer, 500 mM NaCl, 10 mM β-mercaptoethanol pH 8.0 (90% H_2O/10% D_2O)], 3 mm or shaped NMR tubes were used. All NMR experiments were recorded at 20°C on a Bruker AVANCE 600 MHz and 900 MHz spectrometers equipped with a cryogenic triple resonance probe (Bruker BioSpin). The chemical shifts of backbone and side-chains were assigned using standard triple resonance experiments (22,23). All distance restraints were derived from the 3D ^{15}N- and ^{13}C-edited NOESYs (with mixing time of 80 ms) collected at 900 MHz spectrometer. The structure calculations were performed with the automated NOE assignment module implemented in the CYANA program (24). Initial structures were further refined with the CYANA-generated restraints along with the restraints for the tetrahedral zinc coordination with AMBER 10.0 software (25) as described previously (26,27). From 40 refined structures, the 20 conformers with the lowest AMBER energy were selected to form the final ensemble of structures.

Enzymatic assays

Polyadenylation assays were carried out in 15 µl reaction mixtures containing 5–50 ng of affinity purified complex, 50 fmol 5′-end labeled RNA, 0.5 mM ATP, 5 mM $MgCl_2$, 25 mM Tris, pH 7.9, 20 mM KCl, 10% glycerol, 0.1 mg ml^{-1} BSA, 1 mM DTT, 0.02% Nonidet P-40, 5U of RNasin Plus (Promega) with or without addition of 50 ng of affinity purified Rrp6-TAP eluate (5). Reaction mixture was incubated at 30°C, and 5 µl aliquotes were collected at times indicated. The reaction was stopped by the addition of 25 mM EDTA and proteins removed by incubation with ProteinaseK (Promega) for 10 min at 37°C. Sample was mixed with one volume of formamide loading buffer and resolved on denaturing polyacrylamide gel.

Database searches and sequence analysis

BLAST searches were conducted by using the NCBI genome databases (http://www.ncbi.nlm.nih.gov/BLAST/) or other nonredundant or EST databases at the Swiss EMBnet node (28). The multiple sequence alignment was performed by using ClustalW2 (29) at the http://www.ebi.ac.uk/Tools/services/web_clustalw2 interface. The Boxshade server (http://www.ch.embnet.org/software/BOX_form.html) was used to color the similarity in the alignment.

RESULTS

Identification of Air2p regions required for intermolecular contacts with Trf4p and Mtr4p

Trf4/5p, unlike canonical PAPs, possess no identifiable RNA-binding domain and thus are inactive on their own. We have previously demonstrated that Air2p activates Trf4p, likely through its RNA-binding properties (5). To identify Air2p regions that contact RNA, interact with Trf4p or Mtr4p, and are required for yeast viability, we have prepared Air2p mutants carrying point mutations in the second zinc-coordinating cysteine in each of the five ZnKs (ZnK1 to ZnK5) to disrupt the ZnK fold (Figure 1A and Supplementary Figure S1A). In order to tackle the importance of the surrounding N- and C-termini, we generated deletion mutants lacking either the N- or C- terminus or both (Figure 1A). Mutants ZnK4, ZnK5, CΔ180 and ZnK1-5L were not able to rescue growth defect of the *air1Δair2Δ* strain at higher temperature and mutants in the first three ZnKs and NΔ57 showed reduced fitness (Figure 1B). The growth defects observed at 30°C were not caused by reduced expression nor stability of Air2p variants nor other TRAMP subunits, as the western blot analysis of protein extracts prepared from cells grown at 30°C revealed comparable protein levels for the mutants tested (Supplementary Figure S1B and S1C). This pointed to the importance of the last two zinc knuckles and the Air2p C-terminus. Our previous yeast two-hybrid results indicated the minimal region in Air1 and Air2 proteins required for Trf4p interactions positioned across the zinc knuckles ZnK3 to ZnK5 (5). Accordingly, in NMR titration experiments, we observed that the fourth and the fifth ZnKs of Air2p associated with recombinant Trf4p (data not shown). Furthermore, we observed chemical shift perturbation in the third (L3) and fourth linker (L4) regions upon Trf4p binding. It concerned mainly residues 117–120 (L3) and residues S138, I139, W140, R141 and Y143 (L4), most of them being highly conserved between Air homologs from distant species (Supplementary Figure S1A). The analysis of additional mutants in those regions revealed that double mutants in the last two ZnKs ZnK4+5 and in the fourth linker (W140G+R141E), respectively, exhibited poor fitness comparable to the WT strain at all temperatures tested (Figure 1B). The growth defects observed at 24°C correlated with deficiencies in TRAMP function *in vivo* as northern blot analyses revealed a strong accumulation of NEL025c CUT and snR33 pre-snoRNA in ZnK4+5, L4, CΔ180 and ZnK1-5L mutants cultivated at 24°C (Figure 1C). Cold sensitivity and to a certain extent RNA degradation defects seen at 24°C were observed also in all the other ZnK point mutants indicating that all ZnKs contribute to some degree to TRAMP activity *in vivo* (Figure 1B and C).

To address whether the *in vivo* phenotypes reflected defects in TRAMP assembly or catalytic activity, we isolated and analyzed yeast complexes containing the individual mutated Air2p versions. In agreement with our NMR studies, L4 and ZnK5 mutants failed to efficiently copurify Trf4p and Mtr4p, whereas the other point mutations did not affect TRAMP4 assembly (Figure 1D). Western blot analysis of the input fractions of individual Air2p mutants revealed comparable levels of all three TRAMP subunits in WT, ZnK and L4 mutant strains. Thus the lack of Trf4p and Mtr4p copurification with L4 and ZnK5 was not caused by their decreased stability or expression in these mutant strains. As the absence of Mtr4p does not disrupt the interaction with Trf4p (Figure 1D, HSW) (4,5), we conclude that the fourth linker and last zinc knuckle regions of Air2p are crucial for the binding to Trf4p. The lack of Trf4p and Mtr4p copurification in ZnK5 samples was reflected in the lack of PAP activity of the purified protein sample (Figure 2A). Interestingly, although the ZnK4 mutant was able to maintain the TRAMP integrity (Figure 1D), it was strongly defective in polyadenylation *in vitro* (Figure 2A). Accordingly, recombinant Air2 proteins mutated in ZnK4 or ZnK5 failed to activate yeast Trf4p *in vitro* (Figure 2B).

We have previously reported that purified TRAMP4 complex activates exosomes *in vitro* (5). In order to test whether some of the mutants may affect exosome activation, we performed coupled polyadenylation/exosome assays. In this assay, the 5′-end labeled hypomodified $tRNA_i^{Met}$ (a well-defined *in vivo* target of this pathway) is incubated in the presence of WT or mutant TRAMPs and yeast nuclear exosome purified via the Rrp6p subunit (5). Figure 2C shows Rrp6p-TAP purified fractions that were used in all subsequent degradations assays in this work compared to exosome obtained via the core subunit Rrp4p. The coupled polyadenylation/degradation assays revealed that none of the polyadenylation-competent Air2p point mutants further affected exosome activation *in vitro* (Figure 2D).

The analysis of deletion mutants indicated the importance of the first 25 N-terminal aminoacids for the

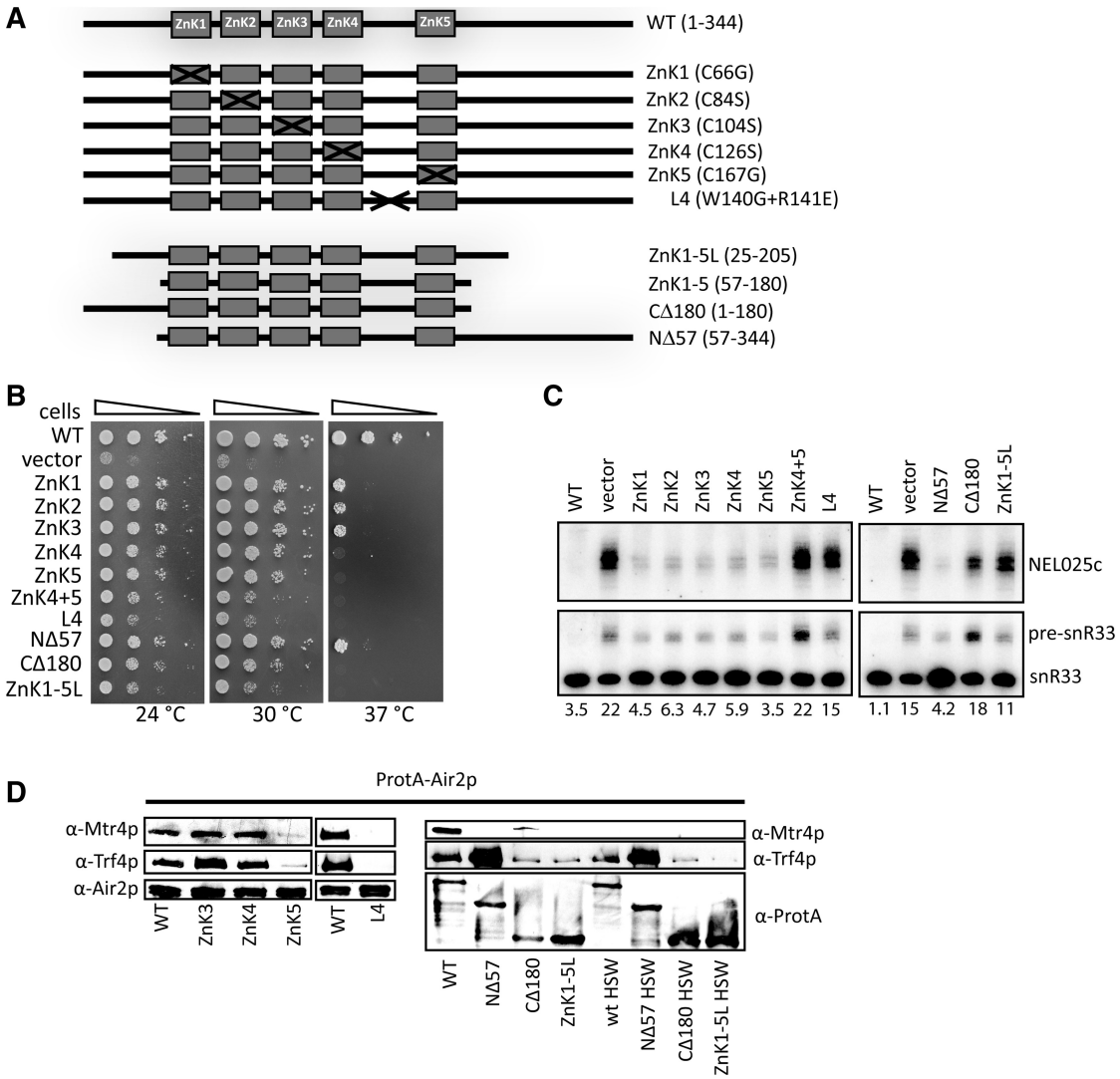


Figure 1. Air2p N- and C-termini and ZnK4 and ZnK5 domains are important for TRAMP functions and integrity. (**A**) Overview of mutant forms of *AIR2* used in this study: ZnK, zinc knuckle. The aminoacid substitutions of individual point mutants are designated in brackets. (**B**) Growth test analysis of WT and mutant forms of Air2p episomally expressed in *air1Δair2Δ* strain, serially diluted on SD-Leu media and incubated for 3 days at temperatures indicated. (**C**) Northern blot analysis of total RNA isolated from *AIR2* WT and mutant yeast strains grown at 25°C, with specific probes for RNAs indicated. The number below each lane represents percentage of unprocessed snoRNA precursor. (**D**) Western blot analysis of the composition of purified TRAMPs containing WT and mutant Air2p variants. TRAMP subunits were detected with antibodies indicated: HSW, high-salt wash (1M KCl).

interaction with Mtr4p. We observed that both the NΔ57 (residues 57-344) and ZnK1-5L (residues 26-205) proteins did not co-purify Mtr4p (Figure 1D, lanes NΔ57 and ZnK1-5L). Interestingly, the NΔ57 protein on contrary pulled down significantly higher amounts of Trf4p than WT even in the presence of 1M KCl (HSW). At present, we do not know the reason for higher yields of Trf4p in this mutant. These results were not due to significantly altered levels of Mtr4p nor Trf4p in these mutants (Supplementary Figure S1C). In addition, the deletion of the last 139 amino acids weakened the interaction with both Mtr4p and Trf4p (Figure 1D, samples CΔ180 and ZnK1-5L).

Air2p is a genuine RNA-binding protein

To further investigate why TRAMP4 mutant in Air2p ZnK4 lacks PAP activity, we studied the central region of Air2p (ZnK1-5, residues 57-180) by NMR spectroscopy. The Air2p was purified under denaturing conditions and refolded on Ni-NTA column in the presence of N-terminal Smt3 fusion protein. We first confirmed, that this fragment is sufficient to activate the PAP activity of Trf4p (Figure 2B, first lane). For NMR experiments, a high-salt concentration of 500 mM NaCl was used to prevent the protein aggregation. Although this protein construct suffered from a very low solubility limit (150 μM), we were able to achieve ^{1}H, ^{13}C and ^{15}N resonance assignments of ZnK1-5 and determine its three-dimensional structure (Figure 3A). The chemical shift deviations of Cα and Hα of the assigned backbone resonances of ZnK1-5 from the sequence-dependent random coil values show no pattern for the ZnK1-5 region. This agrees well with the absence of secondary structure elements in the

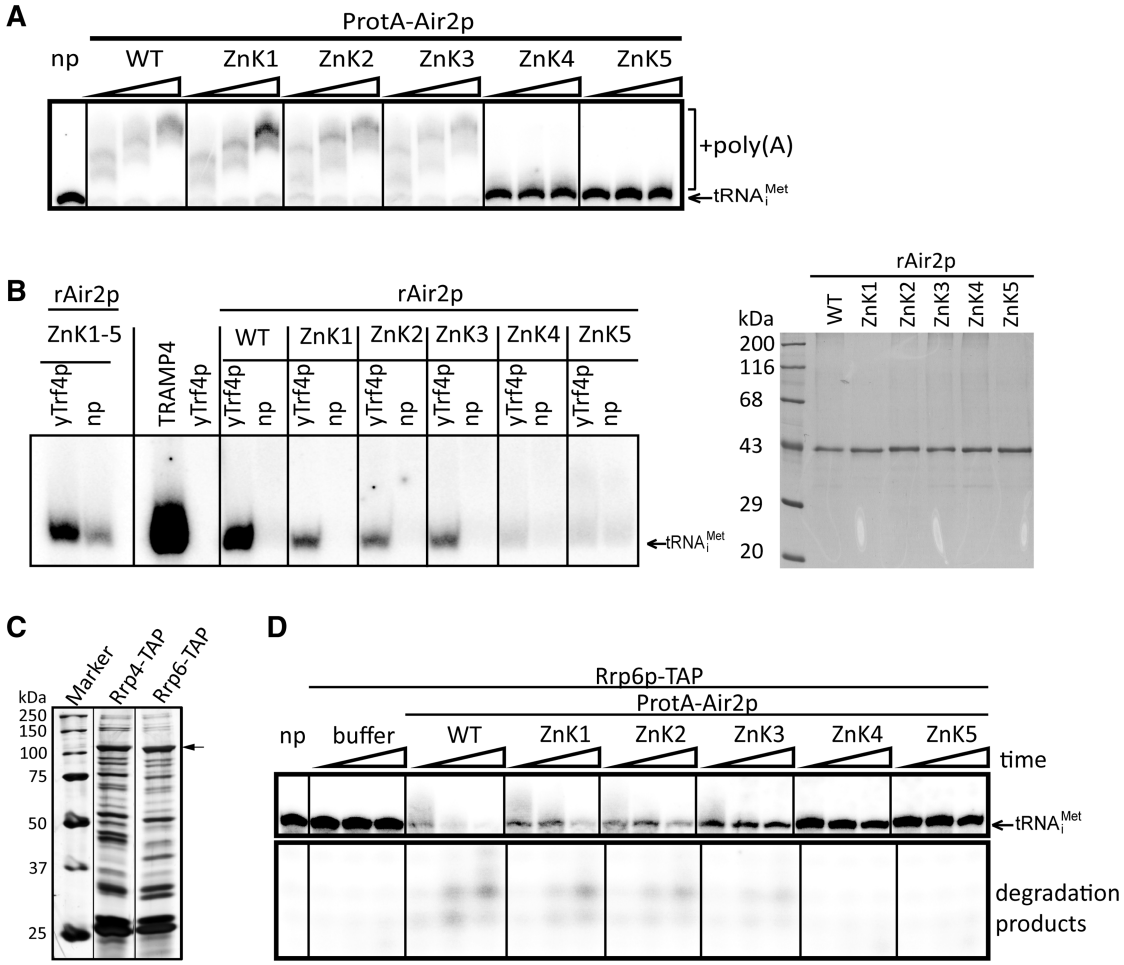


Figure 2. Air2p ZnK4 and ZnK5 are essential for TRAMP polyadenylation activity *in vitro*. (**A**) *In vitro* polyadenylation assay with affinity purified WT and mutant Air2p TRAMPs was performed in the presence of 0.5 mM ATP and 50 fmol of radioactively labeled hypomodified tRNA$_i^{Met}$ (migration indicated by an arrow). Reactions were stopped after 15, 30 and 60 min, respectively. (**B**) Reconstitution of Trf4p PAP activity with recombinant Air2p. *In vitro* polyadenylation assay was performed with 60 ng of WT and mutant recombinant Air2p and 20 ng of WT Trf4p affinity purified from *air1Δair2Δ* yeast strain in the presence of [α-^{32}P]-ATP and 50 fmol unlabeled hypomodified tRNA$_i^{Met}$ (migration indicated by an arrow). The presence of a radioactive signal in a lane is the result of polyadenylation activity. Reactions where Trf4p was omitted (np) serves as a negative control for bacterial PAP contamination. Trf4p purified from *air1Δair2Δ* strain has no detectable activity alone (lane 4, yTrf4p). Trf4p purified from WT yeast strain (TRAMP4) was used as a positive control (lane 3, TRAMP4). Reactions were stopped after 60 min in the case of recombinant proteins or 10 min in the case of TRAMP4. The purity of recombinant Air2 proteins used in the assay separated on 12% polyacrylamide gel and stained with Coomassie blue are shown on the right. (**C**) Rrp6p-TAP purified exosome fraction used in all degradation assays of this work is compared to profile of an exosome purified via the core subunit Rrp4p. The purified complexes were resolved on 10% SDS-PAGE and silverstained. The migration position of Rrp6p is marked by an arrow. M is molecular weight marker. (**D**) ZnK1, 2 and 3 are dispensable for exosome activation. *In vitro* coupled polyadenylation/exosome assays with 30 ng TRAMP4 obtained via affinity purification of WT and mutant Air2p. The reactions were performed in the presence of 0.5 mM ATP and 50 fmol of radioactively labeled hypomodified tRNA$_i^{Met}$ (migration indicated by an arrow). Reactions were stopped after 15, 30 and 60 min, respectively. RNAs were resolved on 20% PAGE.

topology of zinc knuckles (13,22). The histidines of the CC**H**C motif of ZnKs, namely H71, H90, H109, H131 and H172, are involved in coordination of zinc ions as the N-H correlations of these residues span the same region in the 2D ^{1}H-^{15}N HSQC (Figure 3B) that is characteristic for zinc knuckles (22). Furthermore, the presence of folded zinc knuckles was confirmed by the addition of EDTA in excess of zinc ions that resulted in a loss of resonance dispersion in the 2D ^{1}H-^{15}N HSQC (data not shown). The NOESY spectra suffered from a low signal-to-noise ratio as a result of low solubility of the ZnK1-5 protein and the presence of high-salt buffer. Therefore, the resulting NMR structure is of a low resolution (Supplementary Table S1, Supplementary Figure S2). The structure of this Air2p region consists of five independent zinc knuckles that all adopt a canonical zinc knuckle fold of the CCHC type. The domains are connected by flexible linkers, and no inter-domain contacts have been found (Figure 3A). The conserved IWRxY motif of linker 4 is also unstructured in the free form, whereas it forms a short α-helix upon binding to Trf4p (15).

To investigate whether Air2p is a genuine RNA-binding protein, we performed a quantitative solution-binding

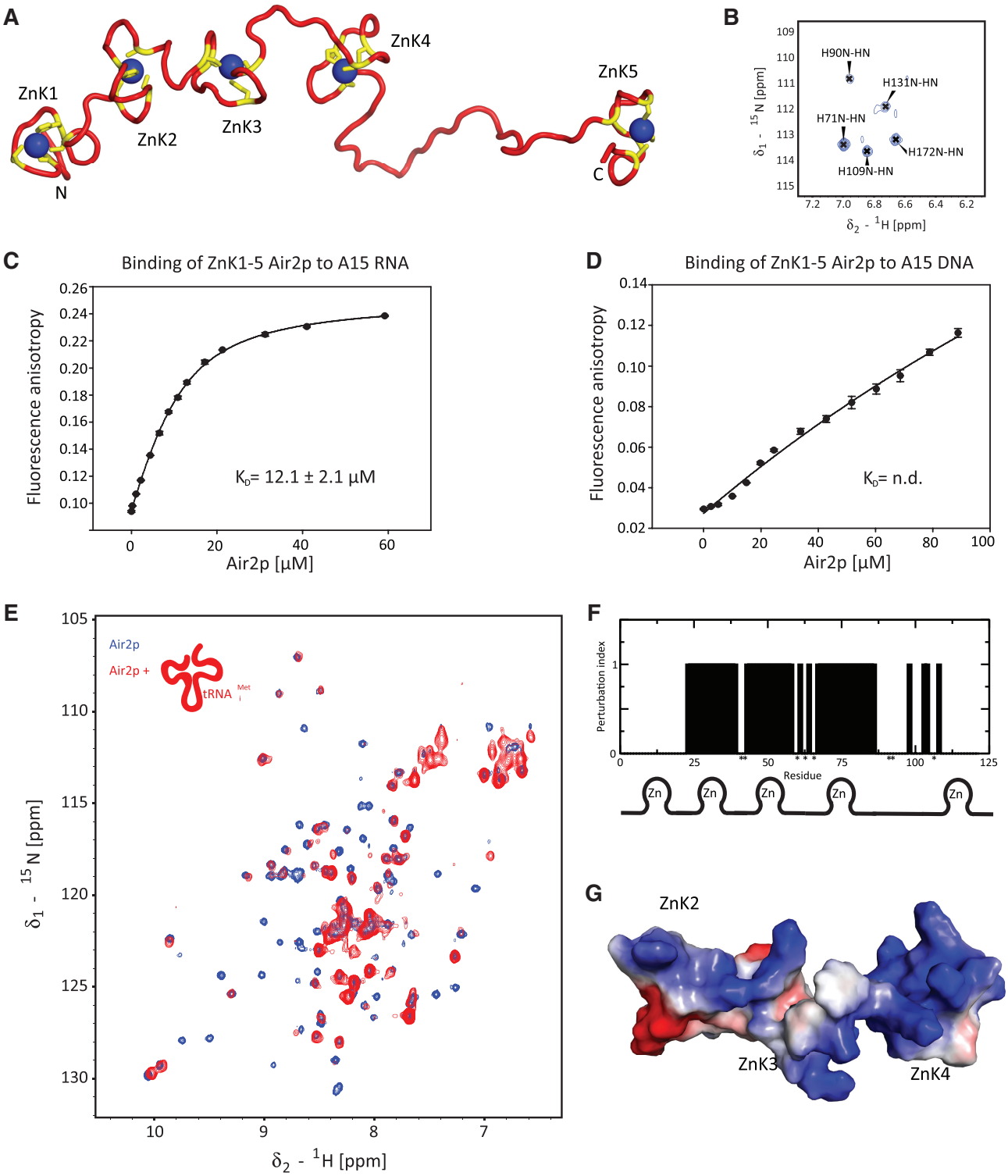


Figure 3. Structure and RNA-binding of ZnK1-5 Air2p. (**A**) The lowest energy structure of the Air2p ZnK1-5. The protein is shown as a ribbon model (in red) with the zinc-coordinating residues (CCHC) shown in yellow. The zinc ions are shown in blue. (**B**) Close-up view of the ^{1}H-^{15}N TROSY spectrum, showing the N-H correlations of histidines that are involved in the coordination of zinc ions. (**C**) Air2p ZnK1-5 binds A15 RNA. Binding isotherms for equilibrium binding of ZnK1-5 Air2p to fluorescently labeled A_{15} RNA, monitored by fluorescence anisotropy. (**D**) Air2p ZnK1-5 binding to dA_{15} DNA. The ionic strength and pH of the binding buffer was the same for both measurements. The dissociation constant (K_d) was calculated from the best fit of data using a single-site binding isotherm. (**E**) ^{1}H-^{15}N TROSY spectra of Air2p ZnK1-5 alone (in blue) and in the presence of one equivalent of hypomodified $tRNA_i^{Met}$ (in red; ^{1}H-^{15}N HSQC) at 20°C. (**F**) Summary of chemical shift perturbations and line broadening of Air2p ZnK1-5 upon binding to hypomodified $tRNA_i^{Met}$. Affected residues (qualitatively described by perturbation index; yes = 1, no = 0) are plotted against the amino-acid residue number. The assignments of residues indicated by asterisks could not be obtained. (**G**) Solvent-accessible surface representation of the RNA-binding ZnK of Air2p (ZnK2, ZnK3 and ZnK4) colored by electrostatic potential (blue, positive; red, negative) of the representative structure.

assay using fluorescence anisotropy (FA) experiments. We found that Air2p ZnK1-5 binds unstructured A_{15} RNA with a K_D in a low micromolar range (Figure 3C), whereas the binding affinity to dA_{15} DNA could not be detected in our experimental setup ($K_D >> 100\mu M$) (Figure 3D). Next, we used NMR spectroscopy to reveal which ZnK domains of Air2p are used to bind to hypomodified $tRNA_i^{Met}$, a substrate that is used for *in vitro* assays in this work. The chemical shift changes and resonance broadening observed in the ^1H-^{15}N HSQC spectrum of Air2p ZnK1-5 upon the addition of $tRNA_i^{Met}$ established that $tRNA_i^{Met}$ interacts with the second, the third, and the fourth ZnKs (Figure 3E and F). In addition, the chemical shifts of the linkers connecting these ZnKs were perturbed as well. This suggested that these highly conserved linkers that are unstructured in the free form, become structured upon RNA-binding akin to the linkers of viral nucleocapsid (NC) proteins (13). Our attempts to compare binding of Air2p to hypomodified $tRNA_i^{Met}$ and native $tRNA_i^{Met}$ (the latter being a poor substrate for TRAMPs) failed, as we were not able to prepare sufficient quantities of native $tRNA_i^{Met}$ from yeast for NMR experiments.

The minimal functional Trf4p

The recent crystal structure of truncated Trf4p bound to a fragment of Air2p encompasses only the conserved catalytic (CAT) and central domains (CD) of Trf4p (residues 161–481) (15). We observed that these regions alone (fragments 182–463 or 161–481) are not functional *in vivo* as they were not able to rescue viability of the double deletion strain *trf4Δtrf5Δ* (Figure 4A and C and Supplementary Figure S3A and S3B). To search for the minimal functional Trf4p and to address the role of individual Trf4p domains, we prepared a set of deletion and point mutant constructs of *TRF4* (Figure 4B). The deletions were designed to remove poorly conserved, potentially unstructured parts of the protein. In addition to the catalytically inactive DXD mutant (D236A and D238A) (5), we introduced two point mutations: (i) D425A residing in a highly conserved aspartate residue in the central domain (30), which is not present in canonical PAPs, and (ii) K552A, which has been proposed to reside in the catalytic center for Trf4p dRP lyase activity (31).

Deletions of the entire N- (NΔ181) or C- (CΔ463) termini surrounding the CAT and CD parts of Trf4p did not support viability of the double *trf4Δtrf5Δ* strain (Figure 4C, Supplementary Figure S3A and S3B) and exhibited an accumulation of three well-characterized TRAMP4 substrates, NEL025c CUT and precursors of SnR13 and U14 snoRNAs when expressed in trf4Δ strain (Figure 4D). Although we were able to detect the expression of these fragments by western blot (Supplementary Figure S3C), our attempts to purify these mutants to assess the interaction with other TRAMP components and their activities failed. Thus, we hypothesize that the deletion resulted in protein misfolding and instability. The analyses of the rest of the deletion mutants indicated the minimal functional Trf4p region (spanning residues 97–499) containing CAT and CD of Trf4p surrounded by short extensions as NΔ96 and CΔ499 were sufficient for TRAMP4 assembly (Supplementary Figure S3D). *In vivo*, CΔ499 displayed cold sensitivity and impaired growth at 30°C when expressed in *trf4Δtrf5Δ* deletion background (Figure 4C, Supplementary Figure S3A and S3B). When expressed in the single *trf4Δ* strain, CΔ499 along with NΔ96 were able to suppress the RNA degradation/processing defects of the *trf4Δ in vivo* (Figure 4D). TRAMPs purified via proteinA-tagged NΔ96 and CΔ499 showed WT-like PAP activities (Figure 4E) and were able to activate $tRNA_i^{Met}$ degradation in the presence of Rrp6p-TAP exosomes *in vitro* (Figure 4F).

In agreement with previously published data, the catalytically inactive DXD TRF4 mutant suppressed only partially the NEL025C accumulation (6,7). This implies that TRAMP can also act in polyadenylation independent manner *in vivo*. In addition to DXD, we have identified D425A as the second point mutant with impaired PAP activity (Figure 4F). Compared to DXD, the D425A exhibited very weak, only above the background accumulation of TRAMP targets seen by northern blot analysis (Figure 4D). Moreover, it was able to support growth of *trf4Δtrf5Δ* strain (Figure 4C). The purified TRAMP4 containing the point mutation in D425 failed to polyadenylate hypomodified $tRNA_i^{Met}$ *in vitro* under standard assay conditions and accordingly to activate the exosome (Figure 4E and F), although the complex integrity was not affected (Supplementary Figure S3D). We found that increasing ATP concentration resulted in a gradual increase of the PAP activity in the D425A mutant up to the 4 mM concentration tested, whereas it did not change any further activity of the saturated WT protein (Figure 4G). This result indicated that the mutation reduced Trf4p affinity for ATP. We propose that lowered ATP substrate affinity of D425A slows down PAP activity of the complex, which is however sufficient for TRAMP4 function *in vivo*.

The Mtr4p-arch domain is required for TRAMP4 complex-mediated exosome activation *in vitro*

We have observed that in the absence of Air proteins the stability of the Trf4p-Mtr4p heterodimer is compromised even under low salt conditions (Figure 5A) implying that Air proteins mediate interaction between Trf4p and Mtr4p. This is in agreement with similar recent findings in *Schizosaccharomyces pombe* (32), suggesting a conserved mode of Mtr4p interaction within TRAMP in two distant yeast species. Based on sequence conservation and molecular surface scanning, several patches were proposed as possible binders for Trf4p-Air2p (17). To identify the interaction surface of Mtr4p with the rest of the TRAMP complex, we prepared a set of terminal and internal deletion variants N-terminally fused with cleavable 2×proteinA tag (Figure 5B). As expected, the mutants in the conserved core were not able to support growth in the absence of WT Mtr4p, while the removal of the first 143 amino acids (NΔ143) resulted in strong growth defect (Figure 5C). The archless mutant shows reduced growth as reported previously (16).

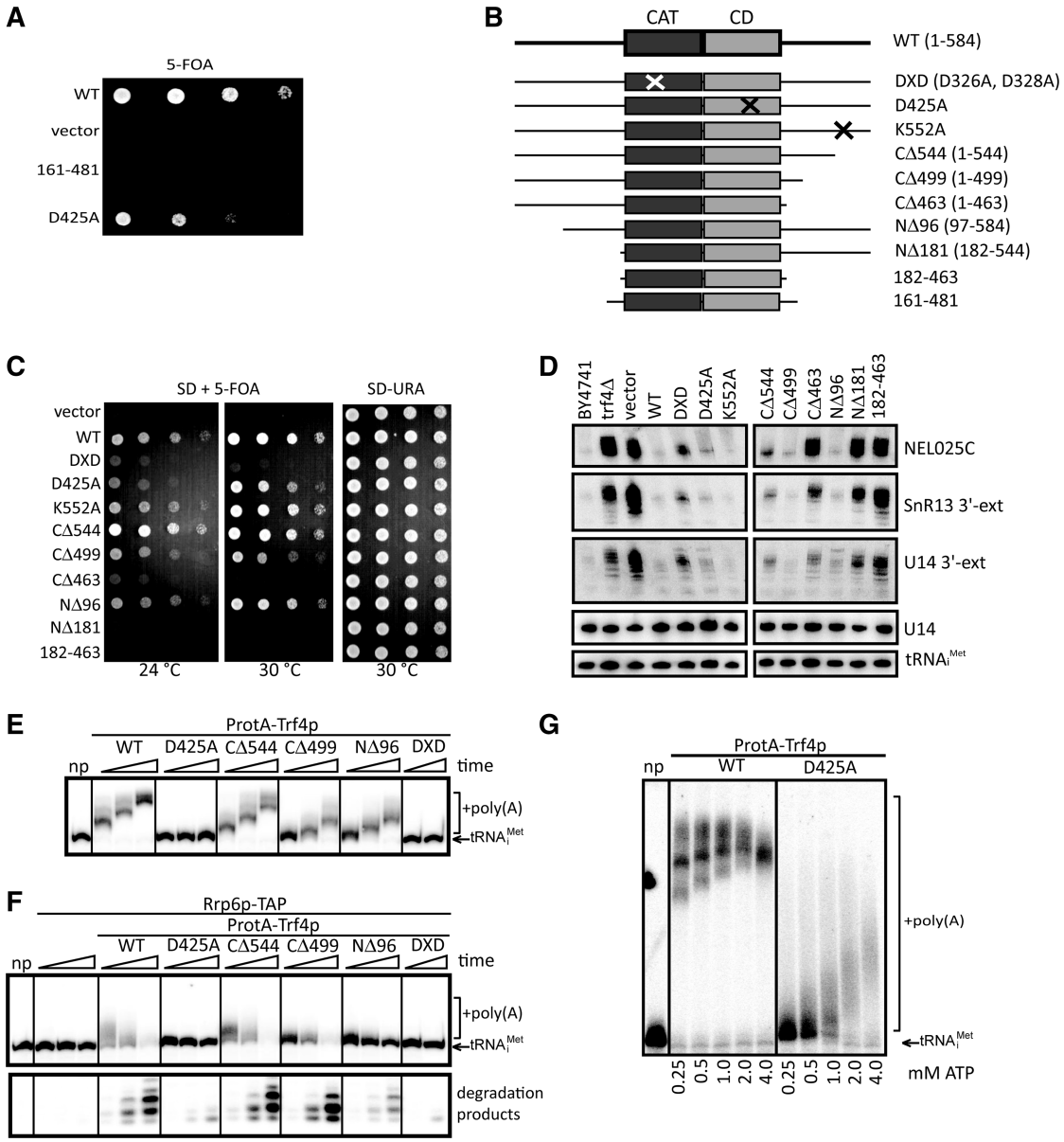


Figure 4. Mutational analysis of Trf4p reveals minimal functional Trf4p and indicates the importance of D425 in PAP activity. (**A**) Fragment of Trf4p (aminoacids 161–481) used for previous structural studies does not support growth in *trf4Δtrf5Δ* background. Growth test analysis of wt and mutants of *TRF4* episomally expressed in *trf4Δtrf5Δ* strain complemented with WT *TRF4* on *URA3* plasmid and serially diluted and spotted on media with 5-FOA. (**B**) Overview of mutant forms of *TRF4* used in this study: CAT, catalytical domain; CD, central domain. (**C**) Minimal viable allele of *TRF4* spans the amino acids 96–499. Drop tests were performed analogously to (A). (**D**) Northern blot analysis of total RNA isolated from WT and mutant *TRF4* yeast strains with DNA probes specific for indicated RNAs. (**E**) *In vitro* polyadenylation assays with 20 ng affinity purified WT and mutant TRAMP4 were performed in the presence of 0.5 mM ATP, 10 ng radioactively labeled hypomodified $tRNA_i^{Met}$ (migration indicated by an arrow). Reactions were stopped after 15, 30 and 60 min, respectively. The first lane (np) represents control sample with no protein added. RNAs were separated on 20% PAGE. (**F**) Coupled polyadenylation and exosome assay. The 5′-end-labeled unmodified $tRNA_i^{Met}$ incubated with 40 ng of affinity-purified TRAMP4 (ProtA-Trf4p) and 40 ng of nuclear exosome (Rrp6-TAP) for 15, 30 and 60 min. The first lane (np) represents control sample with no protein added. (**G**) D425A mutation lowers Trf4p affinity to ATP. Polyadenylation assay was performed with 25 ng of affinity purified D425A Trf4p, 50 fmol 5′-end-labeled unmodified $tRNA_i^{Met}$ in the presence of ATP concentrations indicated. Reactions were stopped after 60 min and resolved on 10% denaturing gel.

Interestingly, the deletion of the KOW domain was sufficient to reproduce this phenotype (Figure 5D).

Using affinity purification from *S. cerevisiae* and subsequent western blotting, we confirmed the previous results from *in vitro* (17) showing that the arch domain is also dispensable for TRAMP formation *in vivo* (Figure 5E). Furthermore, we observed the full complex formation after deletion of the unusual RecA-1N-terminal β-hairpin extension (NΔ143) found in the structure (17), which is not conserved outside of the Ski2 protein family

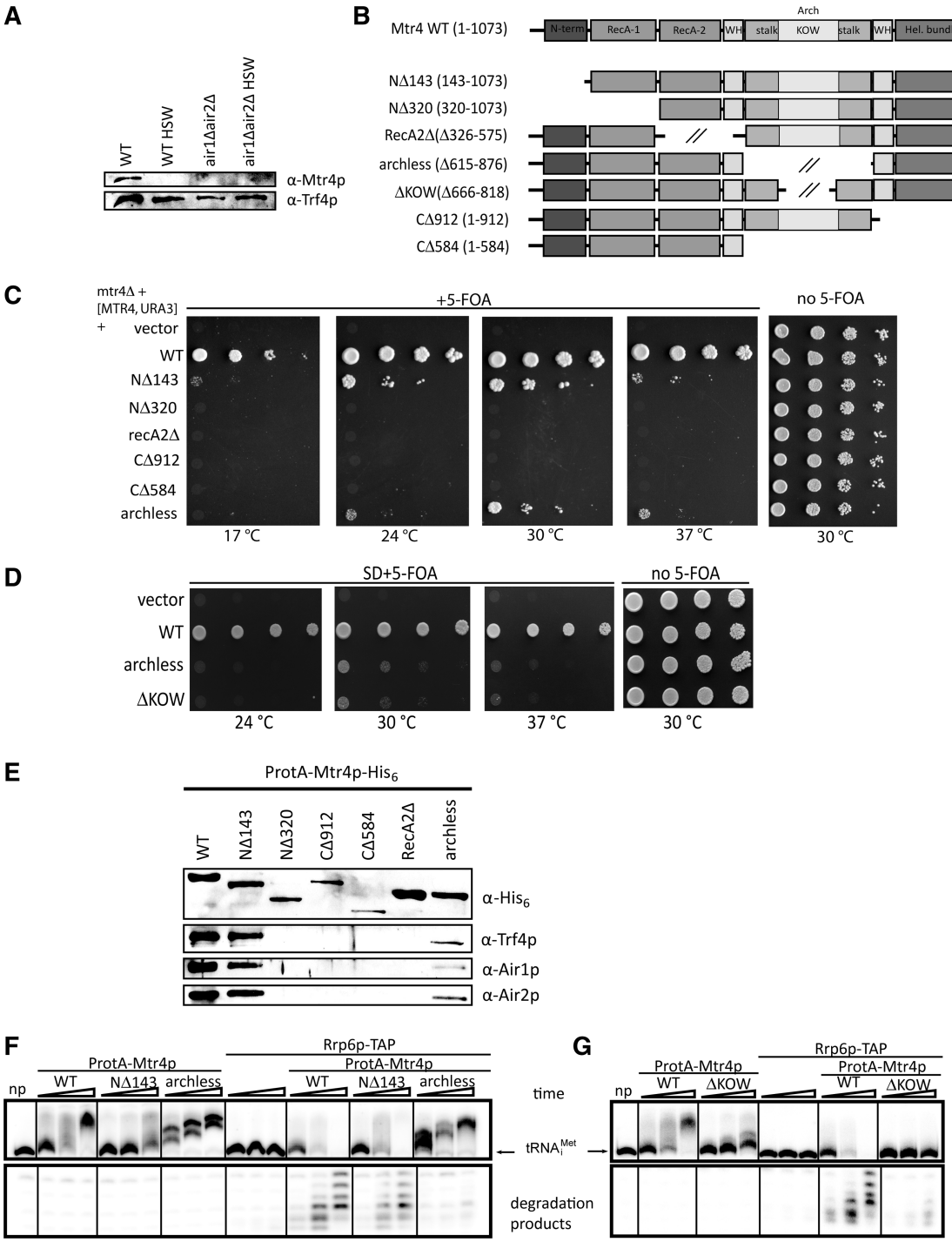


Figure 5. The RNA-binding KOW domain of Mtr4p is required for activation of exosome by the TRAMP complex. (**A**) Air2p is required for Mtr4p association with Trf4p. Western blot analysis of affinity purified ProteinA-Trf4p from WT or *air1Δair2Δ* strains. TRAMP subunits were detected with antibodies indicated. HSW, 1 M KCl high-salt wash used during purification. (**B**) Overview of mutant forms of *MTR4* used in this study: N-term, N-terminal β-hairpin extension, WH, winged-helix domain. (**C** and **D**) Growth test analysis of WT and mutants of *MTR4* episomally expressed in *mtr4Δ* strain complemented with WT *MTR4* on *URA3* plasmid and serially diluted and spotted on media with 5-FOA. (**E**) Western blot analysis of the composition of purified TRAMPs containing different Mtr4p variants. TRAMP subunits were detected with antibodies indicated. (**F** and **G**) Coupled *in vitro* polyadenylation and exosome activation assay with 40 ng affinity purified Mtr4p variants were performed with 0.5 mM ATP and 50 fmol radioactively labeled hypomodified $tRNA_i^{Met}$ (migration position indicated by an arrow). Reactions were stopped after 15, 30 and 60 min, respectively. Samples np stand for controls with no protein included. The entire gel is shown in Supplementary Figure S4A.

(Figure 5E). Deletion of any of the domains that form the helicase core, led to the loss of co-purification of TRAMP components (Figure 5E). Nevertheless, we cannot rule out the possibility of misfolding or cellular mislocalization. Since the base area of the helical core around the RNA exit tunnel was proposed to serve for exosome interaction

(16), we narrowed down the interaction site with the Trf4p/Air2p heterodimer to the side of the barrel formed by RecA-2, winged helix and helical bundle domains.

Since Mtr4p is dispensable for TRAMP polyadenylation activity but is required for the activation of the exosome (4,5), we tested whether the loss of N-terminal extension or arch domains had any effect on the ability of the TRAMP complex to activate degradation in coupled *in vitro* polyadenylation and exosome assay. While the NΔ143 mutant did not show any deviation from WT, we observed almost complete loss of exosome activation in the presence of the archless mutant on two different structured substrates (Figure 5F, Supplementary Figure S4A and S4B). To investigate the importance of the KOW domain, the RNA binding element of the arch (17), we conducted the assay in the presence of Mtr4p variant lacking this domain. As shown in Figure 5G, removal of the KOW domain recapitulated the inability of the archless mutant to efficiently activate exosome *in vitro*. Since we observed apparently lowered polyadenylation in KOWless TRAMPs, we next compared the exosome activation in the presence of lowered amounts of WT Mtr4p. Despite the resulting reduction in observed polyadenylation activity, the WT Mtr4p TRAMP still activated Rrp6p-TAP, unlike KOWless mutant (Supplementary Figure S4C). This together with the slow growth phenotypes (Figure 5D) suggested that arch and particularly the KOW domains are crucial for exosome activation.

DISCUSSION

We have identified several key features of the assembly, RNA binding and exosome activation of the yeast TRAMP RNA surveillance complex. While we focused on the analysis of the TRAMP4 complex, due to the similar composition of TRAMP5 (33) and high level of sequence conservation between Trf5p and Trf4p, and between Air1p and Air2p (5), data presented here are likely valid also to TRAMP5.

Dual function of Air1/2p in the TRAMP complex

Here, we demonstrate that Air2p is a *bona fide* RNA-binding protein. The Air2p ZnKs 2, 3 and 4 bind aberrant tRNA (Figure 3E and F); however, only the fourth ZnK is fully essential for TRAMP PAP activity *in vitro* (Figure 2A and B). Although other study by Hamill *et al.* (15) proposed its involvement in RNA binding, our analyses have not revealed any specific role for ZnK1. It is possible that ZnK1 binds other RNAs, not used in our tests. Alternatively, it may be involved in the interaction with other TRAMP-associated proteins. Structural and sequence alignment analyses indicate that the RNA-binding ZnKs of Air2p use a different binding mode compared to the NC proteins. They lack the hydrophobic pocket, which accommodates guanine base in the NC proteins. Instead, they have large electropositive patches (Figure 3G) created out of lysines, arginines and serines. We speculate that these residues could be responsible for the recognition of a broad spectrum of RNA substrates based on a low-sequence specificity interactions with the sugar-phosphate backbone.

The analysis of deletion mutants indicated the importance of the first 25 N-terminal aminoacids for the interaction with Mtr4p, as both the NΔ57 (residues 57–344) and ZnK1-5L (residues 26–205) proteins did not co-purify Mtr4p (Figure 1D, lanes NΔ57 and ZnK1-5L). The N-terminus is poorly conserved among Air1/2p homologs. In fact, a metazoan Air1-like protein (also designated ZCCHC7) contains a long N-terminal extension (data not shown). Given the observation that the human Mtr4p (SKIV2L2) forms the NEXT (Nuclear EXosome Targeting) complex independently of PAPD5/ZCCHC7 (the human homologs of Trf4p/Air1/2p) (34), it is possible that the interaction between noncanonical PAP (Trf4p/Air2p) and Mtr4p helicase has diverged during evolution depending on the organism complexity.

On the other hand, the Trf4p interacting residues in Air2p identified by NMR titration experiments at positions 139–143, W140G and R141E reside in a stretch of aminoacids forming the (P/T)xIWRxYxL motif conserved among Air-like proteins from distant species including humans (Supplementary Figure S1A). The protein pulldown experiments revealed strong defects in Trf4p and Mtr4p copurification with mutants in this conserved motif (L4 mutant) as well as in the mutant with the disrupted last fifth zinc knuckle fold (ZnK5) (Figure 1D). Mtr4p is not required for Trf4p–Air2p interaction. It is likely that dimerization of Air2p and Trf4p is needed for efficient assembly with Mtr4p, as the absence of either one results in loss of Mtr4p copurification (Figures 1D and 5A). Therefore, we believe that linker 4 and ZnK5 of Air2p are crucial regions responsible for interaction with Trf4p. The last ZnK also displays the highest degree of conservation among the five Air ZnKs (Supplementary Figure S1A). Likewise, the Trf4p surface interacting with the ZnK4–ZnK5 region is conserved among other Trf4-like proteins (15). This implies a phylogenetically conserved mode of interaction between Air1/2p homologs and Trf4p-like noncanonical PAPs. Although the human PAPD5 does not apparently require a cofactor for PAP activity *in vitro* (35), it does form a complex with ZCCHC7 and other putative RNA-binding proteins (34), suggesting that Air-like proteins mainly serve as RNA substrate specificity factors. During the preparation of this manuscript, Fasken and co-workers (36) reported similar results on Air1p as regard to the importance of the last two zinc knuckles for cell viability and IWRxY motif conservation, NEL025C degradation and Trf4p interaction (36). Our analysis goes beyond this work by distinguishing the different key roles of ZnK4 in RNA binding and Trf4p activation (see below) and ZnK5 in Trf4p interaction. Fasken and colleagues found that Air1p and Air2p mutants in ZnK5 are unstable at 37°C and to a lesser extend at 25°C, which is likely the reason of strong growth defects at high temperatures. Our western blot analysis of expression levels of the different Air2p variants shows only slight decrease in ZnK5 Air2p mutant at 25°C (Supplementary Figure S1B). Moreover, our *in vitro* reconstitution assays (Figure 2B) suggest that

this mutant is unable to form functional Trf4p-Air2p PAP even in amounts comparable to WT. Thus we conclude that the absence of Trf4p in ZnK5 immunoprecipitates is not caused by lowered *in vivo* expression of Trf4p, but due to weakened interactions between the two proteins as a response to the disruption of the fifth Air2p zinc knuckle. Furthermore, we show that Air2p mediates the interaction between Trf4 and Mtr4 proteins (Figure 5A). Our work extends to the analysis of the N- and C- termini of Air2p revealing crucial interactions with Mtr4p. The Air2p termini enhance Mtr4p binding, while the main surface for interaction with Trf4p is located in the fifth ZnK and the adjacent linker 4 regions (Figure 1D). Based on these mapping analyses and the prediction that the bottom of the Mtr4p helicase barrel serves for interaction with exosome (16), we propose a model of the quarternary interactions within the TRAMP complex. In this model, Trf4p/Air2p binds to the Mtr4p DExH globular core, at the side of barrel formed by RecA-2, winged helix and helical bundle domains. We suggest that the RNA substrates are first screened by the Air2p ZnK2 – ZnK4 region that protrudes from Mtr4 surface and are consequently loaded to the Trf4p active site. We speculate that the RNA 3′-oligo(A) terminus leaves Trf4p in the direction toward the KOW-RecA-2 interface, where the RNA 3′-end is expected to enter the helicase core (17) (Figure 6).

TRF4 catalytical mechanism

We found that increasing ATP concentration resulted in a gradual increase of the PAP activity in the D425A mutant up to the 4 mM concentration tested, whereas it did not change any further activity of the saturated WT protein (Figure 4G). This result indicated that the mutation reduced Trf4p affinity for ATP.

By examining the available crystal structure (15), we observed that the D425 residue is situated approximately 20 Å from the active site. Close inspection of the mechanism of the homologous yeast canonical poly(A) polymerase Pap1p catalysis revealed that hydrogen bonding in this region is required for stabilization of conformational movements leading from an open to closed conformation of the protein (37). When the N189A (CAT domain) mutation is introduced in Pap1p, the equilibrium between the open and closed state is shifted, resulting in increased K_M for ATP (37). N189 forms a hydrogen bond with the main chain carbonyl of the Y307 (CD) in the closed conformation (Supplementary Figure S3E and S3F). Interestingly, the Y307 of Pap1p resides in both, sequentially (30) and spatially, identical position to D425 of Trf4p (Supplementary Figure S3G). Therefore, we suggest that the side chain of D425 may serve as a hydrogen bond acceptor analogous to Y307 main chain carbonyl. We have not identified any hydrogen bond formed by D425 in the published structure [(15) and Supplementary Figure S3G]. Notably, the structure was solved for the Trf4p apoenzyme; hence it is likely to represent the open state status. We, therefore, suggest that a conformational change analogous to Pap1p may occur in Trf4p upon RNA and ATP binding, thus indicating the conservation of the catalytic mechanism of both related proteins.

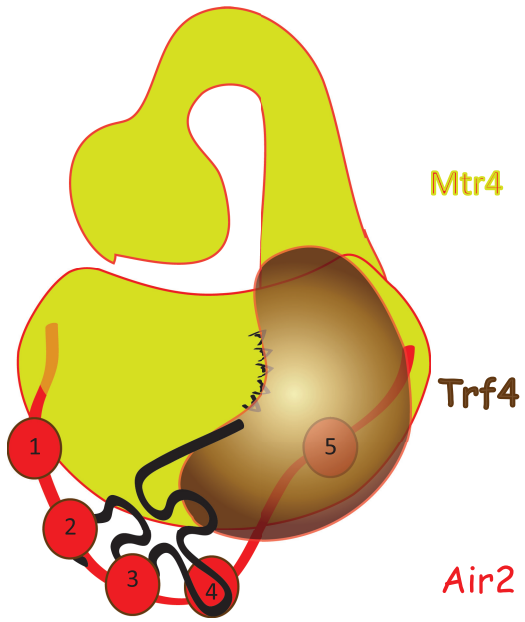


Figure 6. Model of TRAMP assembly and RNA binding. The Air2 protein is bridging the Mtr4p (interaction via N- and C-terminus) and Trf4p (interaction via L4, ZnK5 and C-terminus). The ZnKs 2, 3 and 4 are involved in RNA binding. The KOW domain of Mtr4p is required for exosome activation.

RNA binding KOW domain is involved in exosome activation

The role of the arch domain in exosome function *in vivo* has been proposed previously by Jackson *et al.* (16) who observed the accumulation of 5.8S+30 rRNA and stabilization of 5′-ETS, both being specific substrates of Rrp6p activity (38). Here we report that the arch, particularly the KOW domain, of Mtr4p is necessary for the TRAMP-mediated exosome activation *in vitro*. We show that TRAMPs lacking arch or KOW are able to add poly(A) tails, but they almost completely fail to activate exosomes to degrade structured RNAs *in vitro* (Figure 1F and G; Supplementary Figure S4). Recently Jia *et al.* (14) showed that Mtr4p regulates the individual steps of TRAMP polyadenylation as well as the length of the poly(A) tails added. We do not exclude the possibility that arch/KOW is also involved in this regulation. However, in our assays, NΔ143 and ΔKOW TRAMPs showed comparable polyadenylation, but only NΔ143 was fully able to initiate exosomal degradation, whereas ΔKOW was highly defective. Therefore, we conclude that the observed differences in polyadenylation seen in KOW-less TRAMP cannot explain the lack of exosome activation, and we propose that the KOW domain is crucial for TRAMP-mediated exosome activation.

As neither the helicase nor ATPase activities of Mtr4p are affected by arch removal (16,17), we speculate that the RNA-binding properties of the arch might be necessary to handle the substrate between different activities of the TRAMP-exosome pathway. Both the close proximity of

the KOW to the helicase RNA entry site and the possible flexibility in the elbow of the arm (16) suggest that arch transports RNA between Trf4p-Air2p, the helicase core and potentially the Rrp6p. Alternatively, the arch is required for the displacement of other RNA-binding factors and shuttling the substrate to the helicase core. We also cannot exclude the possibility that the arch serves for protein–protein interactions with Rrp6p or other core exosome components. This scenario is supported by the presence of a similar structure recently uncovered in the cytoplasmic exosome cofactor Ski2p (16,17,39).

The nuclear exosomes contain two confirmed exoribonucleases—the Rrp6p and Rrp44p. Currently, it is still not fully understood how these two subunits interplay on divergent exosome substrates *in vivo*. However, it was shown that in the conditions used in our *in vitro* assays (5 mM Mg^{2+}) the nuclease Rrp44p is inactive (12). Therefore, we propose that nuclease activity observed in our assays can be fully attributed to Rrp6p. Therefore, it will be interesting to address whether RNAs bound by KOW can also undergo processing by the core exosome or whether they are specifically triggered for the degradation by Rrp6p. The exact mechanistic details remain obscure and will be the subject of following studies.

ACCESSION NUMBERS

The atomic coordinates and restraints (code 2lli) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rcsb.org/).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–5, Supplementary Figures 1–4 and Supplementary Reference [40].

ACKNOWLEDGEMENTS

We thank Leona Svajdova for excellent technical assistance; Walter Keller and Georges Martin for helpful advice; and John LaCava, Jonathan Houseley and Andrzej Dziembowski for critical reading and comments on the manuscript. We also thank David Tollervey, Ambro Van Hoof and Domenico Libri for yeast strains and constructs, Patrick Lindner for Mtr4p antibodies, Gerhard Wagner for G-proteinB1 vector and Christopher Lima for Smt3p and Ulp1p expression vectors.

FUNDING

Wellcome Trust (084316/Z/07/Z to S.V.); EMBO Installation Grant (1642 to S.V.); Czech Science Foundation (305/11/1095 to S.V, 305/10/1490 to R.S., P305/12/G034 to S.V. and R.S.); CEITEC–Central European Institute of Technology (CZ.1.05/1.1.00/02.0068) from European Regional Development Fund; the NOESY spectra were obtained at the BMRZ NMR facility supported by the EU-NMR program (RII3-026145); the EC FP-7 (grant no 205872 to M.S.A.); the Brno City Municipality Scholarship for Talented Ph.D. Students (to P.H. and F.H.). Funding for open access charge: Senior International Wellcome Trust Fellowship.

Conflict of interest statement. None declared.

REFERENCES

1. Brosnan,C.A. and Voinnet,O. (2009) The long and the short of noncoding RNAs. *Curr. Opin. Cell. Biol.*, **21**, 416–425.
2. Philips,A.V. and Cooper,T.A. (2000) RNA processing and human disease. *Cell. Mol. Life Sci.*, **57**, 235–249.
3. Kadaba,S., Krueger,A., Trice,T., Krecic,A.M., Hinnebusch,A.G. and Anderson,J. (2004) Nuclear surveillance and degradation of hypomodified initiator tRNAMet in *S. cerevisiae*. *Genes Dev.*, **18**, 1227–1240.
4. LaCava,J., Houseley,J., Saveanu,C., Petfalski,E., Thompson,E., Jacquier,A. and Tollervey,D. (2005) RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell*, **121**, 713–724.
5. Vanacova,S., Wolf,J., Martin,G., Blank,D., Dettwiler,S., Friedlein,A., Langen,H., Keith,G. and Keller,W. (2005) A new yeast poly(A) polymerase complex involved in RNA quality control. *PLoS Biol.*, **3**, e189.
6. Wyers,F., Rougemaille,M., Badis,G., Rousselle,J.C., Dufour,M.E., Boulay,J., Regnault,B., Devaux,F., Namane,A., Seraphin,B. *et al.* (2005) Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell*, **121**, 725–737.
7. San Paolo,S., Vanacova,S., Schenk,L., Scherrer,T., Blank,D., Keller,W. and Gerber,A.P. (2009) Distinct roles of non-canonical poly(A) polymerases in RNA metabolism. *PLoS Genet.*, **5**, e1000555.
8. Wlotzka,W., Kudla,G., Granneman,S. and Tollervey,D. (2011) The nuclear RNA polymerase II surveillance system targets polymerase III transcripts. *EMBO J.*, **30**, 1790–1803.
9. Houseley,J., Kotovic,K., El Hage,A. and Tollervey,D. (2007) Trf4 targets ncRNAs from telomeric and rDNA spacer regions and functions in rDNA copy number control. *EMBO J.*, **26**, 4996–5006.
10. Houseley,J. and Tollervey,D. (2008) The nuclear RNA surveillance machinery: the link between ncRNAs and genome structure in budding yeast. *Biochim. Biophys. Acta*, **1779**, 239–246.
11. Thiebaut,M., Kisseleva-Romanova,E., Rougemaille,M., Boulay,J. and Libri,D. (2006) Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. *Mol. Cell.*, **23**, 853–864.
12. Callahan,K.P. and Butler,J.S. (2010) TRAMP complex enhances RNA degradation by the nuclear exosome component Rrp6. *J. Biol. Chem.*, **285**, 3540–3547.
13. D'Souza,V. and Summers,M.F. (2005) How retroviruses select their genomes. *Nat. Rev. Microbiol.*, **3**, 643–655.
14. Jia,H., Wang,X., Liu,F., Guenther,U.P., Srinivasan,S., Anderson,J.T. and Jankowsky,E. (2011) The RNA helicase Mtr4p modulates polyadenylation in the TRAMP complex. *Cell*, **145**, 890–901.
15. Hamill,S., Wolin,S.L. and Reinisch,K.M. (2010) Structure and function of the polymerase core of TRAMP, a RNA surveillance complex. *Proc. Natl Acad. Sci. USA*, **107**, 15045–15050.
16. Jackson,R.N., Klauer,A.A., Hintze,B.J., Robinson,H., van Hoof,A. and Johnson,S.J. (2010) The crystal structure of Mtr4 reveals a novel arch domain required for rRNA processing. *EMBO J.*, **29**, 2205–2216.
17. Weir,J.R., Bonneau,F., Hentschel,J. and Conti,E. (2010) Structural analysis reveals the characteristic features of Mtr4, a

DExH helicase involved in nuclear RNA processing and surveillance. *Proc. Natl Acad. Sci. USA*, **107**, 12139–12144.
18. Zhou,P., Lugovskoy,A.A. and Wagner,G. (2001) A solubility-enhancement tag (SET) for NMR studies of poorly behaving proteins. *J. Biomol. NMR*, **20**, 11–14.
19. Price,S., Oubridge,C., Varani,G. and Nagai,K. (1998) In: Smith,C. (ed.), *RNA–Protein Interaction: Practical Approach*. Oxford University Press, Oxford, pp. 37–74.
20. Heyduk,T. and Lee,J.C. (1990) Application of fluorescence energy transfer and polarization to monitor *Escherichia coli* cAMP receptor protein and lac promoter interaction. *Proc. Natl Acad. Sci. USA*, **87**, 1744–1748.
21. Mossessova,E. and Lima,C.D. (2000) Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast. *Mol. Cell.*, **5**, 865–876.
22. Sanudo,M., Jacko,M., Rammelt,C., Vanacova,S. and Stefl,R. (2011) 1H, 13C, and 15N chemical shift assignments of ZCCHC9. *Biomol. NMR Assign.*, **5**, 19–21.
23. Sattler,M., Schleucher,J. and Griesinger,C. (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Magn. Reson. Spectrosc.*, **34**, 93–158.
24. Herrmann,T., Guntert,P. and Wuthrich,K. (2002) Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J. Biomol. NMR*, **24**, 171–189.
25. Case,D.A., Darden,T.A., Cheatham,T.E. III, Simmerling,C.L., Wang,J., Duke,R.E., Luo,R., Crowley,M., Walker,R.C., Zhang,W. *et al.* (2008) *AMBER 10, Users' Manual*. University of California, San Francisco, CA, USA.
26. Stefl,R., Oberstrass,F.C., Hood,J.L., Jourdan,M., Zimmermann,M., Skrisovska,L., Maris,C., Peng,L., Hofr,C., Emeson,R.B. *et al.* (2010) The solution structure of the ADAR2 dsRBM-RNA complex reveals a sequence-specific readout of the minor groove. *Cell*, **143**, 225–237.
27. Hobor,F., Pergoli,R., Kubicek,K., Hrossova,D., Bacikova,V., Zimmermann,M., Pasulka,J., Hofr,C., Vanacova,S. and Stefl,R. (2011) Recognition of transcription termination signal by the nuclear polyadenylated RNA-binding (NAB) 3 protein. *J. Biol. Chem.*, **286**, 3645–3657.
28. Falquet,L., Bordoli,L., Ioannidis,V., Pagni,M. and Jongeneel,C.V. (2003) Swiss EMBnet node web server. *Nucleic Acids Res.*, **31**, 3782–3783.
29. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
30. Rogozin,I.B., Aravind,L. and Koonin,E.V. (2003) Differential action of natural selection on the N and C-terminal domains of 2′-5′ oligoadenylate synthetases and the potential nuclease function of the C-terminal domain. *J. Mol. Biol.*, **326**, 1449–1461.
31. Gellon,L., Carson,D.R., Carson,J.P. and Demple,B. (2008) Intrinsic 5′-deoxyribose-5-phosphate lyase activity in *Saccharomyces cerevisiae* Trf4 protein with a possible role in base excision DNA repair. *DNA Repair (Amst)*, **7**, 187–198.
32. Keller,C., Woolcock,K., Hess,D. and Buhler,M. (2010) Proteomic and functional analysis of the noncanonical poly(A) polymerase Cid14. *RNA*, **16**, 1124–1129.
33. Houseley,J. and Tollervey,D. (2006) Yeast Trf5p is a nuclear poly(A) polymerase. *EMBO Rep.*, **7**, 205–211.
34. Lubas,M., Christensen,M.S., Kristiansen,M.S., Domanski,M., Falkenby,L.G., Lykke-Andersen,S., Andersen,J.S., Dziembowski,A. and Jensen,T.H. (2011) Interaction profiling identifies the human nuclear exosome targeting complex. *Mol. Cell.*, **43**, 624–637.
35. Rammelt,C., Bilen,B., Zavolan,M. and Keller,W. (2011) PAPD5, a noncanonical poly(A) polymerase with an unusual RNA-binding motif. *RNA*, **17**, 1737–1746.
36. Fasken,M.B., Leung,S.W., Banerjee,A., Kodani,M.O., Chavez,R., Bowman,E.A., Purohit,M.K., Rubinson,M.E., Rubinson,E.H. and Corbett,A.H. (2011) Air1 zinc knuckles 4 and 5 and a conserved IWRXY motif are critical for the function and integrity of the Trf4/5-Air1/2-Mtr4 polyadenylation (TRAMP) RNA quality control complex. *J. Biol. Chem.*, **286**, 37429–37445.
37. Balbo,P.B. and Bohm,A. (2007) Mechanism of poly(A) polymerase: structure of the enzyme-MgATP-RNA ternary complex and kinetic analysis. *Structure*, **15**, 1117–1131.
38. Allmang,C., Kufel,J., Chanfreau,G., Mitchell,P., Petfalski,E. and Tollervey,D. (1999) Functions of the exosome in rRNA, snoRNA and snRNA synthesis. *EMBO J.*, **18**, 5399–5410.
39. Halbach,F., Rode,M. and Conti,E. (2012) The crystal structure of *S. cerevisiae* Ski2, a DExH helicase associated with the cytoplasmic functions of the exosome. *RNA*, **18**, 124–134.
40. Bard,J., Zhelkovsky,A.M., Helmling,S., Earnest,T.N., Moore,M.J. and Bohm,A. (2000) Structure of yeast poly(A) polymerase alone and in complex with 3′-dATP. *Science*, **289**, 1346–1349.

⌘ *Author's Choice*

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 286, NO. 5, pp. 3645–3657, February 4, 2011

Recognition of Transcription Termination Signal by the Nuclear Polyadenylated RNA-binding (NAB) 3 Protein*

Received for publication, June 25, 2010, and in revised form, November 9, 2010 Published, JBC Papers in Press, November 17, 2010, DOI 10.1074/jbc.M110.158774

Fruzsina Hobor[‡1,2], Roberto Pergoli[‡1,3], Karel Kubicek[‡], Dominika Hrossova[‡], Veronika Bacikova[‡], Michal Zimmermann[§2,4], Josef Pasulka[‡], Ctirad Hofr[§], Stepanka Vanacova[‡], and Richard Stefl[‡5]

From the [‡]National Centre for Biomolecular Research and [§]Department of Functional Genomics and Proteomics, Central European Institute of Technology, Faculty of Science, Masaryk University, Brno CZ-62500, Czechia

Non-coding RNA polymerase II transcripts are processed by the poly(A)-independent termination pathway that requires the Nrd1 complex. The Nrd1 complex includes two RNA-binding proteins, the nuclear polyadenylated RNA-binding (Nab) 3 and the nuclear pre-mRNA down-regulation (Nrd) 1 that bind their specific termination elements. Here we report the solution structure of the RNA-recognition motif (RRM) of Nab3 in complex with a UCUU oligonucleotide, representing the Nab3 termination element. The structure shows that the first three nucleotides of UCUU are accommodated on the *β*-sheet surface of Nab3 RRM, but reveals a sequence-specific recognition only for the central cytidine and uridine. The specific contacts we identified are important for binding affinity *in vitro* as well as for yeast viability. Furthermore, we show that both RNA-binding motifs of Nab3 and Nrd1 alone bind their termination elements with a weak affinity. Interestingly, when Nab3 and Nrd1 form a heterodimer, the affinity to RNA is significantly increased due to the cooperative binding. These findings are in accordance with the model of their function in the poly(A) independent termination, in which binding to the combined and/or repetitive termination elements elicits efficient termination.

RNA Polymerase II (RNA Pol II)[6] transcribes messenger RNA (mRNA), but also a subset of small nuclear and small nucleolar RNAs (snRNAs/snoRNAs), micro-RNA precursors, and a class of intergenic and antisense RNAs (1). RNA Pol II uses two different mechanisms for transcription termination of these "coding" and "non-coding" RNAs. Although the RNA Pol II termination of mRNA requires a large multiprotein complex that recognizes the poly(A) signal in the nascent transcript (2), the termination of the non-coding RNAs requires no poly(A) signal (2–4).

In the poly(A)-independent mechanism, transcription termination requires a specific factor, the Nrd1 complex. This complex consists of three proteins: the nuclear pre-mRNA down-regulation (Nrd) 1 protein, the nuclear polyadenylated RNA-binding (Nab) 3 protein, and the putative RNA helicase Sen1 (5–7). The Nrd1 complex interacts with the exosome, a complex of 10–12 exoribonucleolytic and RNA-binding proteins (8) and the Trf4-Air2-Mtr4 polyadenylation (TRAMP) complex (9–11), which are involved in the 3′ end processing of non-coding RNA transcripts (3, 4, 7).

In yeast, transcription termination mediated by the Nrd1 complex requires binding to both the nascent RNA and the carboxyl-terminal domain of RNA Pol II, which consists of 26 repeats of the sequence Tyr[1]-Ser[2]-Pro[3]-Thr[4]-Ser[5]-Pro[6]-Ser[7] (1, 12). Interestingly, the Nrd1 complex binds the carboxyl-terminal domain when it is phosphorylated at Ser[5], a typical feature of the early elongation phase of the transcription cycle. The Ser[5]-phosphorylated carboxyl-terminal domain is recognized by the carboxyl-terminal domain-interacting domain of Nrd1 (13, 14). The RNA-binding subunits of the Nrd1 complex, Nrd1 and Nab3, recognize their specific RNA sequences (called terminator elements) in the nascent transcripts of RNA Pol II. It is believed that this specific binding of Nrd1 complex to the terminator elements is the initial step in the assembly of termination machinery.

A number of studies narrowed the sequence regions with terminator elements (5, 6, 15–17) that were subsequently identified as GUAR (where R stands for purine) and UCUU sequences (18). GUAR and UCUU terminator elements are recognized by Nrd1 and Nab3, respectively, via their fragments encompassing RNA recognition motifs (RRMs) (18). These terminator sequences are located downstream of snRNA and snoRNA genes (18) although their relative orientation and spacing are not highly conserved. In addition, it was demonstrated that Nrd1 and Nab3 form a stable het-

* This work was supported in part by a Howard Hughes Medical Institute/European Molecular Biology Organization start-up grant, a Human Frontier Science Program Career Development Award, Ministry of Education of the Czech Republic Grants MSM0021622413, MSM0021622415, and Ingo LA08008, Czech Science Foundation Grants 204/08/1212 and 305/10/1490, Grant Agency of the Academy of Sciences of the Czech Republic IAA401630903, Wellcome Trust Grant 084316/Z/07/Z, and EMBO Installation Grant 1642.

The atomic coordinates and structure factors (codes 2KVI and 2L41) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rcsb.org/).

⌘ *Author's Choice*—Final version full access.

[1] Both authors contributed equally to this work.

[2] Supported by Brno City Municipality Scholarships for Talented Ph.D. Students.

[3] Supported by the European Community FP-7 Grant 205872.

[4] Supported by the GACR Grant 204/08/H054.

[5] To whom correspondence should be addressed: University Campus Bohunice, Kamenice 5/A4, Brno, CZ-62500 Brno, Czechia. Tel.: 420549492436; Fax: 420549492556; E-mail: stefl@chemi.muni.cz.

[6] The abbreviations used are: RNA Pol II, RNA polymerase II; FA, fluorescence anisotropy; Nab3, nuclear polyadenylated RNA-binding 3; Nrd1, nuclear pre-mRNA down-regulation 1; RRM, RNA-recognition motif; snoRNAs, small nucleoar RNAs; HSQC, heteronuclear single quantum coherence; RNP, ribonucleoprotein; TAMRA, *N*,*N*,*N*′,*N*′-tetramethyl-6-carboxyrhodamine; PTB, polypyrimidine tract-binding protein.

erodimer and bind to snoRNA terminators that contain multiple Nrd1- and Nab3-binding sequences (19).

Both Nrd1 and Nab3 contain RRM that likely mediates the binding to their specific RNA sequences. The RRM is the most abundant RNA-binding domain in higher vertebrates; *e.g.* the RRM is present in about 2% of human genes (20). It is a small protein domain of ~90 amino acids with a typical $\beta\alpha\beta\beta\alpha\beta$ topology that forms a four-stranded β-sheet packed against two α-helices (21–23). The structure of this domain is relatively well defined despite a little sequence conservation among various RRMs. The solved structures of RRM bound to RNA show the complexity of protein-RNA recognition mediated by the RRM, which often involves not only RRM-RNA interactions but also RRM-RRM and other RRM-protein interactions. The main protein surface of the RRM involved in the interaction with the RNA is the four-to-five-stranded β-sheet, which typically contacts two or three nucleotides. Frequently, RRM-containing proteins bind more than three nucleotides and recognize longer single-stranded RNA or even internal RNA loops by employing of β-strand loops and N- or C-terminal flanking regions of RRMs (21–23).

To better understand the structural basis behind the poly(A) independent transcription termination pathway, we initiated an NMR study of *Saccharomyces cerevisiae* Nab3. Here, we present the three-dimensional solution structure of the Nab3 RRM in free form and in complex with the 5′-UCUU-3′ RNA substrate. The structure of the complex reveals recognition of the YCU sequence (where Y stands for pyrimidine) by the Nab3 RRM. We confirmed the sequence-specific intermolecular contacts by site-directed mutagenesis and fluorescence anisotropy (FA) measurements, and their physiological role was also confirmed by yeast phenotypic analyses. Finally, we demonstrate that the weak RNA binding of the isolated RRMs of Nab3 and Nrd1 is greatly enhanced when Nab3 and Nrd1 form a heterodimer and bind the RNA cooperatively.

EXPERIMENTAL PROCEDURES

Cloning, Expression, and Purification of Proteins—The coding sequence corresponding to the RRM of the *Nab3* gene from *S. cerevisiae* (961–1245) was amplified by polymerase chain reaction (PCR), and cloned into a pET22b expression vector (Novagen) via NdeI and XhoI restriction sites. The resulting C-terminal His_6-tagged construct was verified by DNA sequencing. The protein was overexpressed in *Escherichia coli* BL21-Codon Plus (DE3)-RIPL (Stratagene), transformed with the pET22b-RRM Nab3 construct at 37 °C in M9 minimal medium, supplemented with 50 mg/liter of ampicillin. For isotope labeling, the medium was supplemented with $^{15}NH_4Cl$ and [U-$^{13}C_6$]glucose. Cells were grown at 37 °C to A_{600} ~1 and induced with 1 mM isopropyl β-D-thiogalactoside. Cells were harvested by centrifugation (6000 × *g* for 10 min), resuspended in lysis buffer (50 mM sodium phosphate, 300 mM NaCl, 10 mM β-mercaptoethanol, pH 8), and disrupted by sonication. The cell debris was cleared by centrifugation (14,000 × *g* for 60 min). Soluble lysate was loaded on a nickel-nitrilotriacetic acid column (Qiagen), equilibrated with lysis buffer, washed with a high salt buffer (50 mM sodium phosphate, 500 mM NaCl, 10 mM β-mercaptoethanol, 5 mM imidazole, pH 8), and eluted with imidazole gradient (50–500 mM) of elution buffer (50 mM sodium phosphate, 300 mM NaCl, 10 mM β-mercaptoethanol, pH 8). The protein was subsequently loaded on a Superdex 75 gel filtration column (GE Healthcare), equilibrated with lysis buffer. The protein fractions from gel filtration were dialyzed against lysis buffer. The purified protein was 99% pure, as judged by Coomassie-stained SDS-PAGE. For NMR measurements the pure protein was concentrated to 2.5 mM in 550 μl of 50 mM sodium phosphate (pH 8.0), containing 300 mM NaCl, and 10 mM β-mercaptoethanol. The cloning, expression, and purification of Nrd1 RRM-(340–410) were carried out in the same way as for Nab3 RRM.

The expression and purification of the Nrd1-Nab3 heterodimer have been done in a similar manner as reported previously (19). To improve the yield of expression, we used *E. coli* BL21-Codon Plus (DE3)-RIPL (Stratagene). We used the following final buffer (50 mM Tris (pH 8.0), containing 150 mM NaCl, and 10 mM β-mercaptoethanol) to have the same conditions for all fluorescence anisotropy measurements. Prior to RNA titration, all proteins were tested for the residual RNase activity using RNaseAlert Lab Test (Ambion). RNA oligonucleotides were purchased from Thermo Fisher Scientific/Dharmacon and Sigma.

Generation of Nab3 RRM Mutants—Site-specific mutagenesis was performed using the QuikChange site-directed mutagenesis kit (Stratagene) with complementary sense and antisense (AS) oligonucleotide primers as follows: R331A (S), 5′-gcacaatattcctccgaagtcagcattattcattggtaatttgccg-3′ and (AS), 5′-cggcaaattaccaatgaataatgctgacttcggaggaatattgtgc-3′; N361A (S), 5′-tccatacggtcatatcatgcaaatcgctatcaaaaatgcctttggattcatt-3′ and (AS), 5′-aatgaatccaaaggcatttttgatagcgatttgcatgatatgaccgtatgga-3′; E397A (S), 5′-gcaaaaagttgatcctggcagtttctagctcgaatgc-3′ and (AS), 5′-gcattcgagctagaaactgccaggatcaacttttgc-3′; E397K (S), 5′-tggcaaaaagttgatcctgaaagtttctagctcgaatgc-3′ and (AS) 5′-gcattcgagctagaaactttcaggatcaacttttgcca-3′; S399A (S), 5′-agttgatcctggaagttgctagctcgaatgctcgt-3′ and (AS), 5′-acgagcattcgagctagcaacttccaggatcaact-3′; S399K (S), 5′-actttggcaaaaagttgatcctggaagttaaaagctcgaatgctcgtcc-3′ and (AS), 5′-ggacgagcattcgagcttttaacttccaggatcaacttttgccaaagt-3′. All mutations were verified by DNA sequence analysis.

NMR Spectroscopy—All NMR spectra of 2.5 mM uniformly ^{15}N,^{13}C-labeled Nab3 RRM in 50 mM sodium phosphate buffer (pH 8.0), 300 mM NaCl, 10 mM β-mercaptoethanol (90% H_2O/10% D_2O) were recorded on Bruker AVANCE 600 and 900 MHz spectrometers equipped with a cryoprobe at a sample temperature of 30 °C. All spectra were processed with Topspin 2.1 (Bruker BioSpin) and analyzed with Sparky 3.0 (T. G. Goddard and D. G. Kneller, University of California, San Francisco). The ^1H, ^{13}C, and ^{15}N chemical shifts of Nab3 RRM were assigned as described previously (24). All distance restraints were derived from the three-dimensional ^{15}N- and ^{13}C-edited NOESYs and two-dimensional ^1H-^1H-labeled NOESY (with mixing time of 150 ms) collected at 900 MHz spectrometer.

Akin to the free Nab3 RRM, the backbone resonance assignments of Nab3 RRM in the bound form were achieved

using three-dimensional triple resonance experiments: HNCA, HNCACB, and CBCA(CO)NH (24–26). Resonances of the aliphatic side chains were assigned by a combination of three-dimensional HCCH TOCSY, three-dimensional HNHA, and three-dimensional ^{13}C-edited NOESY spectra. Resonances of the aromatic side chains were assigned using two-dimensional homonuclear NOESY, three-dimensional $^{13}C_{arom}$-edited NOESY, and two-dimensional (HB)CB(CGCD)HD spectra. All distance restraints were derived from the three-dimensional ^{15}N- and ^{13}C-edited NOESYs (with mixing time of 150 ms) collected at 900 MHz spectrometer. The RNA resonances in complex were assigned using a combination of standard through-space and through-bond experiments (27, 28). We could not assign all sugar resonances unambiguously due to a high resonance overlap in the spectra (only unlabeled RNA was used). The sugar pucker conformation for all nucleotides is *C2′-endo*, as identified in the two-dimensional homonuclear TOCSY spectrum (strong cross-peaks between the H1′ and H2′ resonances). Intermolecular distance constraints were obtained from the three-dimensional ^{13}C F_1-edited, F_3-filtered NOESY-HSQC experiment (29), which was recorded in H_2O (with WATERGATE water suppression) as a two-dimensional filter NOESY omitting ^{13}C chemical shift evolution.

Structure Calculations—The preliminary structure determinations of the free and bound Nab3 RRM were performed with the automated NOE assignment module implemented in the CYANA program (30). This automated NOE assignment procedure is a re-implementation of the former CANDID algorithm (31) on the basis of a probabilistic treatment of the NOE assignment. CYANA carries out automated assignment and distance calibration of NOE intensities, removal of meaningless restraints, structure calculation with torsion angle dynamics, and automatic upper distance limit violation analysis. The resultant NOE cross-peak assignments were subsequently confirmed by visual inspection of the spectra. The predicted protein backbone ϕ and ψ torsion angle from the chemical shifts (32) for the secondary structure elements were also included in the calculations. In the next step, CYANA-generated restraints along with manually assigned protein-RNA intermolecular restraints were used for further refinement of the preliminary structures with AMBER 10.0 software (33). This calculations employed a modified version (AMBER ff99SB) of the force field described by Cornell *et al.* (34) along with a refinement protocol described in Padrta *et al.* (35), and an explicit solvent. From 40 refined structures, the 20 conformers with the lowest AMBER energy were selected to form the final ensemble of structures. Structural quality was assessed using PROCHECK (36) and WHAT IF (37). Molecular graphics were generated using MOLMOL (38) and PyMOL (57).

Fluorescence Anisotropy Measurements—The equilibrium binding of Nab3 RRM to different oligonucleotides was analyzed by fluorescence anisotropy. The RNA oligonucleotides were either 5′-labeled with TAMRA or fluorescein attached via a hexyl linker. The measurements were conducted on a FluoroMax-4 spectrofluorometer (Horiba Jobin-Yvon Edison, NJ). The instrument was equipped with a thermostatted cell holder with a Neslab RTE7 water bath (Thermo Scientific). The whole system was operated using FluorEssence software (version 2.5.3.0, Horiba Jobin-Yvon). The TAMRA fluorophore was excited at 561 nm and its emission was collected at 581 nm. The widths of both excitation and emission monochromatic slits were 8 nm and integration time was set to 3 s. The fluorescein fluorophore was excited at 488 nm and its emission was collected at 520 nm. For measurement with the individual domains (Nrd1-(340–410) and Nab3-(331–415)) the width of both excitation and emission monochromatic slits were 7 nm and the integration time was set to 3 s. 10 nM labeled oligonucleotide (volume 1.4 ml) was titrated with increasing amounts of the protein in 50 mM sodium phosphate buffer (pH 7.5), supplemented with 150 mM NaCl and 10 mM β-mercaptoethanol. For measurement with the heterodimer (Nab3-(191–565)–Nrd1-(1–548)), 1 nM fluorescein-labeled RNA was used. Both excitation and emission monochromatic slits were 14 nm, the integration time was set to 3 s.

In all measurements, an identical concentration of the oligonucleotide was included in the protein stock solution to prevent dilution of the RNA during titration. All experiments were carried out at 25 °C in a stirred 1.5-ml quartz cuvette. Protein aliquots were added stepwise until the cuvette was filled. After this point, a certain volume of the sample was always removed from the cuvette before addition of a protein aliquot of the same volume. A fixed delay of 30 s was set between each aliquot addition and start of the measurement to allow the reaction to reach equilibrium. This delay was sufficient, as no further change in anisotropy was observed. Each data point is an average of five measurements. Neither TAMRA nor fluorescein labels showed binding with any of the protein constructs.

The data were analyzed in SigmaPlot 11 software (Systat Software). The experimental isotherms were fit to a single-site binding model according to Heyduk and Lee (39) using nonlinear least squares regression. The data were normalized for visualization purposes.

Yeast Strains and Plasmids—The plasmid for the expression of wild-type Nab3 in yeast contain 550 bp of NAB3 promoter upstream of the AUG and a 286-bp sequence downstream of the stop codon. SV320 contains the wild-type NAB3 inserted in pRS415 (a LEU2 CEN plasmid) (40). Plasmids SV321-SV326 contain point mutants R331A, N361A, E397A, E397K, S399A, and S399K, respectively (see above for the primers). Strain DLY889 containing the endogenous NAB3 under control of the GAL1 promoter (3) was transformed with plasmids containing either wild-type Nab3 or Nab3 RRM point mutants with a LEU2 selectable marker (SV320, SV321, SV322, SV323, SV324, SV325, and SV326, respectively). The resulting strains were used for growth tests and Western blot analyses.

Growth Test Analyses—To test whether the mutated residues were essential for growth, the resulting transformants were grown in SC-LEU-HIS + 2% galactose at 30 °C to an A_{600} 1.0. The cultures were then serially diluted in 96-well plates by a factor of 10, and spotted onto SC-LEU-HIS medium containing 2% glucose to repress the expression of the endogenous NAB3 or control medium (SC-HIS + 2% galac-

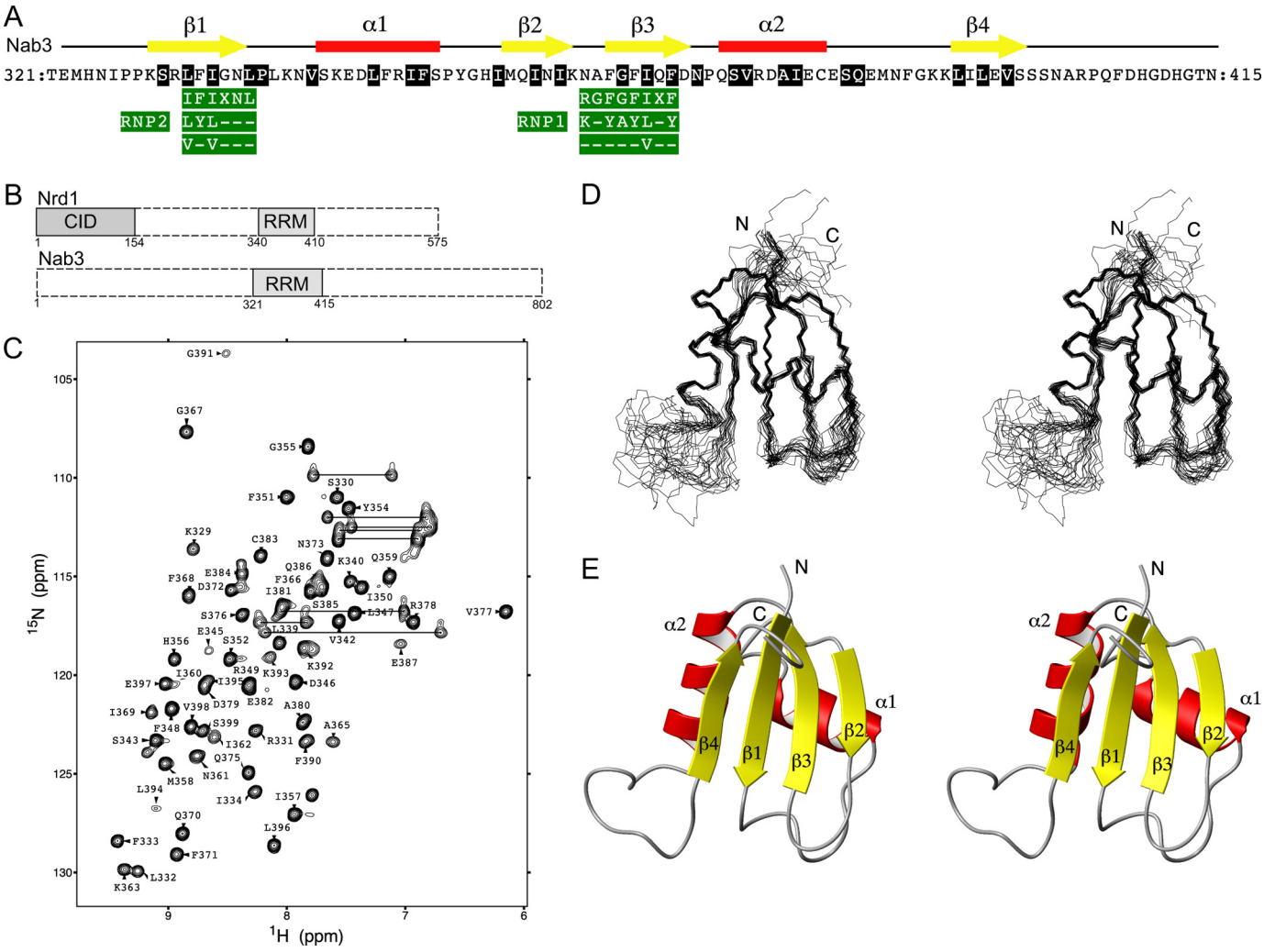


FIGURE 1. **Overview of the RRM of Nab3 sequence, topology, NMR spectra, solution structure, and domain structure in Nab3 and Nrd1.** *A*, amino acid sequence of the *S. cerevisiae* Nab3 RRM along with its secondary structure elements and general consensus of RNP1 and RNP2 motifs. *B*, a schematic drawing of the domain structure of Nab3 and Nrd1. *C*, two-dimensional ^{1}H-^{15}N HSQC spectrum of 2.5 mM uniformly ^{15}N,^{13}C-labeled Nab3 RRM in 50 mM sodium phosphate buffer (pH 8.0), 300 mM NaCl, and 10 mM β-mercaptoethanol (90% H_2O, 10% D_2O). The spectrum was acquired at 303 K on a Bruker Avance 600 MHz spectrometer. The assignments are labeled by the one-letter code of amino acids accompanied by a sequence number. The side chain resonances of asparagine and glutamine are connected by *horizontal lines*. *D*, stereo view of the 20 lowest energy structures of Nab3 RRM. The protein backbone is shown as a wire model. *E*, stereo view of the representative (the lowest energy) structure of Nab3 RRM shown as a *ribbon diagram*. The figure was generated with MOLMOL (38).

tose). These plates were incubated at 24, 30, and 37 °C for 3 days.

Western Blot Analysis—Protein extracts were prepared from cultures grown either on galactose containing medium and cultures shifted to glucose containing medium (as described above). Proteins were resolved on a 12% SDS-PAGE gel, transferred to nitrocellulose membrane by a semi-dry electroblotter (Bio-Rad), and probed for the presence of Nab3p with the mAb 2F12 (41), or antibodies directed against the HA epitope (sc805, Santa Cruz Biotechnology) present on the endogenous Nab3. For loading control we used the antibodies against Air2 protein (42).

RESULTS

Structure of Nab3 RRM—The RRM of *S. cerevisiae* Nab3 was examined by NMR spectroscopy (Fig. 1, *A* and *B*). The ^{1}H, ^{13}C, and ^{15}N chemical shift assignments were obtained as described previously (24). All NMR experiments were measured at a high salt concentration (300 mM NaCl, 50 mM sodium phosphate (pH 8.0), and 10 mM β-mercaptoethanol) to prevent protein precipitation. The solution structure determination of Nab3 RRM employed homonuclear and heteronuclear NMR techniques. The ^{15}N-^{1}H HSQC experiment shows a well dispersed spectrum (Fig. 1*C*), indicating a folded domain. There are a number of missing peaks in this spectrum that mainly correspond to the N- and C-terminal regions of the studied protein construct (outside of the RRM domain). These regions were included in the study as they often form additional structural elements (β-strand or α-helix) in RRMs or contribute to the RNA binding. In addition, several residues in the loops showed no NMR signals. These missing signals are likely a result from the relatively high pH used in the NMR study that was necessary to prevent the precipitation of Nab3 RRM. The three-dimensional structure of Nab3 RRM was determined by combined automated NOESY cross-peak assignment (30) and structure calculations with

TABLE 1
NMR and refinement statistics for Nab3 RRM and Nab3 RRM–UCUU complex

NMR distance and dihedral angle restraints	Nab3 RRM	Nab3 RRM–UCUU complex
Distance restraints		
Total NOEs	857	852
Intra-residue	201	228
Inter-residue		
Sequential ($\|i-j\| = 1$)	218	187
Medium range ($1 < \|i-j\| < 5$)	164	116
Long range ($\|i-j$[bar ≥ 5)	274	310
Hydrogen bond restraints	25	25
Intermolecular		11
Dihedral angle restraints		
ϕ and ψ	76	88[a]
Structure statistics[b]		
Residual NOE violations (mean ± S.D.)		
Number > 0.20 Å	1.5 (± 0.83)	4 (± 2)
Maximum (Å)	0.24 (± 0.03)	0.46 (± 0.08)
Residual dihedral angle violations		
Number > 10.0°	0	0
Maximum (°)	0	0
Ramachandran plot statistics[b,c,d]		
Residues in most favored regions (%)	90.8	88.6
Residues in additionally allowed regions (%)	9.1	9.6
Residues in generously allowed regions (%)	0.1	1.7
Residues in disallowed regions (%)	0.0	0.1
Deviations from idealized geometry		
Bond length (Å)	0.0011 ± 0.0001	0.0010 ± 0.0001
Bond angles (Å)	1.46 ± 0.02	1.48 ± 0.02
Average root mean square deviation to mean structure (Å)[b]		
Protein		
Backbone atoms[b]	0.57 ± 0.12	0.49 ± 0.10
Heavy atoms[b]	1.46 ± 0.16	1.24 ± 0.12
RNA		
All RNA heavy atoms[e]		0.98 ± 0.21
Complex		
Protein[b] and RNA heavy atoms[e]		1.27 ± 0.12
WHAT IF[f] structure *Z*-scores[d,g]		
Packing quality	−2.3	−1.9
Ramachandran plot appearance	−3.5	−3.1

[a] Includes *C2′-endo* sugar pucker and *anti* conformation of the glycosidic bond was used for all nucleotides (56).
[b] Calculated for an ensemble of the 20 lowest energy structures.
[c] Based on PROCHECK analysis (32).
[d] Calculated for the structured part of the protein construct.
[e] Calculated for $U_1C_2U_3$.
[f] Based on WHAT IF analysis (37).
[g] *Z*-score (54, 55) is defined as the deviation from the average value for this indicator observed in a database of high-resolution crystal structures.

torsion angle dynamics implemented in the program CYANA 2.1 (43), followed by refinement in explicit solvent using AMBER 10 (33). An ensemble of the 20 lowest energy structures along with the best energy structure are shown in Fig. 1, *D* and *E*, respectively. These structures have an average backbone root mean square deviation of 0.57 ± 0.12 Å for the secondary structure elements. A full summary of structural statistics including the backbone ϕ-ψ angle distribution is given in Table 1.

The three-dimensional structure of Nab3 RRM adopts a compact fold with an $\beta1\alpha1\beta2\beta3\alpha2\beta4$ topology that is similar to the canonical fold of RRM family (21, 22). The fold is composed of two α-helices and a 3_{10} helix that are packed along a face of a four-stranded antiparallel β-sheet. A central hydrophobic core composed of the residues shown in Fig. 1*A* stabilizes the fold of the domain. Nab3 RRM contains a well conserved signature of the RRM family, RNP1 and RNP2 sequences (44–46). These two conserved amino acid sequences found between Leu^{332}–Leu^{337} and Asn^{364}–Phe^{371} are located on the β3- and β1-strands, respectively. Their sequence compositions correspond to the general RNP2 and RNP1 consensus (ILV)-(FY)-(ILV)-*X*-N-L and (RK)-G-(FY)-(GA)-(FY)-(ILV)-*X*-(FY), respectively, except for the first two amino acids of the RNP1 (Fig. 1*A*). Nab3 RRM has asparagine and alanine in these positions (Fig. 1*A*). The presence of aromatic residues in RNP1 and RNP2 sequences, which usually mediates the stacking interaction with RNA bases, along with a number of basic and polar residues on the β-sheet surface, indicate a potential role of Nab3 RRM in RNA binding.

Characterization of the Nab3-UCUU Interactions by NMR—To investigate the interaction and binding mode between Nab3 RRM and RNA, we carried out an NMR chemical shift perturbation study with a UCUU element, which has been shown to elicit transcription termination via the Nrd1 pathway. In the RNA titration experiment, we observed that the protein amide resonances moved upon RNA binding from their initial positions, corresponding to the free form, in a stepwise directional manner until they reach their final positions that correspond to the fully bound state, with stoichiometry of 1:1 (Fig. 2). Additional RNA aliquots resulting in excess RNA resulted in no further change of chemical shifts, confirming the 1:1 stoichiometry of the complex. These titration data suggest that protein amide resonances are in a fast exchange regime between

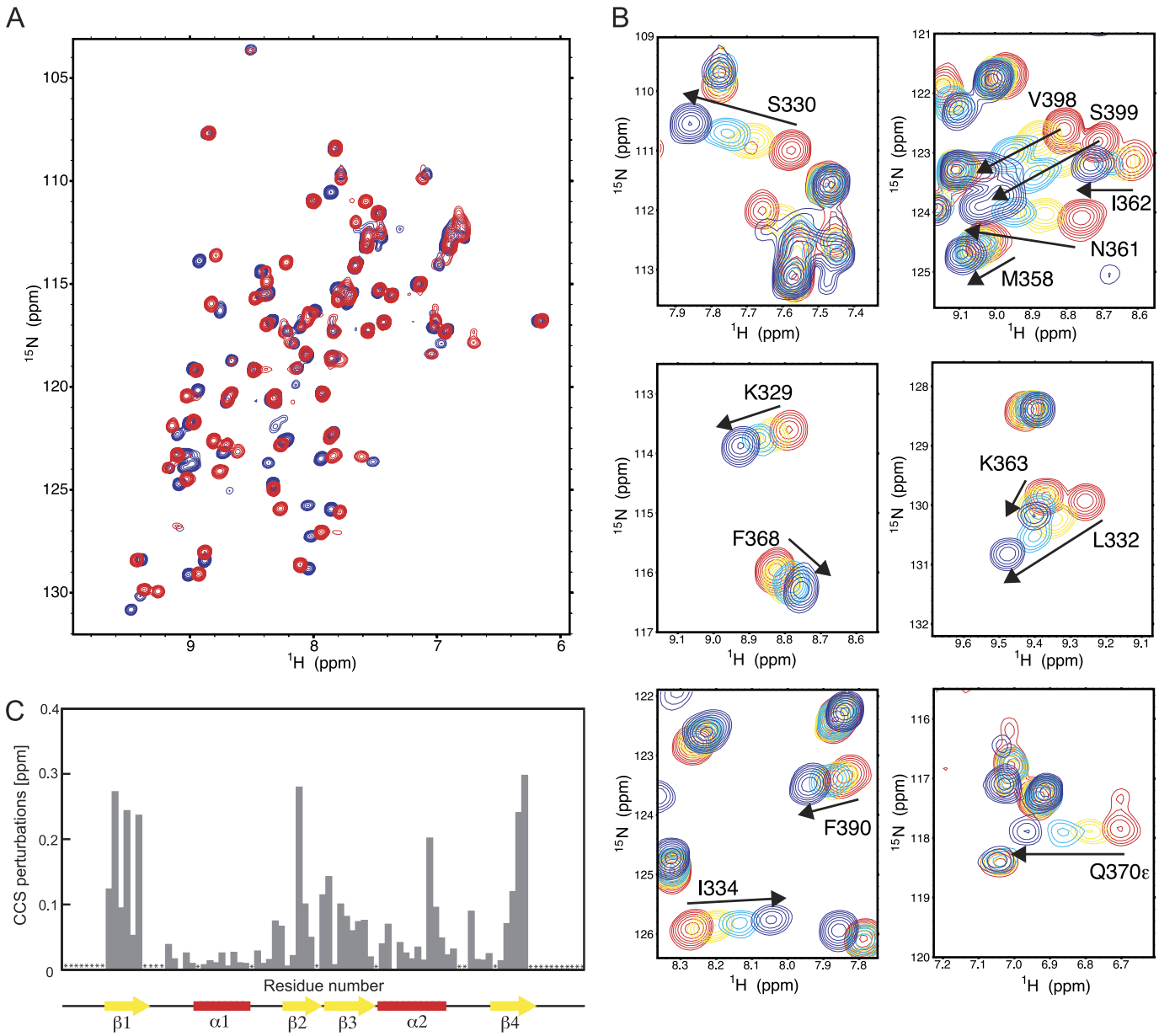


FIGURE 2. **NMR titration experiments of Nab3 RRM with UCUU RNA.** *A*, ^{1}H-^{15}N HSQC spectra of Nab3 RRM alone (in *red*) and in the presence of 1 eq of 5′-UCUU-3′ (in *blue*) at 303 K. *B*, close-up views of the ^{1}H-^{15}N HSQC spectra, showing selected chemical shift changes during the titration. *C*, quantification of chemical shift perturbations of Nab3 RRM upon binding to UCUU RNA. The combined chemical shift perturbations $([\omega_{HN}\Delta\delta_{HN}]^2 + [\omega_N\Delta\delta_N]^2)^{1/2}$, where $\omega_{HN} = 1$ and $\omega_N = 0.154$ are weight factors of the nucleus (52), are plotted *versus* the amino acid residue number. Large changes occur on the β-sheet surface. The assignments of residues indicated by *asterisks* could not be obtained for neither the free nor bound protein, or indicates proline residues.

their free and bound forms relative to NMR time scale. The binding of UCUU to the RRM of Nab3 induces chemical shift perturbation of the residues shown in Fig. 2.

These chemical shift changes indicate that the above mentioned residues are involved in binding to the RNA, or alternatively, could undergo a conformational change upon RNA binding. Mapping the perturbed residues on the sequence of Nab3 RRM delineates that the Nab3 RRM binds the RNA through its β-sheet surface and also through the $\beta2\beta3$ loop (Fig. 2*C*).

Structure of Nab3 RRM in Complex with UCUU—When solving the structure of Nab3 RRM bound to RNA, we extensively tested different lengths of RNA, buffer conditions, and temperatures with the aim to optimize the NMR spectral quality of the complexes. Longer RNA substrates, a UCUU core motif with flanking sequences, resulted in the significant broadening of NMR signals of the complexes. Interestingly, we obtained the NMR spectra of better quality (for both protein and RNA in complex) with a four-nucleotide UCUU despite the fact that this RNA has lower affinity to Nab3 RRM compared with the longer substrates (see below). Similar improvement of the NMR spectral quality by using of a minimal specific RNA sequence has also been observed for other protein-RNA complexes investigated by NMR (47–49). Therefore, we pursued the structure determination of the Nab3 RRM-UCUU complex.

The RRM of Nab3 in complex with RNA display the canonical RRM-fold with an $\beta1\alpha1\beta2\beta3\alpha2\beta4$ topology and is similar to that of the unbound form (Figs. 1 and 3). Akin to the free

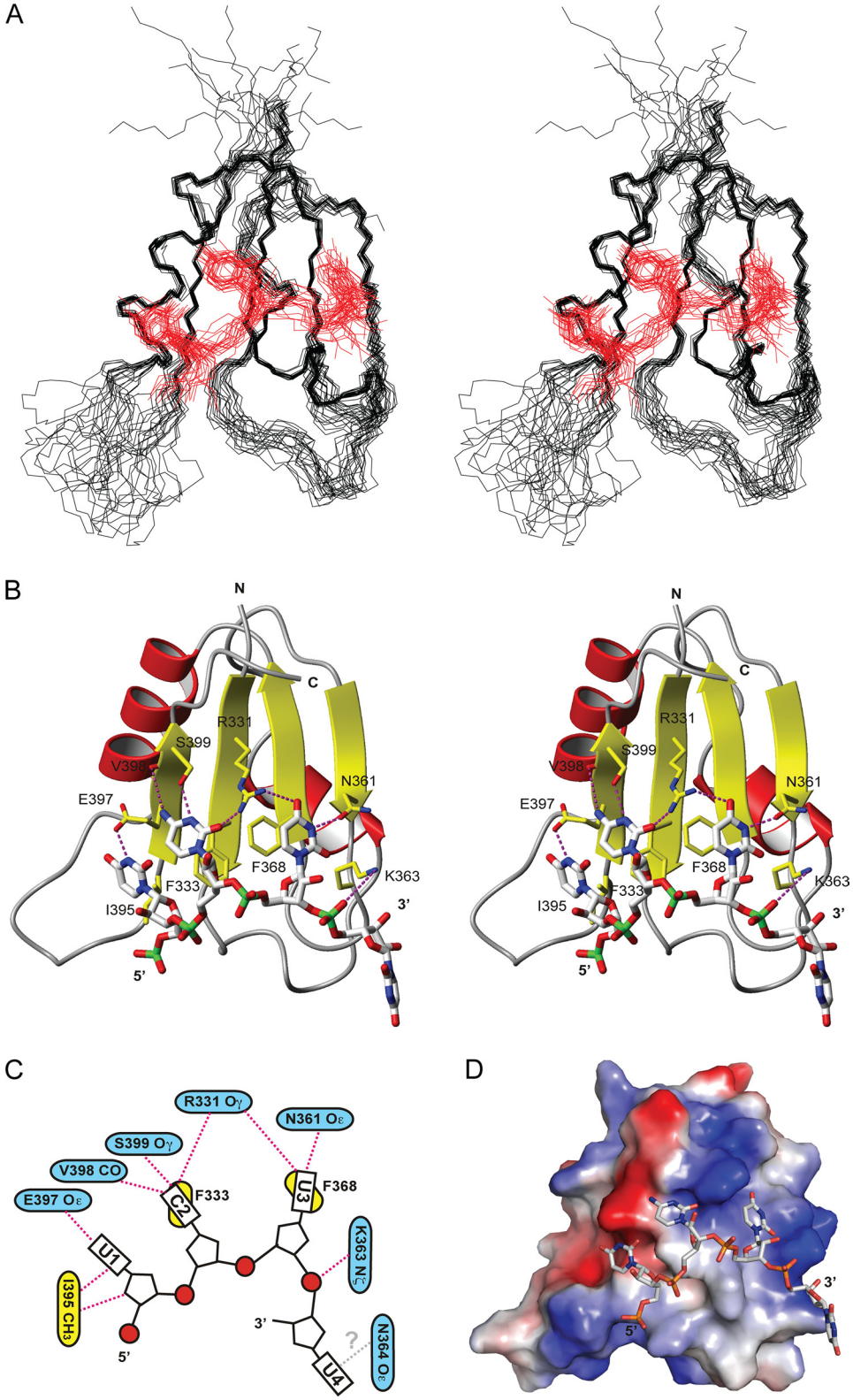


FIGURE 3. **Overview of the solution structure of the Nab3 RRM in complex with UCUU.** *A*, stereo view of the 20 lowest energy structures of the Nab3 RRM-UCUU complex. The protein backbone is shown as a wire model in *black*. The RNA heavy atoms are shown as a wire model in *red*. *B*, stereo view of the representative (the lowest energy) structure of the Nab3 RRM-UCUU complex. The RNA is represented as a *white stick model* and the protein is shown as a *ribbon* model with residues that contact the RNA shown in *yellow*. Putative hydrogen bonds are shown by *dotted magenta lines*. *C*, scheme showing contacts between Nab3 RRM and the UCUU RNA. Protein residues that form putative hydrogen bonds to the RNA are shown in *blue* and the one having hydrophobic interactions are in *yellow*. A hypothetic recognition of U4 is labeled by a *gray question mark*. *D*, solvent-accessible surface representation of Nab3 RRM colored by electrostatic potential (*blue*, positive; *red*, negative) and stick representation for the RNA of the representative structure of the complex. Figures were generated with MOLMOL (38).

form of Nab3 RRM, the N- and C-terminal regions as well as the long $\alpha 2\beta 4$ loop are structurally undefined due to the lack of experimental data. The UCUU RNA adopts a single-stranded conformation and the first three nucleotides are positioned over the whole β-sheet surface in a canonical arrangement in which the 5′ end is located on the first half of the β-sheet ($\beta 4\beta 1$) and the 3′ end on the second half ($\beta 3\beta 2$) (21, 22) (Fig. 2). The overall position of RNA on the β-sheet coincides with the perturbed residues from the titration experiment (Figs. 2 and 3). All bases have an *anti* conformation of the glycosidic bond and *C2′-endo* conformation of the sugar pucker.

The NMR spectra provided a limited number of intermolecular NOEs (11 unambiguous intermolecular NOEs) that loosely define the position of $U_1C_2U_3$ on the β-sheet surface of Nab3 RRM (Fig. 3*A*), but are sufficient to reveal the molecular basis of $U_1C_2U_3$ recognition by Nab3 RRM (Fig. 3, *B* and *C*; the protein-RNA hydrogen bonds described below are inferred from the final ensemble of structures and thus they should be considered as putative hydrogen bonds). We could not define the position of U_4 due to the lack of intermolecular NOEs. Based on NMR titration data, we speculate that U_4 could be recognized by the asparagine side chain or a main chain of the $\beta 2$–$\beta 3$ loop that are in proximity to the base of U_4, as displayed in the representative structure (Fig. 3*B*). In our NMR structure, C_2 and U_3 are involved in base stacking with the aromatic rings of Phe^{333} and Phe^{368}, respectively. The Watson-Crick edge of C_2 is recognized by the main chain carbonyl group of Val^{398} and the hydroxyl group of Ser^{399} that form hydrogen bonds with the amino and imino groups of C_2, respectively. One-half of the 20 structures in the final ensemble has Arg^{331} in a position in which it contacts the O_2 oxygen of C_2 (for the importance of Arg^{331} see below). In addition, it is likely also that Ser^{400} could be involved in the recognition of C_2 as the resonances of this residue broadened beyond detection upon RNA binding. Our structure also rules out the possibility that a purine could be accommodated in the C_2 position due to a steric restriction imposed by the Glu^{397} side chain.

The recognition of U_3 is mediated by the Arg^{331} and Asn^{361} side chains. The side chain NH_2 group of Arg^{331} contacts the O4 carbonyl functional group of the base and the side chain carbonyl group of Asn^{361} forms a hydrogen bond with the imino proton of U_3. Akin to C_2, the position of U_3 cannot be exchanged by a purine due to a steric hindrance of Arg^{331}. The sugar of the U_3 residue is further contacted by the aliphatic region of the Lys^{363} side chain.

In contrast to C_2 and U_3, the recognition of U_1 is less evident from the structure. A single hydrogen bond is formed between the imino proton of U_1 and the Oϵ of Glu^{397}. There are also hydrophobic contacts between the sugar and the base of U_1 and the side chain of Ile^{395}. However, these contacts do not explain fully the sequence specificity of a uridine nucleotide. A cytidine nucleotide in this position could also form a similar interaction with the glutamate. Altogether, our NMR structure indicates that Nab3 RRM recognizes the YCU sequence (where Y stands for pyrimidine).

Nab3 RRM Binds the Nab3 Termination Element with Low Affinity—The FA measurements were carried out to further characterize the binding of Nab3 RRM to various RNA substrates. In FA measurements, formation of the protein-RNA complex is monitored directly from an increase of the FA value that occurs when the protein binds fluorescently labeled RNA. Binding curves were recorded in the course of titration experiments, where protein aliquots were added to 10 nM fluorescently labeled RNAs (Fig. 4).

First, we assayed the binding affinity of Nab3 RRM to UCUU that has been reported as the minimal Nab3 terminator element. A tetranucleotide GUAA was used as a nonspecific control substrate. The comparison of the anisotropy data for specific and nonspecific four-nucleotide substrates along with determined equilibrium dissociation constants (K_d) are shown in Fig. 4*A*. Corresponding logarithmic values of K_a ($K_a = 1/K_d$) are shown in graph in Fig. 4*E*. Nab3 RRM binds the specific recognition sequence UCUU with more than 6-fold higher affinity compared with the nonspecific substrate GUAA.

As terminator elements often occur in multiple repeats, we tested a longer substrate with three UCUU repeats. The substrate sequence was derived from the snR47 that is terminated by the Nrd1 pathway (Fig. 4*F*). Nab3 RRM binds the three UCUU-containing substrate with a K_d of 48 ± 2 μM, 1 order of magnitude stronger than we observed for a single UCUU motif (Fig. 4*B*). Furthermore, we assayed snR13, which is another naturally occurring Nrd1-dependent terminator that contains two UCUU and one CCU motifs (Fig. 4*F*). As expected, Nab3 RRM binds this substrate with a K_d of 46 ± 1 μM, a similar binding affinity to that of snR47 (Fig. 4*B*). As a control, we used SL RNA, which has a similar size and its sequence lacks UCUU or even CU recognition motifs (Fig. 4*F*). For this nonspecific substrate, Nab3 RRM binding is reduced more than 30-fold compared with the snR13 and snR47 terminators (Fig. 4, *B* and *E*).

Nrd1 RRM Binds the Nrd1 Termination Element with Low Affinity—As Nab3 RRM binds its termination motif with a low affinity, we therefore decided to investigate the RNA-binding properties of Nrd1 RRM, which is the second RNA-binding domain occurring in the Nrd1-dependent termination complex (Fig. 1*B*). Akin to Nab3 RRM, we assayed the binding affinity of Nrd1 RRM to RNA using FA. We found that Nrd1 RRM-(340–410) binds to the minimal termination RNA sequence GUAA (GUAR is known as the Nrd1-termination element (where R stands for purine) (18) with a low affinity in the mid-micromolar range (Fig. 4*C*; K_d of 66 ± 1 μM). Next, we assayed a longer RNA substrate, snR13, which contains two GUAR motifs (Fig. 4*F*). The titration curve for snR13 yielded a K_d of 11 ± 1 μM (Fig. 4*C*).

A Complex of Nrd1-Nab3 Binds RNA with High Affinity—As the Nrd1 and Nab3 proteins form a heterodimer *in vivo* and *in vitro* (19), we assayed RNA binding of the co-expressed Nrd1-Nab3 heterodimer (19) using FA. As a substrate, we used RNA derived from the snR13 terminator that contains two copies of each recognition element (Fig. 4*F*). This RNA substrate binds the Nrd1-Nab3 heterodimer with a K_d of 2.4 ± 0.3 nM (Fig. 4*D*), about 4–5 orders of magnitude stron-

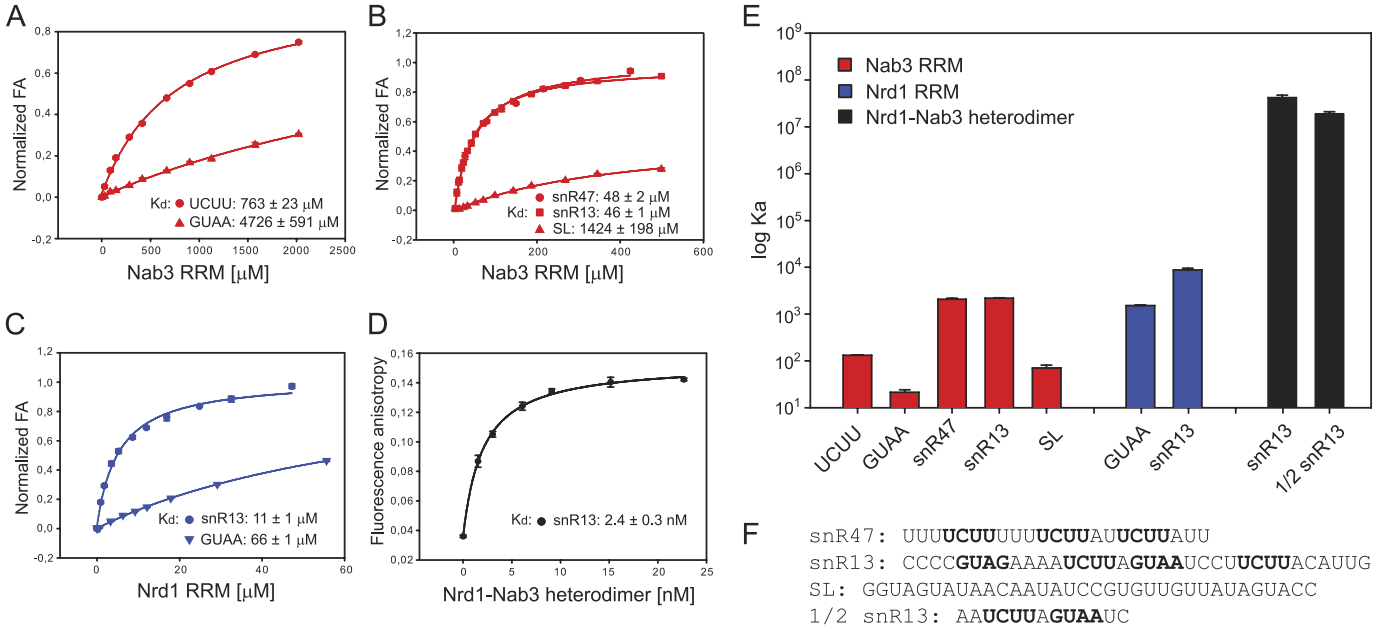


FIGURE 4. **Equilibrium binding of Nab3 RRM, Nrd1 RRM, and Nrd1-Nab3 heterodimer with fluorescently labeled RNA monitored by fluorescence anisotropy.** *A*, Nab3 RRM was titrated with UCUU and GUAA (each 10 nM), and their binding isotherms are shown as *red circles* and *triangles*, respectively. *B*, Nab3 RRM was titrated with snR47, snR13, and SL substrates (each 10 nM), and their binding isotherms are shown as *red circles*, *squares*, and *triangles*, respectively. *C*, Nrd1 RRM was titrated with GUAA and snR13 (each 10 nM) and their binding isotherms are shown as *blue inverted triangles* and *circles*, respectively. *D*, Nrd1-Nab3 heterodimer was titrated with snR13 (100 pM). *E*, summary of the association constants (K_a) for the RRM of Nrd1 (in *blue*) and Nab3 (in *red*) in their free forms as well as for the Nrd1-Nab3 heterodimer (in *black*). Logarithmic scale of K_a is shown to cover a wide range of affinities. *F*, RNA sequences used in the affinity measurements. The buffers contained the same ion strength and pH values for all proteins. Equilibrium dissociation constant (K_d) was calculated from the best fit to the data using a single-site binding isotherm. *Error* is denoted as S.E. The data were normalized for visualization purposes (*A–C*).

ger compared with the individual RRMs of Nrd1 and Nab3 (Fig. 4*E*). Another snR13-derived substrate, ½ snR13, containing only one copy of each recognition element (Fig. 4*F*), binds the Nrd1-Nab3 heterodimer with only a slightly lower affinity (K_d of 5.3 ± 0.6 nM). However, when one of the recognition elements (either Nrd1 or Nab3) is removed from the ½ snR13 substrate, the affinity is reduced such that the dissociation constant cannot be determined (saturation of the binding curve cannot not be reached due to the low solubility limit of the Nrd1-Nab3 heterodimer; data not shown).

Impacts of Nab3 RRM Point Mutations on RNA Binding Affinity—It has been shown previously that mutations of C_2 and U_3 in the UCUU motif reduce the binding affinity to Nab3 (18) or to the Nrd1-Nab3 heterodimer (19). We performed the converse experiments in which we assessed Nab3 RRM mutants for their ability to bind snR47 RNA in a quantitative solution binding assay by fluorescence anisotropy titration experiments. We mutated the non-canonical amino acid residues on the β-sheet surface (R331A, N361A, S399A, and E397A) that specifically recognize the bases of the $U_1C_2U_3$ sequence (Fig. 3, *B* and *C*). These residues surround the conserved residues of RNP1 and RNP2 consensuses. We found that mutants R331A and S399A showed a 3–4-fold decrease in binding affinity of that demonstrated by the wild-type protein (Fig. 5*A*). In contrast, mutants N361A and E397A showed binding affinity similar to the wild-type Nab3 RRM (Fig. 5*A*).

Functional Significance of the Nab3 RRM Residues That Contact RNA—To address the importance of the specific contacts identified in the Nab3 RRM-UCUU complex for Nab3 function *in vivo*, single amino acid mutants (R331A, S399A, S399K, E397A, and E397K) were prepared in a yeast expression vector and introduced into a yeast strain in which the endogenous NAB3 promoter was replaced with the GAL1 promoter (3). To test whether the mutated residues were essential for growth, the resulting transformants were spotted onto glucose containing plates. The shift to glucose represses the expression of the GAL1-driven endogenous NAB3, which completely impairs cell viability (Fig. 5*B*). This lethality was rescued by wild-type Nab3 (Fig. 5*B*). Mutating the three residues involved in the recognition of C_2 and U_3 (Fig. 3), R331A, N361A and S399A, or S399K, caused lethality (Fig. 5*B*), providing further support for the functional significance of these contacts. In contrary, the Glu^{397} mutant displayed only slow growth at 24 and 37 °C.

To further confirm that mutant lethality was not due to impaired protein expression, we performed Western blot analysis with antibodies directed against Nab3 and antibodies against the HA tag present on the endogenous Nab3 only. We detected similar levels of Nab3 in all proteins expressed episomally, whereas no HA-tagged endogenous Nab3 was detected in yeast grown on glucose medium (Fig. 5*C*).

DISCUSSION

RNA Recognition by Nab3 RRM and Its Comparison with Other RRMs—We have solved the structure of Nab3 RRM in free form and in complex with UCUU. In both forms, the fold of Nab3 RRM is very similar, with minor adjustments upon RNA binding, and resembles canonical RRM (21–23). However, it has shorter α-helices than observed in the structures

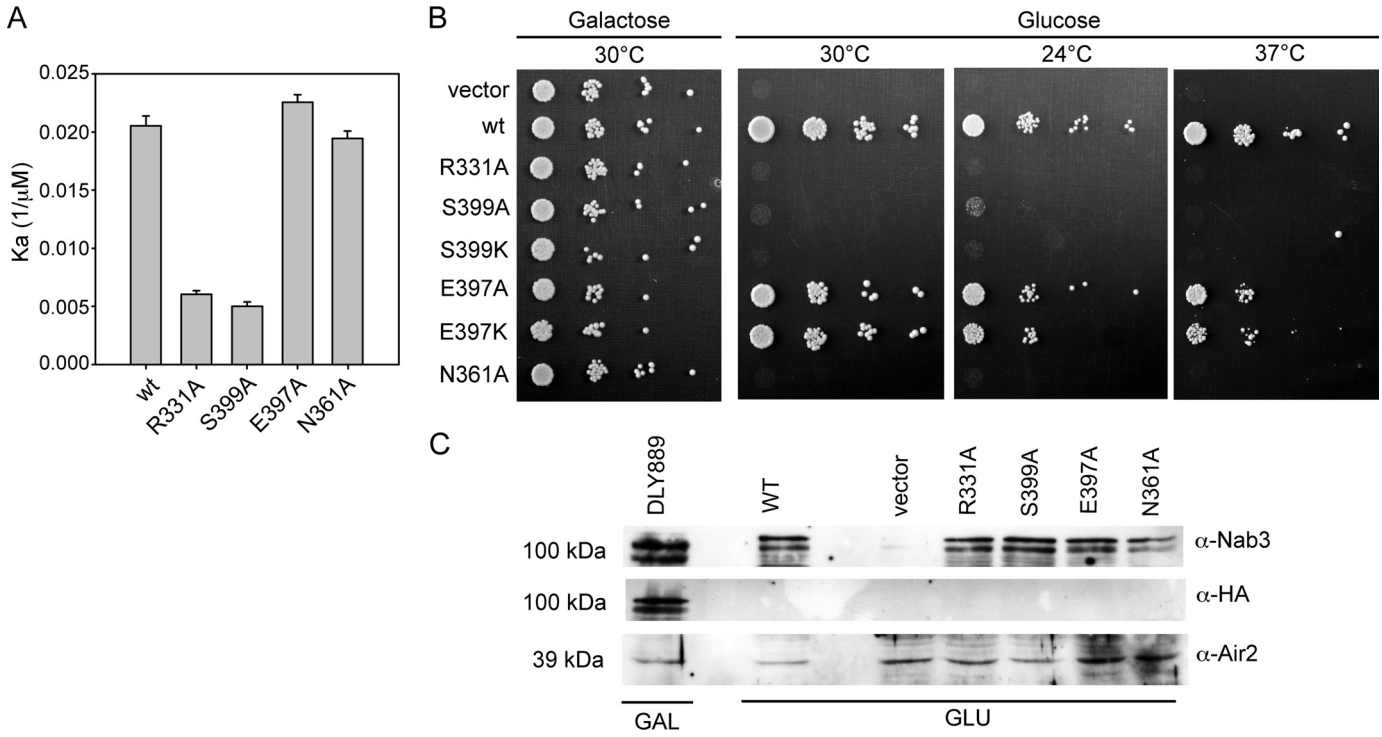


FIGURE 5. **The important residues of Nab3 RRM that are required for RNA binding and cell viability.** *A*, equilibrium binding of Nab3 RRM mutants with fluorescently labeled RNA monitored by fluorescence anisotropy. The Nab3 RRM R331A, N361A, S399A, and E397A mutants along with the wild-type of Nab3 RRM were titrated with fluorescently labeled snR47 substrate. Equilibrium association constants (K_a) are shown for individual mutants with S.E. *B*, residues Arg^{331}, Ser^{399}, and Asn^{361} are required for yeast viability. The indicated Nab3 RRM mutants were expressed episomally from pRS415 plasmids in the yeast strain with the endogenous NAB3 driven by the GAL1 promoter. Mutant strains were spotted on plates containing 2% glucose and a control galactose plate and incubated for 3 days at the indicated temperatures. Growth on glucose-containing plates leads to the repression of GAL1-driven wild-type Nab3, and thus shows the functionality of the different Nab3 mutants. *Vector* is a control where the GAL1::NAB3 strain contains an empty pRS415 plasmid, *wt* is the wild-type NAB3. *C*, expression of Nab3 proteins from pRS415 in glucose-containing medium. Western blot analysis was perfomed with protein extracts from the original GAL1::NAB3 strain (DLY889) grown in galactose-containing medium and extracts from DLY889 transformed with plasmids carrying wild-type and mutant NAB3 grown for 20 h in glucose-containing medium. *Air2* was used as a loading control.

of canonical RRMs, and an extra 3_{10} helix between the α1 helix and β2 strand (22).

Nab3 RRM binds UCUU in a canonical manner as observed in other RRM-RNA complexes (21–23). Specifically, Nab3 RRM contains a well conserved signature of the RRM family, RNP1 and RNP2 sequences (44–46). Two conserved phenylalanine residues are used in RNP1 (Phe^{333}) and RNP2 (Phe^{368}) to mediate the stacking interaction with RNA bases C_2 and U_3, respectively. These aromatic residues are surrounded by basic and polar amino acid residues that mediate the sequence-specific recognition of C_2U_3 and in part of U_1. The specifically recognized nucleotides are accommodated on the β-sheet only and neither loops nor N-/C-terminal regions to the RRM are involved in the recognition process.

Interestingly, the binding preference for CU has also been reported for the polypyrimidine tract-binding protein (PTB) (49). Unlike Nab3 RRM, PTB RRMs display additional topology elements to the canonical RRM involved in RNA recognition and significantly differ from RNP1 and RNP2 consensus (Fig. 6*A*). In particular, PTB RRMs lack aromatic residues in RNP1 and RNP2 that usually make extensive stacking interactions with the RNA bases and sugars as in the case of Nab3 RRM (Fig. 6). In addition, the structures of Nab3 and PTB RRMs revealed that these domains are sequentially unrelated on the entire RNA interaction surface, except for the last serine residue of β-strand 4 that is present in all PTB and Nab3 RRMs (Fig. 6*A*). The serine residue is involved in recognition of a cytosine in the structures of PTB (49) and Nab3 bound to RNA (Fig. 6). Furthermore, in the structures of PTB RRM1-RNA and Nab3 RRM-RNA, recognition of the 5′-end uridine is mediated in a similar way in which its imino proton is contacted by glutamine and glutamate, respectively (Fig. 6). In both structures, a cytidine could be tolerated instead of a uridine in this position. In contrast, recognition of the 3′-end uridine is mediated differently in these two structures. Whereas Nab3 RRM utilizes the side chains of arginine (in β1) and asparagine (in β2) to recognize the uridine, PTB RRM1 uses the main chain of leucine and lysine located in the C-terminal extension to the RRM to facilitate the uridine binding.

The importance of the serine residue, which is discussed above, is demonstrated by our affinity measurements with the Nab3 RRM S399A mutant that decreases the binding affinity to its UCUU-containing substrate (Fig. 5*A*). Correspondingly, *in vivo* analyses of Nab3 S399A or S399K mutants show that the serine residue is essential for yeast viability (Fig. 5*B*). Similarly, the decreased affinity of the Nab3 RRM R331A mutant to the RNA is manifested by the lethal effect in yeast (Fig. 5). Interestingly, Asn^{361}, which specifically recognizes U_3 along with Arg^{331}, shows lethality *in vivo* for the N361A mutant, whereas the affinity of this mutant is only slightly reduced compared with the wild-type (Fig. 5). Furthermore, the

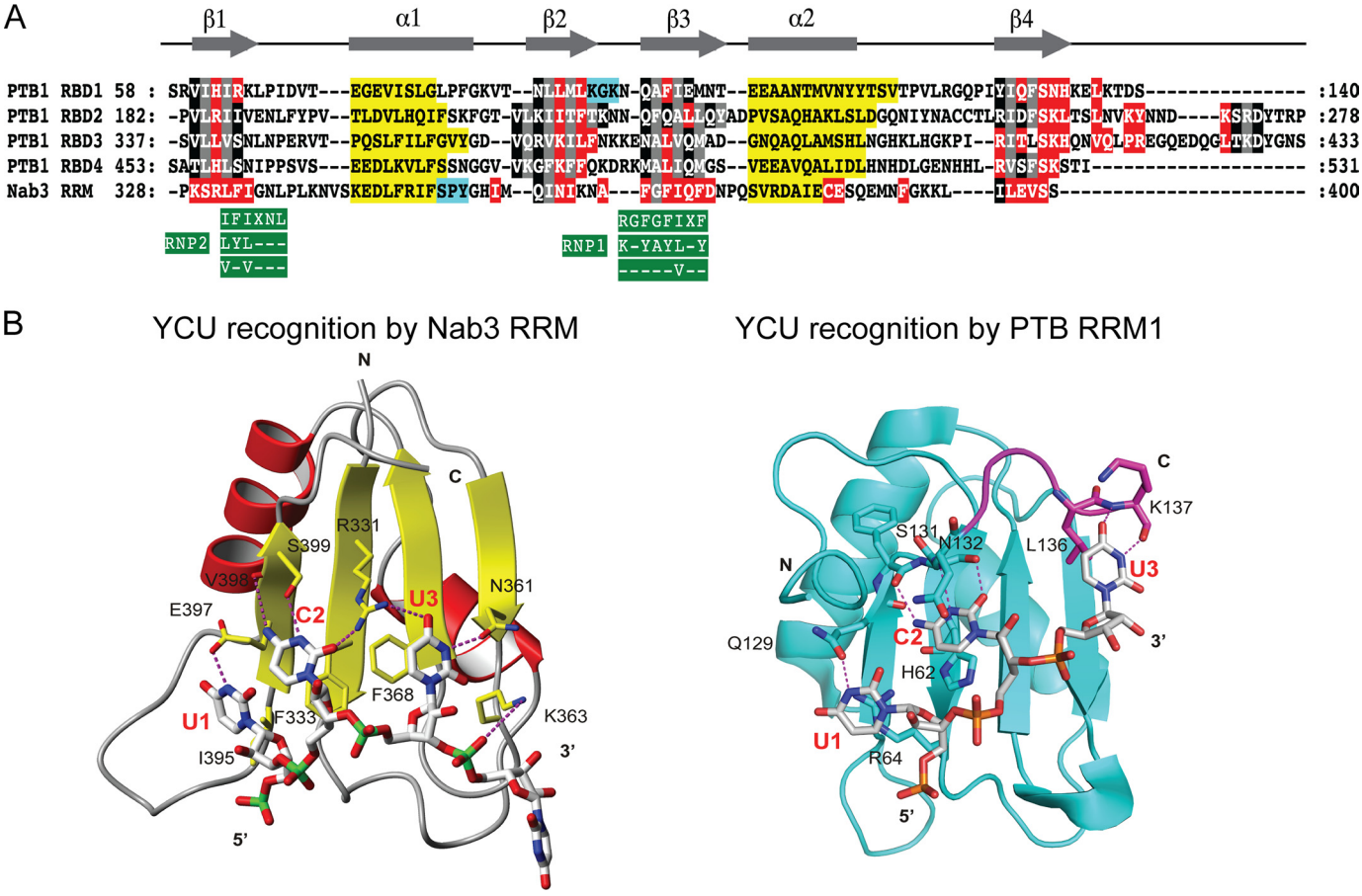


FIGURE 6. **Recognition of YCU by Nab3 and PTB RRMs.** *A*, sequence alignment of PTB RRM1, -2, -3, and -4 and Nab3 RRM whose structures have been solved. The alignment was performed using ClustalW (53) and manually optimized using the three-dimensional structural information (49). For the RRMs of PTB, amino acids interacting with the RNA are shown in *red boxes*, residues in *gray* and *black boxes* are located in the β-sheet and residues in *yellow* and *cyan boxes* are in the α- or 3_{10} helices, respectively (49). Residues in *gray boxes* form the hydrophobic core of the domains. For the RRM of Nab3, residues in *red boxes* are significantly perturbed upon RNA binding. *B*, comparison of Nab3 RRM (*left*: in *yellow* and *red schematics*) and PTB RRM1 (*right*: in *cyan schematics*) binding to UCU nucleotides (represented as a *stick model*). The protein residues that mediate the specific recognition are highlighted as a stick model. The C-terminal region of PTB RRM1 that mediates the recognition of U_3 is shown in *magenta*.

N361A mutant yeast strain shows accumulation of cryptic unstable transcripts when shifted to glucose,[7] supporting the functional significance of this contact in the termination via the Nrd1 pathway. The Glu^{397} mutant does not have an impact on RNA binding *in vitro* and causes only slow growth at higher temperatures, indicating a minor role of this contact for the function of Nab3. Altogether, these functional data corroborate with our structural findings that Nab3 RRM specifically recognizes the YCU sequence.

Association between Nab3 and Nrd1 Increases the Affinity to the Termination Sequences—Our FA experiments showed that Nab3 RRM binds UCUU with the high micromolar range of equilibrium dissociation constant (K_d), yet the Nab3 RRM binds nonspecific four-nucleotide RNA with a K_d in the low millimolar range (Fig. 4). The observed affinity for the specific RNA substrate is weaker than the affinity usually observed for single canonical RRMs (21, 47). The affinity of Nab3 RRM binding to UCUU is lower by 2-fold compared with the affinity of PTB RRM1 to CUCU that also recognizes the YCU motif (Fig. 6) (47). Interestingly, the apparent K_d of Nab3 RRM binding to longer RNAs with multiple UCUU motifs (snR47 and snR13) are ~50 μM, more than 1 order of magnitude stronger than we observed for a single UCUU motif (Fig. 4). The increased binding affinity likely originates from the presence of multiple binding sites and due to the presence of flanking sequences to the UCUU motif in snR47 and snR13 RNA substrates. The effect of flanking sequences is likely nonspecific, mediated by electrostatic interactions of additional phosphate groups. A similar increase in the binding affinity, when multiple binding motifs are present, has also been observed for the RRMs of PTB (47). The formation of multiple complexes between the snR47-derived RNA substrate and an RRM-containing Nab3 construct (Nab3-(277–565)) has also been observed previously using electrophoretic mobility shift assay (18). It has been also demonstrated that mutations in the 2nd, 3rd, and 4th position of the UCUU motif decrease the binding affinity to the Nab3-(277–565) (18). Furthermore, our data showed that Nrd1, the second RNA-binding subunit of the Nrd1 complex containing a single RRM, binds its termination element, GUAA, with a low affinity (K_d of 66 ± 1 μM; Fig. 4).

Therefore, it is very likely that Nrd1 and Nab3 bind RNA in a cooperative manner to achieve the nanomolar range of af-

[7] F. Hobor, D. Hrossova, S. Vanacova, and R. Stefl, unpublished data.

finity to snR13 RNA, as previously estimated using electrophoretic mobility shift assay (EMSA) (19). Indeed, our FA measurements showed that the Nrd1-Nab3 heterodimer binds snR13 RNA with a K_d in the low nanomolar range, 4–5 orders of magnitude stronger compared with the individual RRMs of Nrd1 and Nab3 (Fig. 4*E*). The RRMs of Nrd1 and Nab3 do not bind each other (data not shown) and the regions that mediate the formation of the heterodimer are located near the N-terminal of each RRM (13).[7] It is difficult to assay the binding affinity of full-length Nrd1 and Nab3 individually due to their instability (19); the Nrd1 alone rapidly aggregates.[7] However, other regions outside of the RRMs of Nrd1 and Nab3 are not expected to contribute significantly to RNA binding as they do not contain an identifiable RNA-binding domain. In a similar way, the cooperative RNA binding of two RRM-containing proteins $U2AF^{65}$ and $U2AF^{35}$ is utilized to enhance the affinity and selectivity in the process of defining the site of spliceosomal assembly (50, 51).

Implication for Poly(A) Independent Transcription Termination—We have shown that the RRM of Nab3 binds specifically the YCU sequence (where Y stands for pyrimidine). This is in good agreement with previous functional data that led to the proposal of the UCUU sequence as the Nab3 termination element (6, 18). However, the first position of the UCUU motif is not fully conserved in some snRNA downstream sequences (18); it can be either U or C that perfectly matches our structural findings. Furthermore, the last position of UCUU is also not fully conserved but we cannot explain the recognition of another nucleotide due to the lack of experimental data for the recognition of this nucleotide in our structure. Considering the specificity only for the CU dinucleotide and relatively weak affinity of Nab3 RRM to YCU-like sequences, it is evident that Nab3 alone cannot recruit the Nrd1 complex to the correct termination sites. Indeed, we show the first quantitative evidence that the association of Nrd1 and Nab3 (each protein contains a single RRM) facilitates high affinity binding and sequence selectivity. It remains to be seen whether additional sequence elements are recognized, in addition to YCU and GUAR, upon the association of Nrd1-Nab3.

Acknowledgments—We thank to Dr. D. Libri for the GAL1::Nab3 strain and Dr. J. L. Corden for the generous gift of the Nab3 2F12 antibodies and the pST39 plasmid, containing Nab3-(191–565) and Nrd1-(1–548). We are grateful to Massimo Lucci for assistance with NMR measurements at CERM Florence, Italy. The NOESY spectra including the intermolecular filter experiments were obtained at the CERM NMR facility supported by European Union-NMR program Grant RII3–026145.

REFERENCES

1. Richard, P., and Manley, J. L. (2009) *Genes Dev.* **23,** 1247–1269
2. Bentley, D. (2002) *Curr. Opin. Cell Biol.* **14,** 336–342
3. Thiebaut, M., Kisseleva-Romanova, E., Rougemaille, M., Boulay, J., and Libri, D. (2006) *Mol. Cell* **23,** 853–864
4. Arigo, J. T., Eyler, D. E., Carroll, K. L., and Corden, J. L. (2006) *Mol. Cell* **23,** 841–851
5. Steinmetz, E. J., and Brow, D. A. (1996) *Mol. Cell. Biol.* **16,** 6993–7003
6. Steinmetz, E. J., Conrad, N. K., Brow, D. A., and Corden, J. L. (2001) *Nature* **413,** 327–331
7. Vasiljeva, L., and Buratowski, S. (2006) *Mol. Cell* **21,** 239–248
8. Mitchell, P., Petfalski, E., Shevchenko, A., Mann, M., and Tollervey, D. (1997) *Cell* **91,** 457–466
9. Vanácová, S., Wolf, J., Martin, G., Blank, D., Dettwiler, S., Friedlein, A., Langen, H., Keith, G., and Keller, W. (2005) *PLoS Biol.* **3,** e189
10. LaCava, J., Houseley, J., Saveanu, C., Petfalski, E., Thompson, E., Jacquier, A., and Tollervey, D. (2005) *Cell* **121,** 713–724
11. Wyers, F., Rougemaille, M., Badis, G., Rousselle, J. C., Dufour, M. E., Boulay, J., Régnault, B., Devaux, F., Namane, A., Séraphin, B., Libri, D., and Jacquier, A. (2005) *Cell* **121,** 725–737
12. Proudfoot, N. J., Furger, A., and Dye, M. J. (2002) *Cell* **108,** 501–512
13. Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S., and Meinhart, A. (2008) *Nat. Struct. Mol. Biol.* **15,** 795–804
14. Gudipati, R. K., Villa, T., Boulay, J., and Libri, D. (2008) *Nat. Struct. Mol. Biol.* **15,** 786–794
15. Steinmetz, E. J., and Brow, D. A. (1998) *Proc. Natl. Acad. Sci. U.S.A.* **95,** 6699–6704
16. Conrad, N. K., Wilson, S. M., Steinmetz, E. J., Patturajan, M., Brow, D. A., Swanson, M. S., and Corden, J. L. (2000) *Genetics* **154,** 557–571
17. Morlando, M., Greco, P., Dichtl, B., Fatica, A., Keller, W., and Bozzoni, I. (2002) *Mol. Cell Biol.* **22,** 1379–1389
18. Carroll, K. L., Pradhan, D. A., Granek, J. A., Clarke, N. D., and Corden, J. L. (2004) *Mol. Cell Biol.* **24,** 6241–6252
19. Carroll, K. L., Ghirlando, R., Ames, J. M., and Corden, J. L. (2007) *RNA* **13,** 361–373
20. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D.,

Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001) *Science* **291,** 1304–1351
21. Maris, C., Dominguez, C., and Allain, F. H. (2005) *FEBS J.* **272,** 2118–2131
22. Stefl, R., Skrisovska, L., and Allain, F. H. (2005) *EMBO Rep.* **6,** 33–38
23. Cléry, A., Blatter, M., and Allain, F. H. (2008) *Curr. Opin. Struct. Biol.* **18,** 290–298
24. Pergoli, R., Kubicek, K., Hobor, F., Pasulka, J., and Stefl, R. (2010) *Biomol. NMR Assign.* **4,** 119–121
25. Bax, A., and Grzesiek, S. (1993) *Acc. Chem. Res.* **26,** 131–138
26. Sattler, M., Schleucher, J., and Griesinger, C. (1999) *Prog. Nucl. Magn. Reson. Spectrosc.* **34,** 93–158
27. Varani, G., Aboulela, F., and Allain, F. H. (1996) *Prog. Nucl. Magn. Reson. Spectrosc.* **29,** 51–127
28. Peterson, R. D., Theimer, C. A., Wu, H., and Feigon, J. (2004) *J. Biomol. NMR* **28,** 59–67
29. Zwahlen, C., Legault, P., Vincent, S. J., Greenblatt, J., Konrat, R., and Kay, L. E. (1997) *J. Am. Chem. Soc.* **119,** 6711–6721
30. Güntert, P. (2004) *Methods Mol. Biol.* **278,** 353–378
31. Herrmann, T., Güntert, P., and Wüthrich, K. (2002) *J. Mol. Biol.* **319,** 209–227
32. Shen, Y., Delaglio, F., Cornilescu, G., and Bax, A. (2009) *J. Biomol. NMR* **44,** 213–223
33. Case, D. A., Darden, T. A., Cheatham, T. E., III, Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Crowley, M., Walker, R. C., Zhang, W., Merz, K. M., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossváry, I., Wong, K. F., Paesani, F., Vanicek, J., Wu, X., Brozell, S. R., Steinbrecher, T., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Mathews, D. H., Seetin, M. G., Sagui, C., Babin, V., and Kollman, P. A. (2008) *Amber 10*, University of California, San Francisco, CA
34. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995) *J. Am. Chem. Soc.* **117,** 5179–5197
35. Padrta, P., Stefl, R., Králík, L., Zídek, L., and Sklenár, V. (2002) *J. Biomol. NMR* **24,** 1–14
36. Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) *J. Biomol. NMR* **8,** 477–486
37. Vriend, G. (1990) *J. Mol. Graph.* **8,** 52–56, 29
38. Koradi, R., Billeter, M., and Wuthrich, K. (1996) *J. Mol. Graph.* **14,** 51–55, 29–32
39. Heyduk, T., and Lee, J. C. (1990) *Proc. Natl. Acad. Sci. U.S.A.* **87,** 1744–1748
40. Sikorski, R. S., and Hieter, P. (1989) *Genetics* **122,** 19–27
41. Wilson, S. M., Datar, K. V., Paddy, M. R., Swedlow, J. R., and Swanson, M. S. (1994) *J. Cell Biol.* **127,** 1173–1184
42. San Paolo, S., Vanacova, S., Schenk, L., Scherrer, T., Blank, D., Keller, W., and Gerber, A. P. (2009) *PLoS Genet.* **5,** e1000555
43. Güntert, P., Mumenthaler, C., and Wüthrich, K. (1997) *J. Mol. Biol.* **273,** 283–298
44. Swanson, M. S., Nakagawa, T. Y., LeVan, K., and Dreyfuss, G. (1987) *Mol. Cell. Biol.* **7,** 1731–1739
45. Adam, S. A., Nakagawa, T., Swanson, M. S., Woodruff, T. K., and Dreyfuss, G. (1986) *Mol. Cell. Biol.* **6,** 2932–2943
46. Dreyfuss, G., Swanson, M. S., and Piñol-Roma, S. (1988) *Trends Biochem. Sci.* **13,** 86–91
47. Auweter, S. D., Oberstrass, F. C., and Allain, F. H. (2007) *J. Mol. Biol.* **367,** 174–186
48. Hargous, Y., Hautbergue, G. M., Tintaru, A. M., Skrisovska, L., Golovanov, A. P., Stevenin, J., Lian, L. Y., Wilson, S. A., and Allain, F. H. (2006) *EMBO J.* **25,** 5126–5137
49. Oberstrass, F. C., Auweter, S. D., Erat, M., Hargous, Y., Henning, A., Wenter, P., Reymond, L., Amir-Ahmady, B., Pitsch, S., Black, D. L., and Allain, F. H. (2005) *Science* **309,** 2054–2057
50. Kielkopf, C. L., Rodionova, N. A., Green, M. R., and Burley, S. K. (2001) *Cell* **106,** 595–605
51. Singh, R., Valcárcel, J., and Green, M. R. (1995) *Science* **268,** 1173–1176
52. Mulder, F. A., Schipper, D., Bott, R., and Boelens, R. (1999) *J. Mol. Biol.* **292,** 111–123
53. Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003) *Nucleic Acids Res.* **31,** 3497–3500
54. Spronk, C. A. E. M., Nabuurs, S. B., Krieger, E., Vriend, G., and Vuister, G. W. (2004) *Prog. Nucleic Magn. Res. Spectro.* **45,** 315–337
55. Hooft, R. W., Sander, C., and Vriend, G. (1997) *Comput. Appl. Biosci.* **13,** 425–430
56. Trantírek, L., Stefl, R., Masse, J. E., Feigon, J., and Sklenár, V. (2002) *J. Biomol. NMR* **23,** 1–12
57. DeLano, W. L. (2002) *The PyMOL Molecular Graphics System,* DeLano Scientific, Palo Alto, CA

Biomol NMR Assign (2011) 5:203–205
DOI 10.1007/s12104-011-9300-y

ARTICLE

^{1}H, ^{13}C, and ^{15}N resonance assignments for the CTD-interacting domain of Nrd1 bound to Ser5-phosphorylated CTD of RNA polymerase II

Karel Kubíček · Josef Pasulka · Hana Černá · Frank Löhr · Richard Štefl

Received: 11 November 2010 / Accepted: 16 February 2011 / Published online: 26 February 2011

Abstract In this article, we report the resonance assignment of CTD-interacting domain (CID) of pre-mRNA down-regulation (Nrd)1 bound to Ser5-phosphorylated CTD (pSer5) of RNA Polymerase II. The presented assignment of backbone and side-chain resonances of the Nrd1 CID proton, carbon and nitrogen nuclei will allow studies of the structure and interaction of CID with carboxy-terminal domain (CTD) of the RNA polymerase II.

Keywords CTD-interacting domain (CID) · Carboxy-terminal domain (CTD) · RNA-processing · Transcription termination

Biological context

In yeast, termination of sno/snRNA and other functional RNAs is dependent on the nuclear pre-mRNA down-regulation (Nrd)1 complex, consisting of two RNA-binding proteins (Nrd1 and the nuclear polyadenylated RNA-binding (Nab)3), and a putative RNA helicase Sen1. Nrd1 complex interacts with carboxy-terminal domain (CTD) of RNA polymerase II (Pol II) through the CTD-interacting domain (CID) of Nrd1 (Richard and Manley 2009). Pol II CTD contains 26 repeats of the sequence YSPTSPS in yeast and is dynamically phosphorylated and dephosphorylated at serines in position 2, 5, and 7 (Richard and Manley 2009). Different phosphorylation status creates so called CTD-code that controls binding of processing factors at various stages of the transcription (Egloff and Murphy 2008). Nrd1 CID preferentially binds pSer5 CTD (Vasiljeva et al. 2008). Interestingly, pSer5 CTD coincides with early phase of elongation (Buratowski 2009).

Recently, the structure of the Nrd1 CID was determined using X-ray crystallography in the free form (Vasiljeva et al. 2008). The structure shows a similar fold of that of Pcf11 CID whose structure has been also derived from X-ray data (Meinhart and Cramer 2004). However, there are structural differences between these structures such as an insertion located within the loop region between helices 1 and 2, an expanded region in helix 4, which is extended by an additional fifth helical turn. It is very likely that some of these regions of the Nrd1 CID could be used to alter binding preferences to pSer5 instead of pSer2 or pSer7 CTD.

To provide a structural insight on how the Nrd1 CID recognizes pSer5 CTD and to explore the Nrd1 termination pathway, we have initiated an NMR study of the Nrd1 CID protein bound to pSer5 CTD peptide. Here, we report ^{13}C, ^{15}N and ^{1}H resonance assignment of Nrd1 CID in complex with pSer5 CTD.

Methods and experiments

The DNA encoding the *Saccharomyces cerevisiae* Nrd1 CID domain (residues 1–153) was amplified and cloned into a pET22b expression vector (Novagen) via *NdeI* and

K. Kubíček (✉) · J. Pasulka · H. Černá · R. Štefl (✉)
National Centre for Biomolecular Research, Central European Institute of Technology, Faculty of Science, Masaryk University, 62500 Brno, Czech Republic
e-mail: karelk@chemi.muni.cz

R. Štefl
e-mail: stefl@chemi.muni.cz

F. Löhr
Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, Goethe-University, 60438 Frankfurt am Main, Germany

XhoI restriction sites. The protein was overexpressed in *Escherichia coli* BL21 (DE3) strain overnight at 16 °C in M9 minimal medium. For isotope labelling, the medium was supplemented with $^{15}NH_4Cl$ and U–$^{13}C_6$-glucose. Cells were grown at 37 °C to $OD_{600} \sim 0.6$ and induced with 66 µM Isopropyl-β-D-thiogalactopyranoside (IPTG). Cells were harvested by centrifugation, resuspended in lysis buffer (50 mM TRIS pH = 7.9, 500 mM NaCl, 20 mM β-mercaptoethanol, 5 mM Imidazole, 1 mM PMSF, 2 µg/ml Leupeptin, 2 µg/ml Pepstatin) and disrupted by sonication. The cell debris was cleared by centrifugation and supernatant was applied to Ni–NTA Column (QIAGEN). Protein fractions were further purified by gel filtration on a Superdex 75 column (GE Healthcare) equilibrated with 50 mM Na_2HPO_4, 100 mM NaCl, 10 mM β-mercaptoethanol, pH = 8.0. Pure protein was concentrated to 2.0 mM using a Vivaspin 20 concentrator (Sartorius Stedim Biotech S.A).

Two repeats of the heptapeptide sequence Tyr^1-Ser^2-Pro^3-Thr^4-Ser^5-Pro^6-Ser^7 phosphorylated at Ser^5 (pSer5) were used as the mimic of RNA polymerase II CTD (purchased from Clonestar Peptide Services, Brno, CZ). The equimolar complex of Nrd1 CID-pSer5 CTD was prepared by mixing of 2.0 mM uniformly [^{15}N,^{13}C]-labelled Nrd1 CID protein with unlabeled pSer5 CTD (ca. 15% excess of the peptide was used to assure complete complex formation).

All NMR experiments used for the resonance assignment were recorded at 293 K on Bruker AVANCE 600 and 900 MHz spectrometers both equipped with cryogenic triple resonance probeheads (Bruker BioSpin). Classical set (Sattler et al. 1999) of double- and triple-resonance experiments was acquired for the assignment of backbone resonances, namely, HNCA, HN(CO)CA, CBCA(CO)NH, HNCACB. The experiments that correlate amide resonances with the carbon alpha of the same residue and/or with the carbon alpha of the preceding residue were acquired with 1,024 complex points in hydrogen dimension and 48 and 64 complex points in nitrogen and carbon dimension, respectively. In the experiments, where Cβ was additionally correlated, 96 complex points were acquired in the carbon dimension. During the processing of the spectra for the backbone assignment maximum entropy calculation on the 2D indirect (N–Cα/β) planes was applied and data were zero-filled to provide matrices of 512 × 64 × 128 points. Side-chain resonances were assigned utilizing 3D HCCH-TOCSY, 3D ^{13}C separated NOESY-HSQC, ^{15}N separated NOESY-HSQC, 3D HNHA, and 2D (HB)CB (CGCD)HD. The 3D spectra were acquired as matrices 2048 × 64 × 128 complex points. The correlation of Cβ and Hδ was acquired as 2,048 × 96 complex points. Spectra were processed with linear prediction applied in the ^{15}N or ^{13}C dimension of the 3D experiments. All spectra were processed using NMRPipe package (Delaglio

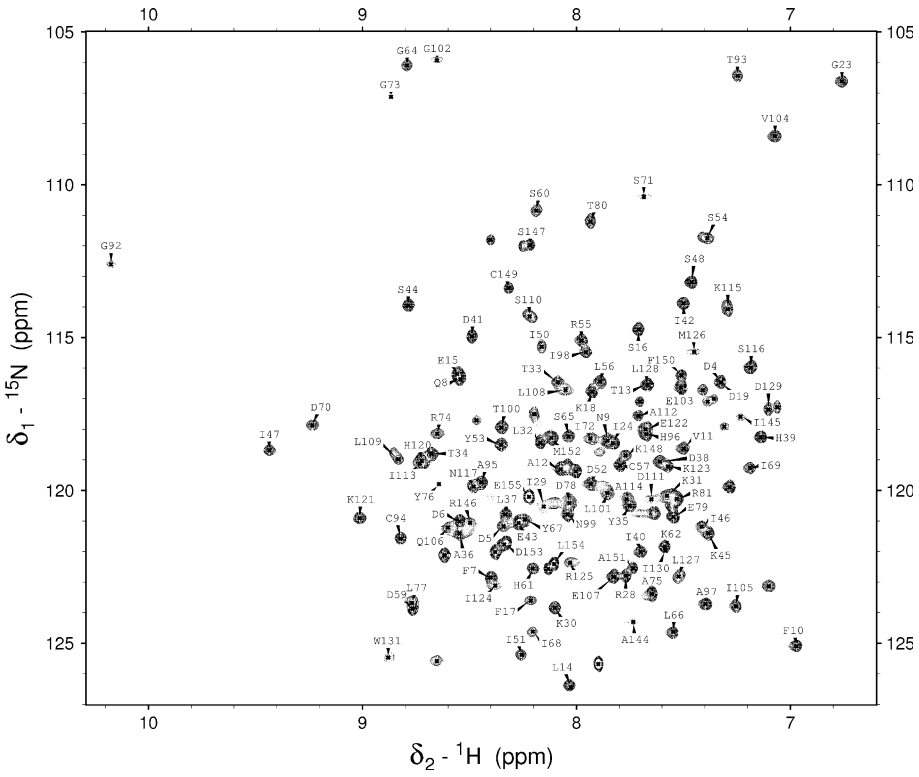


Fig. 1 2D [^{1}H–^{15}N]-TROSY-HSQC spectrum of 2.0 mM uniformly [^{15}N,^{13}C]-labeled Nrd1 CID complexed with unlabeled pSer5 CTD in 50 mM sodium phosphate buffer (pH 8.0), 100 mM NaCl, 10 mM β-mercaptoethanol, (90% H_2O/10% D_2O). The spectrum was acquired at 293 K on a Bruker Avance 900 MHz spectrometer equipped with triple-resonance cryogenic probehead. The assignments are labeled by the *one-letter* code of amino acids accompanied by a sequence number

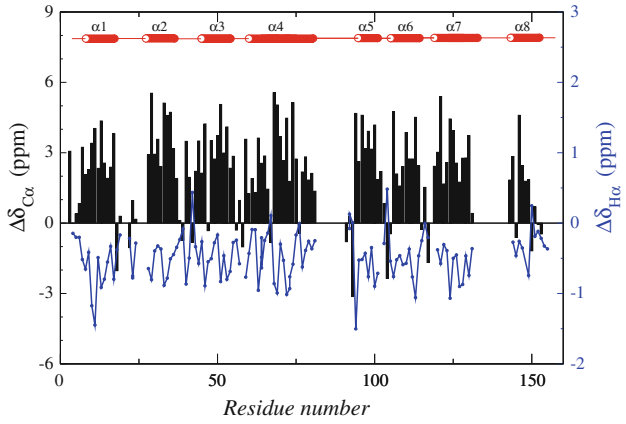


Fig. 2 Hα and Cα chemical shift deviations of the Nrd1 CID-pSer5 CTD complex from the sequence-dependent random coil values along with the secondary structure. *Red cylinders* at the *top* of the chart indicate α-helices

et al. 1995) and the protein resonances were assigned manually using Sparky software (Goddard and Kneller 2004).

Extent of assignments and data deposition

The spectral resonances were assigned by following the well-established procedure (Sattler et al. 1999). This allowed us to assign nearly all observable resonances. The 2D ^{1}H–^{15}N TROSY-HSQC spectrum of Nrd1 CID with amino acids labeled by one-letter code and sequential number is shown in Fig. 1.

For the structured domain, 90.2% of backbone amide protons and ^{15}N nuclei (prolines not counted), 92.9% of the Cα and 87.5% of the Cβ, and 69.1% of ^{1}H, ^{13}C, and ^{15}N side-chain resonances were assigned. The chemical shift deviations of Cα and Hα of the assigned backbone resonances of Nrd1 CID bound to pSer5 CTD from the sequence-dependent random coil values are shown in Fig. 2. These data are in good agreement with the secondary structure elements of the X-ray structure of Nrd1 CID. It is important to note that no resonances could be assigned for expectedly highly flexible unstructured loop regions 82–90 and 132–142. The former overlaps with region of the X-ray model of Nrd1 CID, for which no electron density could be assigned (Vasiljeva et al. 2008).

All assignment for ^{1}H, ^{15}N and ^{13}C backbone and side-chain chemical shifts of Nab3 RRM have been deposited at the BMRB (http://www.bmrb.wisc.edu) and can be accessed under the accession number 17173.

Acknowledgments This work was supported by HHMI/EMBO start-up grant, HFSP Career Development Award, Ministry of Education of the Czech Republic (MSM0021622413 and Ingo LA08008), GACR (204/08/1212 and 305/10/1490), and GAAV (IAA401630 903). The NOESY spectra were obtained at the BMRZ NMR facility supported by the EU-NMR program (RII3-026145).

References

Buratowski S (2009) Progression through the RNA polymerase II CTD cycle. Mol Cell 36:541–546

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

Egloff S, Murphy S (2008) Cracking the RNA polymerase II CTD code. Trends Genet 24:280–288

Goddard TD, Kneller DG (2004) SPARKY 3. University of California, San Francisco

Meinhart A, Cramer P (2004) Recognition of RNA polymerase II carboxy-terminal domain by 3′-RNA-processing factors. Nature 430:223–226

Richard P, Manley JL (2009) Transcription termination by nuclear RNA polymerases. Genes Dev 23:1247–1269

Sattler M, Schleucher J, Griesinger C (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Prog Nucl Magn Reson Spectrosc 34:93–158

Vasiljeva L, Kim M, Mutschler H, Buratowski S, Meinhart A (2008) The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. Nat Struct Mol Biol 15:795–804

Cell

The Solution Structure of the ADAR2 dsRBM-RNA Complex Reveals a Sequence-Specific Readout of the Minor Groove

Richard Stefl,[1,4,6] Florian C. Oberstrass,[1,6,7] Jennifer L. Hood,[3] Muriel Jourdan,[1,8] Michal Zimmermann,[5] Lenka Skrisovska,[1] Christophe Maris,[1] Li Peng,[2] Ctirad Hofr,[5] Ronald B. Emeson,[2] and Frédéric H.-T. Allain[1,*]

[1]Institute of Molecular Biology and Biophysics, ETH Zurich, CH-8093 Zürich, Switzerland
[2]Department of Pharmacology, Vanderbilt University, Nashville, TN 37232, USA
[3]Neuroscience Graduate Program, Vanderbilt University, Nashville, TN 37232, USA
[4]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, CZ-62500 Brno, Czechia
[5]Department of Functional Genomics and Proteomics, Institute of Experimental Biology, Faculty of Science, Masaryk University, CZ-62500 Brno, Czechia
[6]These authors contributed equally to this work
[7]Present address: Department of Bioengineering, Stanford University, 318 Campus Drive, Stanford, CA 94305, USA
[8]Present address: Departement de Chimie Moleculaire, 38041 Grenoble Cedex09, France
*Correspondence: allain@mol.biol.ethz.ch
DOI 10.1016/j.cell.2010.09.026

SUMMARY

Sequence-dependent recognition of dsDNA-binding proteins is well understood, yet sequence-specific recognition of dsRNA by proteins remains largely unknown, despite their importance in RNA maturation pathways. Adenosine deaminases that act on RNA (ADARs) recode genomic information by the site-selective deamination of adenosine. Here, we report the solution structure of the ADAR2 double-stranded RNA-binding motifs (dsRBMs) bound to a stem-loop pre-mRNA encoding the R/G editing site of GluR-2. The structure provides a molecular basis for how dsRBMs recognize the shape, and also more surprisingly, the sequence of the dsRNA. The unexpected direct readout of the RNA primary sequence by dsRBMs is achieved via the minor groove of the dsRNA and this recognition is critical for both editing and binding affinity at the R/G site of GluR-2. More generally, our findings suggest a solution to the sequence-specific paradox faced by many dsRBM-containing proteins that are involved in post-transcriptional regulation of gene expression.

INTRODUCTION

ADARs convert adenosine-to-inosine (A-to-I) by hydrolytic deamination in numerous mRNA and pre-mRNA transcripts (Bass, 2002; Nishikura, 2006). Due to the similar base-pairing properties of both nucleosides, inosine is interpreted as guanosine by cellular machineries during the processes of translation and splicing. In this way, editing-mediated alterations in sequence can alter codon identity or base-pairing interactions within higher-order RNA structures (Bass, 2002; Nishikura, 2006). As a result, ADARs can create protein isoforms or regulate gene expression at the RNA level (Bass, 2002; Nishikura, 2006; Valente and Nishikura, 2005). ADARs are widely expressed in most cell types, yet their expression and activity in neuronal tissues has been shown to be important for proper nervous system function (Higuchi et al., 2000; Palladino et al., 2000). Recent high-throughput sequencing analysis of A-to-I editing identified over 55 editing sites within the coding regions of mRNAs, with 38 of these sites involving a codon change that specifies an alternative amino acid. Many of these changes involve RNA transcripts encoding proteins that are critical for nervous system function (Li et al., 2009).

ADARs from all characterized species have a modular domain organization consisting of one-to-three dsRBMs followed by a conserved C-terminal catalytic adenosine deaminase domain. The structures of the two dsRBMs and of the isolated catalytic domain of ADAR2 have been determined in their free states (Macbeth et al., 2005; Stefl et al., 2006). Among the best-studied ADAR substrates are pre-mRNAs encoding subunits of the α-amino-3-hydroxyl-5-methyl-4-isoxazole-propionate (AMPA)-subtype of ionotropic glutamate receptor (GluR-2, GluR-3 and GluR-4; Higuchi et al., 2000, 1993; Melcher et al., 1996) that contain one or both of two highly edited and functionally relevant sites, namely the R/G and Q/R editing sites (Aruscavage and Bass, 2000; Lomeli et al., 1994; Melcher et al., 1996).

ADARs can edit RNA substrates either specifically or nonspecifically depending upon the structures of the RNA substrates (Bass, 2002). In vitro studies have shown editing of up to 50% of the adenosine residues in both strands using synthetic dsRNAs that are perfectly complementary (Cho et al., 2003; Lehmann and Bass, 2000). Such nonspecific editing can be explained by the presence of dsRBMs which are thought to bind dsRNA in a sequence-independent manner (Tian et al., 2004), yet it remains unclear how certain RNA substrates are edited in a site-specific fashion. Several studies have suggested that the

presence of noncanonical elements in these dsRNAs–such as mismatches, bulges, and loops–could be important for site-selective A-to-I conversion (Bass, 2002; Stefl et al., 2006; Tian et al., 2004).

The dsRBMs of ADARs are not only essential for editing (Stefl et al., 2006; Valente and Nishikura, 2007), but the dsRBM also represents the second most abundant family of RNA recognition motifs. In addition to RNA editing, dsRBMs are involved in numerous post-transcriptional regulatory processes and most prominently in micro RNA (miRNA) biogenesis and function and RNA export (Dreyfuss et al., 2002; Tian et al., 2004). The few solved structures of dsRBM-containing proteins bound to short, synthetic RNA duplexes have suggested that dsRBMs recognize the A-form helix of dsRNA in a sequence-independent manner, since the majority of dsRBM-RNA interactions involve direct contact with the 2′-hydroxyl groups of the ribose sugars and direct or water-mediated contacts with nonbridging oxygen residues of the phosphodiester backbone (Gan et al., 2006; Ramos et al., 2000; Ryter and Schultz, 1998; Wu et al., 2004), and that a subclass of dsRBMs prefer stem-loops over A-form helices (Ramos et al., 2000; Wu et al., 2004).

We previously determined that each of the two dsRBMs of ADAR2 bind to a distinct location on the GluR-2 RNA encompassing the R/G editing site and that the interdomain linker (amino acids 147-231) is unstructured both in the free protein and in the complex (Stefl et al., 2006). To better understand RNA substrate recognition by ADAR2, we have determined the solution structure of the RNA helix surrounding the editing site and the solution structure of the two dsRBMs of ADAR2 bound to the GluR-2 R/G site.

RESULTS

Structure of the GluR-2 R/G RNA Helix Surrounding the Editing Site

The GluR-2 R/G site (A8) is embedded within a 71 nt RNA stem-loop containing three base-pair mismatches and capped by a 5′-GCUAA-3′ pentaloop (Figure 1A). We previously determined the structure of the apical part of the stem-loop and showed that the pentaloop is structured and adopts a fold reminiscent of a UNCG-type family of tetraloops (Stefl and Allain, 2005). Here, we have investigated the structure of the RNA helix surrounding the editing site that contains two A-C mismatches, one at the editing site (A8) and a second one ten base-pairs downstream (A18, Figure 1B). Monitoring adenine C2 chemical shifts (a sensitive probe to monitor the protonation state of N1) during a pH titration, we observed that A8 and A18 are fully protonated below pH 6.5, partially protonated between pH 6.5–8.5, and unprotonated above pH 8.5 (Figures 1H and 1I). The pKa for the adenosines N1 can be estimated between 7 and 7.5 at 310 K, which is 3.3 units higher than the value determined for an isolated AMP (pKa of 4.0; Legault and Pardi, 1994). Using 863 nOe-derived distance restraints, we solved the structure of the free RNA in the protonated state (pH 6.2). The structure is well defined, even for the A-C mismatches (Figure 1E and Table 1) that are stacked inside the stem. Therefore, at pH 6.2, the R/G site has a regular A-form helix structure (Figure 1D) containing two A^+-C base-pairs adopting a wobble conformation, stabilized by two hydrogen bonds each (Figures 1F and 1G).

Structure of ADAR2 dsRBMs Bound to Their Respective RNA Targets

Considering the distinct RNA binding location found previously for each dsRBM (Stefl et al., 2006) and the high molecular weight (over 50 kDa) of the complex formed between the two dsRBMs of ADAR2 and the GluR-2 R/G substrate (Figure 1A), we adopted a modular approach to solve the structure of this complex in solution. To this end, we first solved the structure of dsRBM1 in complex with a modified GluR-2 upper stem–loop (USL, Figure 1C, and Figure S1 available online) and then the structure of dsRBM2 bound to the GluR-2 lower stem-loop that contains the editing site (LSL, Figure 1B, and Figure S2). The use of a GluR-2 R/G USL mutant to determine the structure of dsRBM1 in complex with RNA was dictated by the poor data quality that we obtained with the wild-type (WT) sequence. In changing the loop sequence to that found in the GluR-3 USL (Aruscavage and Bass, 2000), we obtained a smaller and more stable RNA which provided NMR data of higher quality.

A total of 1707 and 1929 nOe-derived distance restraints (including 36 intermolecular ones for each complex) for ADAR2 dsRBM1–GluR-2 R/G USL mutant and ADAR2 dsRBM2–GluR-2 R/G LSL complexes, respectively, were used to obtain well-defined structures (Figure 2 and Table 1). The two dsRBM-RNA complexes are stabilized by a combination of hydrophobic interactions, hydrogen bonding and electrostatic contacts. In both dsRBM–RNA complexes, the dsRBMs adopt the expected $\alpha\beta\beta\beta\alpha$ topology in which the two α helices are packed along the three-stranded antiparallel β sheet. The entire interaction surface spans 12-14 base-pairs covering two minor grooves and a major groove (Figure 2). In both complexes, three distinct regions of the dsRBMs are involved in interaction with RNA. The first region is the helix α1, which interacts with the first minor groove of the RNA. The second region is a well-conserved KKNAK-motif, located at the amino-terminal tip of helix α2 and the preceding loop, that contact the RNA with nonsequence specific contacts between lysine side-chains and the phosphate oxygens across the major groove of the RNA (Lys127, 128, and 131 for dsRBM1 and Lys281, 282, 285 for dsRBM2, Figure 2). In addition, the dipole moment of helices α2 creates a positive charge in the N-terminal tip of these helices that interacts with the negatively charged phosphate backbone. This second set of interactions is mediated by the main-chain amides of K127 and K281, which are hydrogen bonded with the phosphates oxygen of A24 and U11, respectively (Figure 2). The third region of contact is the β1-β2 loop which interacts with the second minor groove of the RNA. The overall architecture of these two complexes resembles other previously determined dsRBM–RNA structures (Blaszczyk et al., 2004; Gan et al., 2008; Gan et al., 2006; Ramos et al., 2000; Ryter and Schultz, 1998; Stefl et al., 2005a; Wu et al., 2004). However, a detailed inspection of the interaction regions revealed striking differences between the two complexes and other dsRBM-RNA complexes, particularly in the first and the third regions where both dsRBMs present unexpected sequence-specific contacts to the RNA minor grooves (Figure 2).

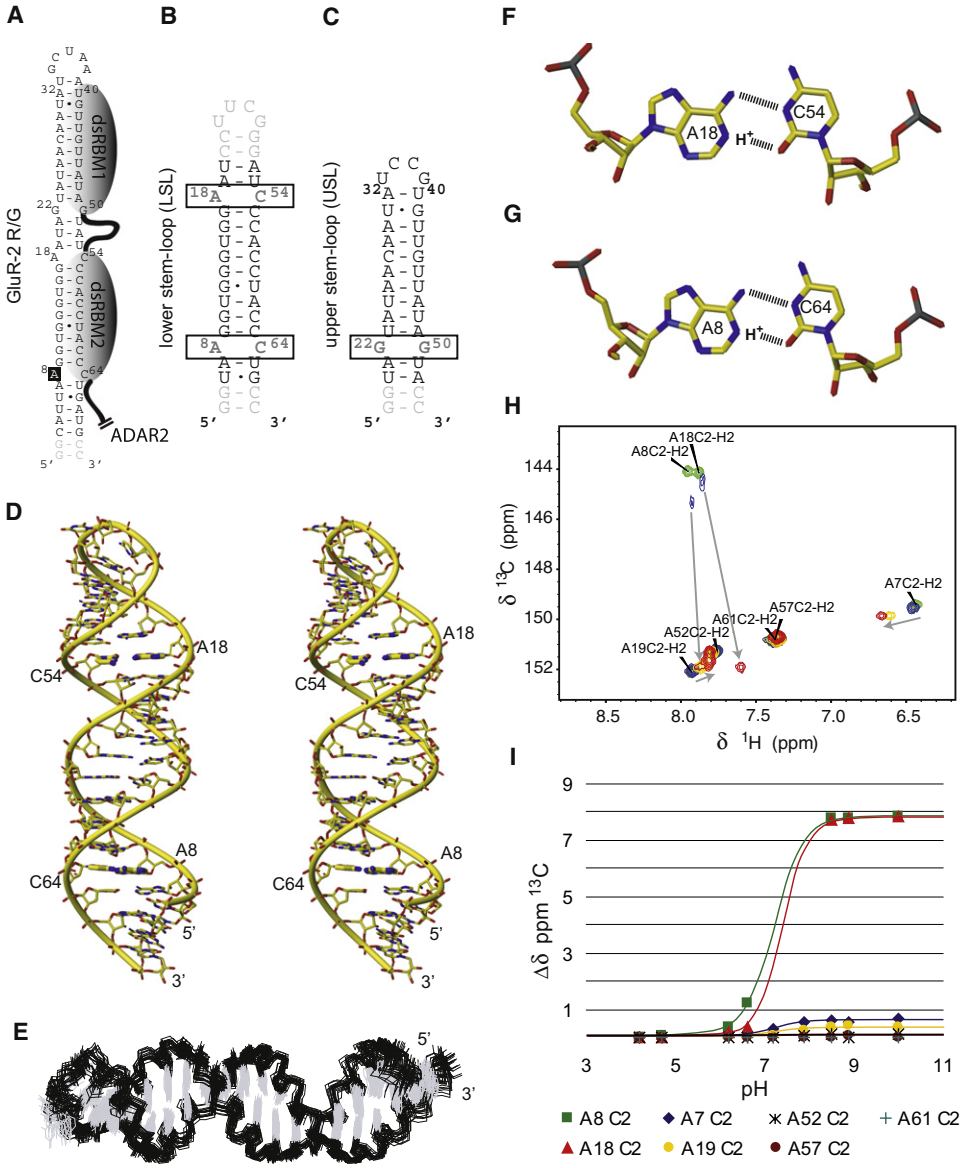


Figure 1. Secondary Structures of the RNAs and Solution Structure of GluR-2 R/G LSL RNA

(A) Secondary structure of GluR-2 R/G RNA. The indicated binding regions for the dsRBMs were proposed previously (Stefl et al., 2006).

(B) Secondary structure of the GluR-2 R/G lower stem-loop (LSL).

(C) Secondary structure of the GluR-2 R/G upper stem-loop (USL).

(D) Stereo view of the most representative structure of GluR-2 R/G LSL RNA. The A^+-C wobble base-pairs are highlighted in bold sticks.

(E–G) (E) Overlay of the 20 lowest energy structures of GluR-2 R/G LSL. The A^+-C wobble base-pairs A18-C54 (F) and A8-C64 (G) are shown.

(H) H2-C2 region of adenines in the ^{13}C-^{1}H-HSQC spectra of the GluR-B R/G LSL is shown at pH 4.7 (green peaks), 6.6 (blue peaks), 7.9 (orange peaks) and 8.9 (red peaks). The two adenines involved in the A^+-C wobble base-pair showed drastic perturbation.

(I) Diagram showing the pH-dependence of ^{13}C chemical shift changes of adenine C2′s.

Sequence-Specific Recognition by ADAR2 dsRBM1

In the ADAR2 dsRBM1–RNA complex, contacts from helix α1 are centered at the A32-U40 base-pair below the UCCG tetraloop (Figures 2A and 2C). Met84 makes a sequence-specific hydrophobic contact with H2 of A32 and Asn87 contacts the 2′-hydroxyl and O2 of U40. The Oε of Glu88 is hydrogen bonded to the amino group of the first cytosine of the tetraloop. In addition, Leu83 makes hydrophobic contacts with the sugar of G41. The entire helix α1 is tightly inserted in the minor groove created by the UCCG tetraloop and two adjacent base-pairs (Figure 2A). The β1-β2 loop of dsRBM1 binds the following minor groove of the RNA. This minor groove is widened as it has to accommodate base-pairing of two guanosines that make an N1 symmetrical G22-G50 mismatch (Figures 2A and 2D) that are the center of this interaction. Val104 side-chain contacts the H8 of G50 (that adopts a *syn* conformation) and a sequence-specific hydrogen

Table 1. NMR and Refinement Statistics for the GluR-2 R/G Upper Stem-Loop RNA Bound to ADAR2 dsRBM1, the Free GluR-2 R/G Lower Stem-Loop RNA, and Its Complex with ADAR2 dsRBM2, and the RDC-Reconstructed Complex of the Full-Length GluR-2 R/G Stem-Loop RNA Bound to ADAR2 dsRBM12

	USL RNA – dsRBM1 Complex		LSL RNA	LSL – dsRBM2 Complex		SL RNA – dsRBM12 complex	
	USL RNA	dsRBM1		LSL RNA	dsRBM2	SL RNA	dsRBM12
NMR Distance and Dihedral Constraints and RDCs							
Distance restraints							
Total NOE	645	927	781	702	1054	1252	1981
Intraresidue	309	201	389	365	216	620	417
Interresidue	336	726	392	337	838	631	1564
Sequential (|i-j| = 1)	270	252	352	306	241	555	493
Nonsequential (|i-j| > 1)	66	474	40	31	597	76	1071
Hydrogen bonds	35	64	81[a]	75	62	132	126
Protein–RNA intermolecular	36			36		72	
Total dihedral angle restraints	180		252	267			
RNA							
Sugar pucker	34		84	84			
Backbone[b]	146		168	183			
RDC restraints							45[d]
Structure Statistics[c]							
Violations (mean and SD)							
Number of distance restraint violations > 0.2 Å	8.45 ± 2.50		0	1.10 ± 1.25		14.31 ± 3.86	
Number of dihedral angle restraint violations > 5°	0.7 ± 0.47		0	0		5.30 ± 3.32	
Max. dihedral angle restraint violation (°)	5.82 ± 1.22		3.28 ± 0.77	2.69 ± 1.12		15.51 ± 2.36	
Max. distance constraint violation (Å)	0.29 ± 0.03		0.16 ± 0.01	0.23 ± 0.06		0.32 ± 0.05	
Deviations from idealized geometry[d]							
Bond lengths (Å)	0.0042 ± 0.00007		0.0046 ± 0.00005	0.0041 ± 0.00005		0.0048 ± 0.00005	
Bond angles (°)	1.989 ± 0.011		2.137 ± 0.017	1.903 ± 0.011		1.995 ± 0.008	
RDCs violations							
Absolute RDC violations (Hz)						1.12 ± 0.82	
Average pairwise r.m.s.d (Å)[c]							
Protein (79-142) for dsRBM1; (221-282) for dsRBM2							
Heavy atoms	1.11 ± 0.17			1.01 ± 0.12		1.60 ± 0.36	
Backbone atoms	0.59 ± 0.14			0.37 ± 0.08		1.22 ± 0.42	
RNA							
All RNA heavy atoms	0.60 ± 0.16		1.15 ± 0.35	1.48 ± 0.51		1.30 ± 0.40	
Complex							
All complex heavy atoms	1.01 ± 0.15			1.49 ± 0.39		1.75 ± 0.31	

[a] In the final structure calculations of the free RNA, H-bond restraints were applied in the two A-C mismatches. This is based on initial structures and on the protonation state of A8/A18. For the structures of the RNA in complex no H-bond restraints for the two A-C mismatches have been applied.
[b] Based on A-form geometry derived from high-resolution crystal structures: α(270°–330°), β(150°–210°), γ(30°–90°), δ(50°–110°), ε(180°–240°), and ζ (260°–320°). These restraints were used only for the double-helical region. No angle restraints were imposed on the two A-C mismatches and the loops.
[c] Calculated for an ensemble of the 20 lowest energy structures.
[d] 16 RDCs of dsRBM1 and 29 RDCs of dsRBM2.

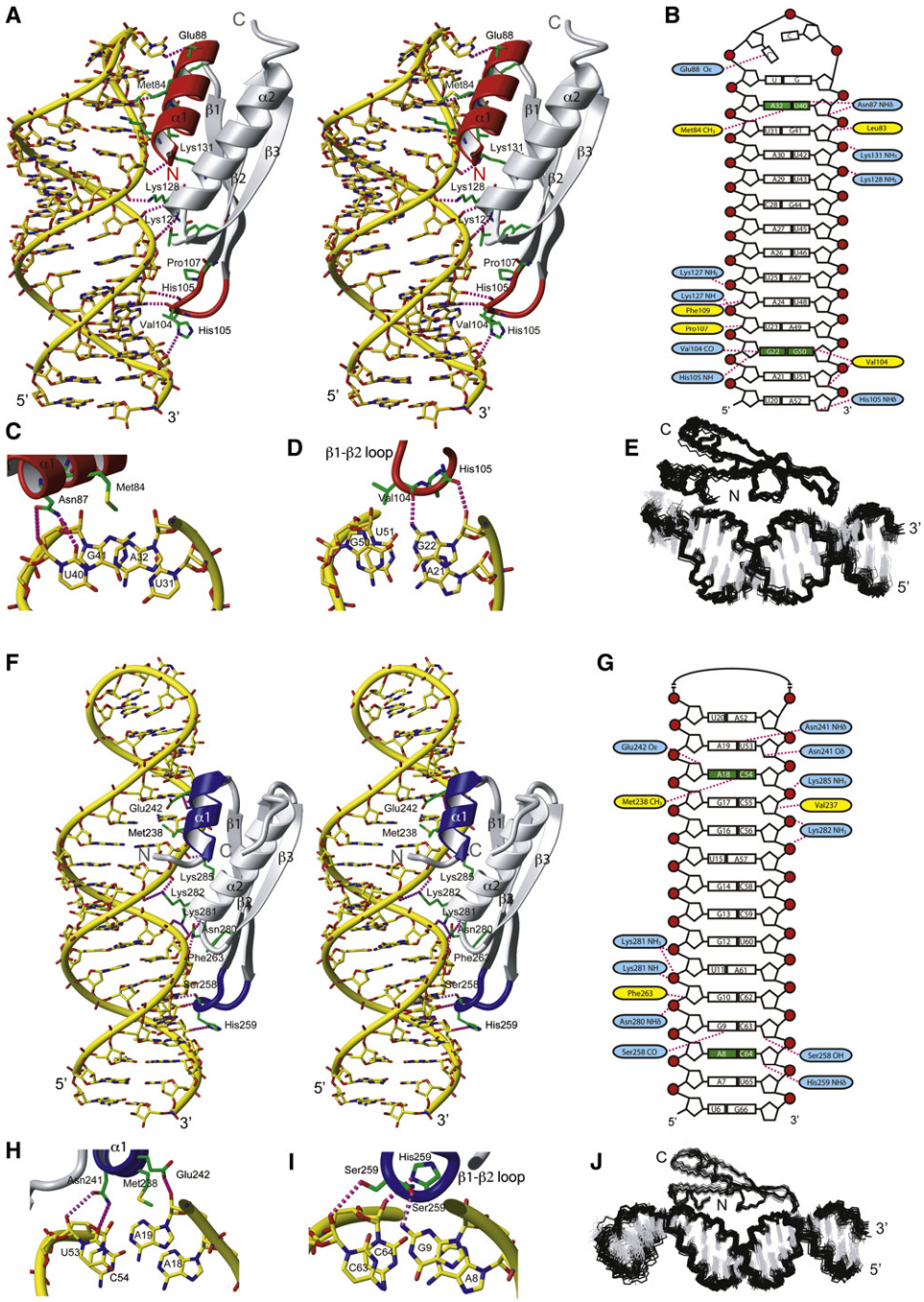


Figure 2. RNA Recognition by ADAR2 dsRBM1 and dsRBM2

(A) Stereo view of the most representative structure of dsRBM1 bound to USL RNA. The RNA is represented as a yellow stick model and the protein is shown as a ribbon model with residues that contact the RNA shown in green. Helix α1 and the β1-β2 loop that mediate the sequence-specific contacts are colored in red. Hydrogen bonds are indicated by magenta dotted lines. (B) Scheme showing contacts between dsRBM1 and the USL RNA. Protein residues that form hydrogen bonds to the RNA are shown in blue and the one having hydrophobic interactions are in yellow. Close-up view of minor groove sequence-specific recognitions mediated by helix α1 (C) and the β1-β2 loop (D) of dsRBM1. (E) Overlay of the 20 lowest energy structures of the dsRBM1-USL complex. (F) Stereoview of the most representative structure of the dsRBM2 bound to LSL RNA. Helix α1 and the β1-β2 loop that mediate the sequence-specific contacts are colored in blue. (G) Scheme showing contacts between dsRBM2 and the LSL RNA. Close-up view of the minor groove sequence-specific recognitions mediated by helix α1 (H) and the β1-β2 loop (I). (J) Overlay of the 20 lowest energy structures of the dsRBM2-LSL complex. For NMR data of these two complexes, see also Figure S1 and Figure S2.

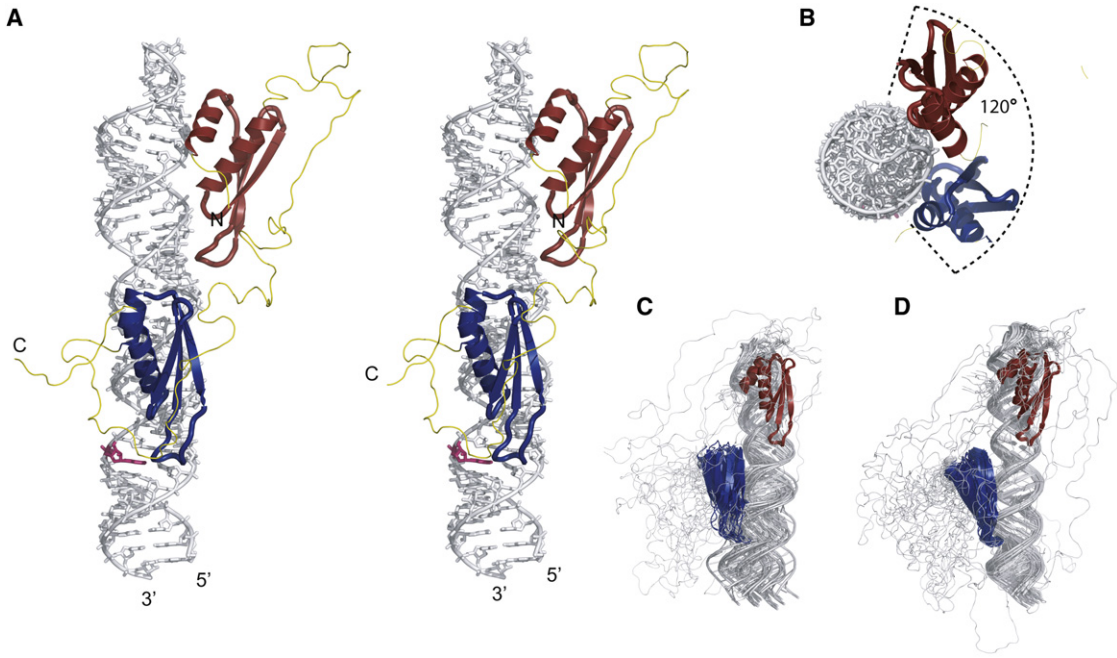


Figure 3. Structure of ADAR2 dsRBM12 Bound to GluR-2 R/G

(A) Stereo view of the most representative RDC-reconstructed structure of the ADAR2 dsRBM12 bound to GluR-2 R/G. The RNA is represented as a stick model (in gray; the edited adenosine is highlighted in pink) and the protein is shown as a ribbon model (dsRBM1 in red; dsRBM2 in blue; linker in yellow). (B) Top view of the complex. Overlay of the 20 lowest energy structures calculated without (C) and with RDCs (D), superimposed on dsRBM1.

bond is formed between the main-chain carbonyl of V104 and the amino group of G22. The widened minor groove accommodates additional interactions between three side-chains (Phe109, Pro107, His105) and the sugars of the base-pairs above and below. Altogether, dsRBM1 binds the RNA stem-loop at a single register via two sequence-specific contacts at two consecutive RNA minor grooves: a hydrogen bond to the amino group of the G22 in the GG mismatch via the β1-β2 loop and an hydrophobic contact to the adenine H2 of A32 via Met84 in helix α1.

Sequence-Specific Recognition by ADAR2 dsRBM2

The dsRBM2 of ADAR2 is adjacent to the deaminase domain and is essential for A-to-I editing at the R/G site (Stefl et al., 2006; Xu et al., 2006). In the ADAR2 dsRBM2–GluR-2 R/G LSL complex, Asn241, Glu242, Met238, Val 237 of helix α1 contact the minor groove region centered at the A18-C54 mismatch (Figures 2F and 2H). At pH 7.6, where the protein-RNA complex has been determined, this mismatch is unprotonated and Met238 makes a sequence-specific hydrophobic contact with A18 H2. Contacts to the base-pair above and below by Asn241 and Glu242, and by Val 237, respectively, further stabilize the interaction of helix α1 in this region (Figure 2H). The β1−β2 loop of dsRBM2 interacts with the second minor groove. The contacts are centered at the G9-C63 Watson-Crick base-pair located above the A8-C64 mismatch containing the editing site. A sequence-specific hydrogen bond is formed between the main-chain carbonyl of Ser258 and the amino of G9 (Figures 2F and 2I). Additionally, nonsequence specific contacts between the side-chains of Ser 258, His 259 and Phe 263 and the G9-C63 base-pair and the base-pairs above and below increase the stability of the interaction with the RNA minor groove (Figure 2G). In the vicinity of the editing site, dsRBM2 contacts C63, while A8 is not contacted by any residue from the β1-β2 loop therefore making A8 accessible to the deaminase domain. Altogether, dsRBM2 similar to dsRBM1, recognizes the RNA helix via two sequence-specific contacts at two consecutive RNA minor grooves: a hydrogen bond to the amino group of the G9 at the GC 3′ to the editing site via the β1-β2 loop and a hydrophobic contact to the adenine H2 of A18 via Met238 in helix α1. In the NMR spectra (*data not shown*), we could observe intermolecular nOes corresponding to dsRBM2 being positioned at a second binding register one base-pair above (although with only 20% occupancy). In this case the β1-β2 loop contact G10 and Met 238 contact A19. Although two consecutive binding sites for dsRBM2 are observed here, they both confirm the sequence-specific nature of the dsRBM2-RNA interaction.

Structure of ADAR2 dsRBM12 in Complex with GluR-2 R/G RNA

Next, we determined the structure of ADAR2 dsRBM12 in complex with GluR-2 R/G RNA (Figures 3A and 3B). To calculate an atomic model of this complex, we used the distance constraints measured in the two sub-complexes described above (Figure 3C). This strategy could be used considering (1) the distinct RNA binding location for each dsRBMs, with no mutual interactions (Stefl et al., 2006), (2) the flexible unstructured linker connecting dsRBM1 and dsRBM2 in the complex

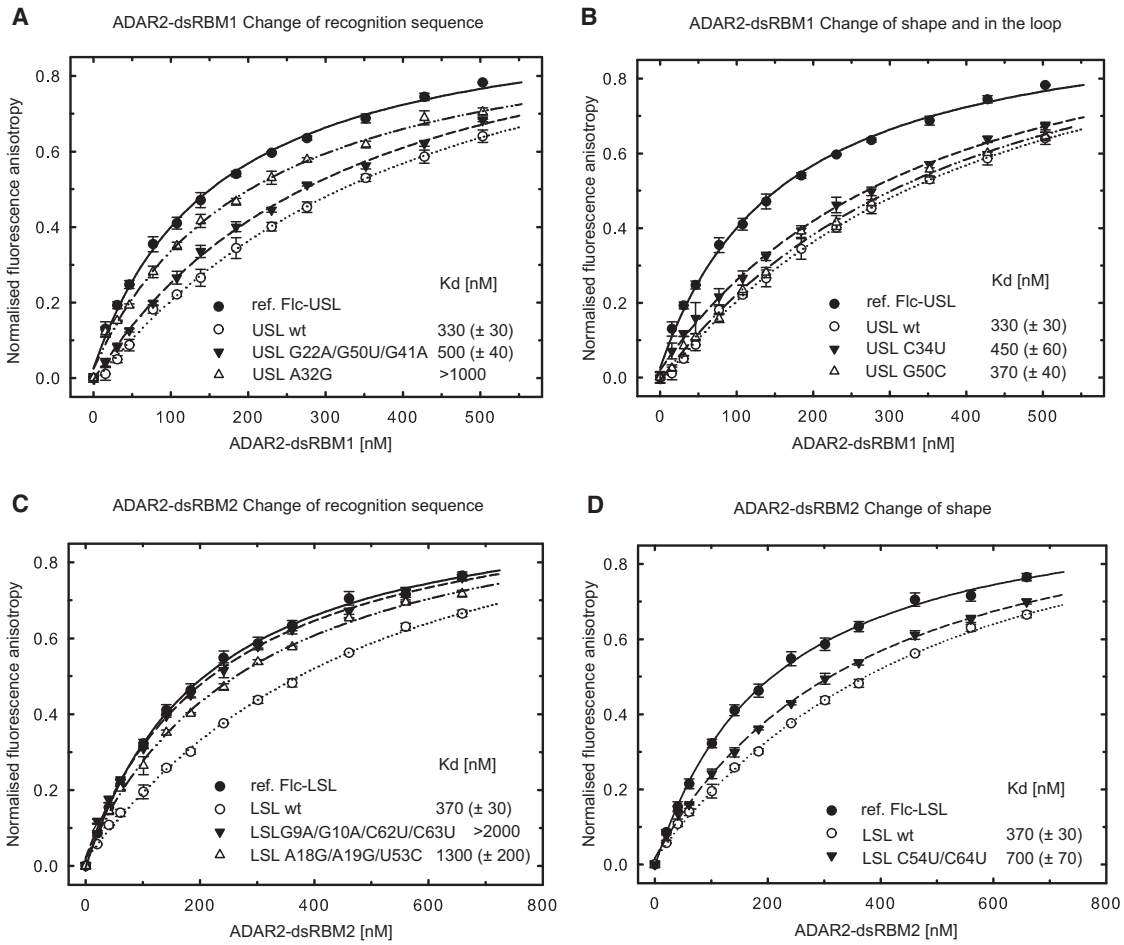


Figure 4. ADAR2 dsRBMs Bind Preferentially to RNAs that Contains Their Sequence-Specific Recognition Motifs

(A) ADAR2 dsRBM1 was titrated with fluorescently labeled USL and binding was measured by fluorescence anisotropy (black circles; fluorescein labeled reference, Flc-USL). The same experiment was then carried out in the presence of competing unlabeled USL wt (○), USL G22A/G50U/G41A mutant (▼), and USL A32G mutant (△). Equilibrium dissociation constants (K_d) were calculated from the best fit to the data as described in Experimental Procedures.

(B) The same assay as shown in (A) but for USL C34U mutant (▼) and USL G50C mutant (△).

(C) ADAR2 dsRBM2 was titrated with fluorescently labeled LSL and binding was measured by fluorescence anisotropy (●; fluorescein labeled reference, Flc-LSL). The same experiment was then carried out in the presence of competing unlabeled LSL wt (○), LSL G9A/G10A/C62U/C63U mutant (▼), and LSL A18G/A19G/U53C mutant (△).

(D) The same assay as shown in (C) but for LSL C54U/C64U mutant (▼).Wild-type and mutant sequences are shown in Figure S3.

(Stefl et al., 2006) and (3) an overlap in the RNA sequence of the joint region of the subcomplexes (Figure 1). Long-range structural constraints for this elongated complex were derived from residual dipolar couplings (RDCs) measured with a deuterated protein on the full-length complex (dsRBM12 bound to GluR-2 R/G RNA, Figure 1A). The pentaloop which is not contacted by dsRBM1 was modeled using the structure that was determined previously (Stefl and Allain, 2005). With this strategy, we could then determine a precise solution structure of this 50 kDa complex using 45 ^{15}N-^{1}H RDCs (Figure 3D, Table 1). In the structure, the two dsRBMs bind one face of the RNA covering approximately 120 degrees of the space around the RNA helix (Figure 3B). This suggests that the binding of an additional molecule of ADAR2 would be sterically possible, consistent with studies indicating that ADAR2 dimerization is necessary for RNA editing (Chilibeck et al., 2006; Cho et al., 2003; Gallo et al., 2003; Valente and Nishikura, 2007).

Sequence-Specific Contacts of ADAR2 dsRBMs Are Important for Binding Affinity

To confirm the ADAR2 dsRBMs sequence-specific preference in a quantitative solution binding assay, we performed fluorescence anisotropy (FA) experiments by titrating dsRBM1 and dsRBM2 against labeled USL and LSL RNAs, respectively. Unlabeled wild-type and mutant RNAs (Figure S3) were used for competition experiments as described in Experimental procedures. The equilibrium dissociation constants were calculated from the displacement of the binding curves (Figure 4). We designed two sets of mutations, one set was designed to change the recognition sequence of USL and LSL RNAs (Figures 4A and

4C and Figure S3) and a second set was designed to maintain the recognition sequence, but change the RNA shape via mismatches of USL and LSL RNAs into Watson-Crick base-pairs (Figures 4B and 4D and Figure S3) to measure their effect on overall binding affinity. In mutating any of the bases that are recognized in a sequence-specific manner by dsRBM1 in USL (G22, A32 or C34), the apparent affinity is reduced compared to the wild-type (Figures 4A and 4B). However when the G22-G50 mismatch is replaced by a Watson-Crick G22-C50 pair, the affinity is almost identical to wild-type RNA, confirming that dsRBM1 recognizes the sequence rather than the shape of the RNA helix (note that G41 was mutated in the first RNA mutant to prevent the sequence-specific recognition of G41 by dsRBM1). Similarly for the LSL, when G9 or A18 are mutated, dsRBM2 binding is reduced more than five-fold (Figure 4C), yet when the two AC mismatches are replaced by Watson-Crick AU pairs, the affinity is only reduced by two-fold (Figure 4D). In this latter context, the sequence-specific contacts are the same for the WT and mutant RNAs, but the presence of a more deformable A18-C54 base-pair in the WT structure could explain the higher affinity of dsRBM2 to the WT RNA (note that additional mutations were introduced in the first two RNA mutants of LSL to abolish the two binding registers found in the wild-type LSL). Altogether, the FA data strongly support the idea that the sequence-specific interactions observed in the structures of ADAR2 dsRBMs-dsRNA are important for the affinity of both dsRBMs and that they finely tune the preferential binding to these recognition motifs.

Sequence-Specific Contacts of ADAR2 dsRBMs Are Important for Editing

To test the functional importance of the four sequence-specific contacts identified in the ADAR2 dsRBM12-GluR-2 R/G RNA complex, single amino acid mutants in helix α1 (M84 or M238) were mutated to alanine or double mutants in the β1−β2 loop in either dsRBM1 or dsRBM2 were evaluated for their ability to edit the wild-type GluR-2 R/G site (Figure 5A). It was necessary to generate double mutants around the carbonyls of V104 in dsRBM1 and S258 in dsRBM2 to change the structure of the main-chain of this loop. All four mutants showed a significant decrease in RNA editing ranging from a near ablation of editing (S258A,H259A in the β1−β2 loop of dsRBM2), to 20% editing (V104A,H105A in the β1−β2 loop of dsRBM1 and M84A in helix α1 of dsRBM1), to 30% editing (M238A in helix α1 of dsRBM2) of that demonstrated by the wild-type protein. These data clearly show that the loss of the sequence-specific contacts of any of the two dsRBMs strongly decreases editing at the R/G site with the contact mediated by the β1−β2 loop of dsRBM2 more strongly affecting editing than the other contacts. In agreement with deletion studies of ADAR2 (Macbeth et al., 2004; Stefl et al., 2006), the S258A,H259A mutations have a stronger effect, likely due to the binding of the β1−β2 loop of dsRBM2 near the editing site.

Converse experiments in which mutations in the sequence-specific recognition motifs of dsRBM2 (mut1 and mut2), dsRBM1 (mut4) or both (mut3) within the GluR-2 RNA (Figure S4) were assessed for their ability to affect R/G editing by wild-type ADAR2 revealed a significant decrease in maximal editing rates (V_{max}) for all RNA mutants tested (Figure 5B) providing further support for the functional significance of these contacts. Best-fit kinetic curves for wild-type and mutant RNAs corresponded to a model of substrate inhibition, consistent with previously observed kinetic models for ADARs in which the formation of a ternary complex containing an ADAR dimer and RNA substrate is required for efficient adenosine deamination.

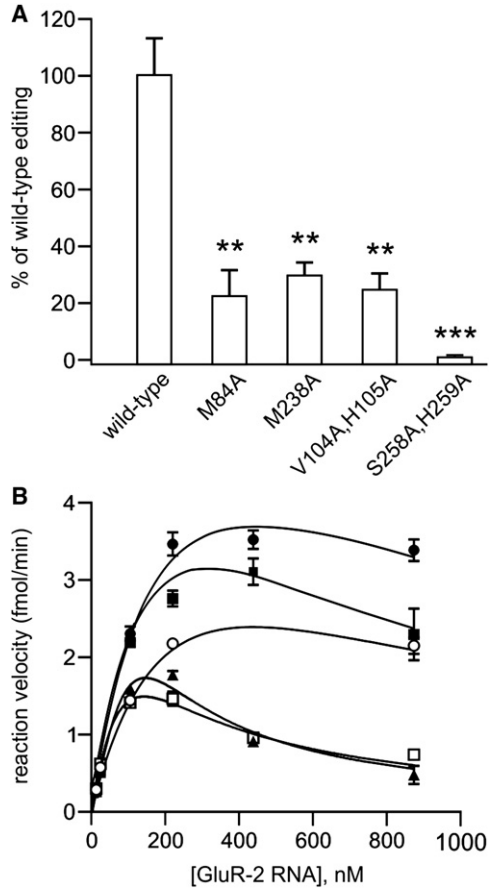


Figure 5. Sequence-Specific Contacts of ADAR dsRBMs Are Important for Editing Activity

(A) Quantitative analysis of in vitro editing efficiency for ADAR2 dsRBM double mutants; all mutants were assayed in duplicate for in vitro editing activity at the GluR-2 R/G site using three independent nuclear extracts (mean ± SEM; *$p < 0.05$, **$p < 0.005$; ***$p < 0.001$).

(B) Kinetic analysis of wild-type ADAR2 editing with GluR-2 R/G mutants. Increasing concentrations of GluR-2 RNAs (see Figure S4; wild-type ●; mut 1 ■, mut 2 ○, mut 3 ▲, mut 4 □) were incubated with wild-type rat ADAR2 protein as described above; all mutant RNAs were assayed in triplicate for determination of in vitro reaction velocity (mean ± SEM). Nonlinear fitting of kinetic curves corresponded to a model of substrate inhibition (R^2 = 0.91-0.98 for all RNAs) with V_{max} values corresponding to 3.92, 3.84, 2.08, 1.20, and 1.29 fmol/min for wild-type, mut1, mut2, mut3, and mut4, respectively. Wild-type and mutant sequences are shown in Figure S4.

DISCUSSION

In solving the structure of ADAR2 dsRBMs bound to the GluR-2 R/G site, we demonstrated that despite forty-four possible

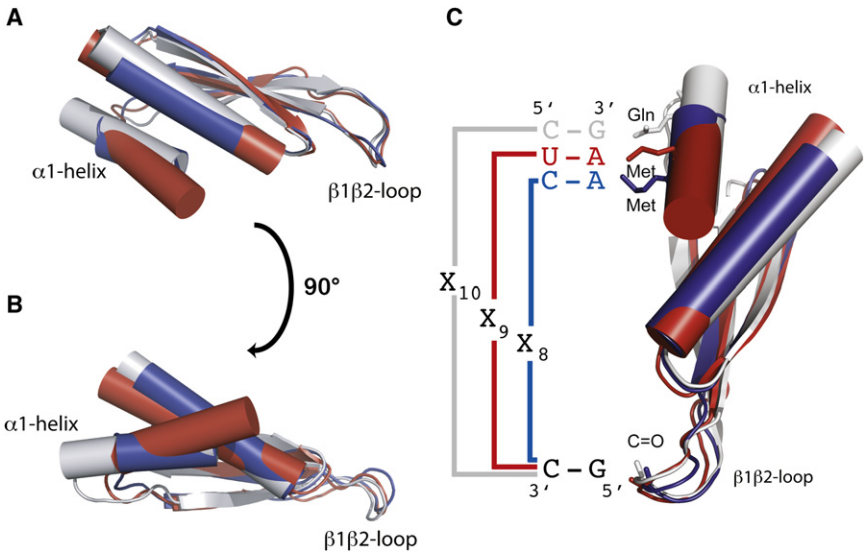


Figure 6. RNA Recognition Code of Various dsRBMs

(A) and (B) Overlay of the ADAR2 dsRBM1 (in blue), ADAR2 dsRBM2 (in red), and Aquifex aeolicus RNaseIII dsRBM (in gray) structures highlights the variability of helix α1 within the dsRBM fold and its importance for the determination of the register length between the two specific contacts on the RNA helix (C). For Aquifex aeolicus RNaseIII dsRBM–dsRNA interactions, see also Figure S5 and for sequence alignments of different dsRBMs, see also Figure S6.

binding sites on the GluR-2 R/G RNA stem-loop (considering a 32 base-pair stem, a 10 base-pair register between the two sequence-specific contacts and two possible orientations for the dsRBM), each dsRBM binds at a very specific register on this large RNA molecule. This binding is achieved by a direct readout of the RNA sequence in the minor groove of the A-form helix. The two dsRBMs of ADAR2 use helix α1 and the β1-β2 loop as molecular rulers to find their binding register in the RNA minor groove of the GluR-2 R/G RNA. Through the β1−β2 loop, the carbonyl oxygens of Val104 in dsRBM1 and Ser258 in dsRBM2 contact the amino groups of base-paired guanines, G22 and G9 respectively. The same type of sequence-specific RNA recognition of GC or GU base-pairs in the minor groove of RNA helices have been observed in several ribosomal proteins of the large subunit (Klein et al., 2004) and in some tRNA synthetases bound to RNA (Rould et al., 1989) although the fold of these proteins and the overall binding mode are different from a dsRBM. Through helix α1, the side-chain methyl groups of Met84 in dsRBM1 and of Met238 in dsRBM2 are in contact with the H2s of A32 and A18, respectively. Recognition of these two anchoring points in the minor-groove, separated by 9 and 8 base-pairs for dsRBM1 and dsRBM2, respectively, illustrates how the two dsRBMs find their sequence-specific binding registers, demonstrating that these dsRBMs have more sequence-specificity than previously thought. Interestingly, in each complex, one of the two anchoring points involves a mismatched base-pair (the G22-G50 base-pair for dsRBM1 and the A18-C54 base-pair for dsRBM2). It is therefore possible that the highly exposed amino or C2H2 groups of these mismatches in the minor groove further assist the dsRBMs of ADAR2 to find their binding register, supporting earlier findings that these two mismatches are important for positioning ADAR2 at the R/G site (Ohman et al., 2000). In addition to sequence-specific interactions between ADAR2 dsRBMs and its GluR-2 target, K127 (dsRBM1) and K281 (dsRBM2) make contacts with phosphate oxygens across the major groove of the RNA (Figure 2). These basic amino acid moieties are conserved in the loop between the β3 and α2 regions for all dsRBMs (Tian et al., 2004) and mutation of these residues in PKR and Staufen have been shown to ablate dsRNA-binding activity (McMillan et al., 1995; Ramos et al., 2000), indicating the importance of both sequence-specific and sequence–independent recognition of the RNA substrate for site-specific adenosine deamination.

Prior to this work, the structures of only four dsRBM-containing proteins in complex with RNA had been determined by X-ray crystallography (XlrbpA and *Aquifex aeolicus* (Aa) RNaseIII) or NMR spectroscopy (Staufen and Rnt1p; Gan et al., 2006; Ramos et al., 2000; Ryter and Schultz, 1998; Wu et al., 2004). In the two solution structures, the dsRBMs appear to recognize primarily the loop of the RNA while in the two crystal structures the dsRBMs are found bound across the junction between coaxially stacked helices. Lack of clear sequence-specific contacts led to the general opinion that dsRBMs are shape-specific rather than sequence-specific RNA binding domains (Stefl et al., 2005a). The two dsRBM-RNA complexes of ADAR2 reported here have revealed that dsRBMs recognize not only the shape of the RNA (a stem-loop for dsRBM1 and an A-form helix for dsRBM2), but also more surprisingly the sequence of the RNA. Interestingly, in a recent crystal structure of an Aa RNaseIII dsRBM bound to a stem-loop, sequence-specific contacts in the minor groove via helix α1 and the β1−β2 loop have been observed (Gan et al., 2008). The helix α1 in Aa RNaseIII is elongated by one turn compared to the helix α1 of the dsRBMs of ADAR2 and a Gln side-chain recognizes a guanine by two sequence-specific hydrogen bonds (Figure S5). The contact mediated by the β1−β2 loop in Aa RNaseIII are similar to the dsRBMs in ADAR2. The β1−β2 loop has the same length (six amino acids) and the main-chain carbonyl of the third residue of the loop is hydrogen bonded to a guanine amino of a GU base-pair. Despite similarities in the mode of binding, the three dsRBMs recognize different sequences and different register lengths. The dsRBM of Aa RNaseIII preferentially recognizes an RNA helix containing a G-X_{10}-G sequence while the dsRBM1 and dsRBM2 of ADAR2 preferentially recognize G-X_9-A and G-X_8-A sequences, respectively (Figure 6). The length and the positioning of helix α1 relative to the dsRBM fold appear to be the key structural elements that determine the register length of the different dsRBMs (Figure 6C).

Our findings regarding the RNA binding specificity of dsRBMs have important implications for the sequence-specificity paradox of ADAR2, but also of many other dsRBM-containing proteins that continue to puzzle investigators (Tian et al., 2004). Apparent differences in the sequences of dsRBMs between mammalian ADAR2 and ADAR1 (Figure S6), where ADAR1 dsRBMs appears to have a longer helix α1 and lack the ADAR2 equivalent of Met 84 and Met 238, could explain why ADAR1 and ADAR2 have different substrate specificities (Bass, 2002; Lehmann and Bass, 2000). Furthermore, our structure shows how dsRBM2 of ADAR2 binds the GluR-2 R/G site near the editing site in recognizing the amino group of the guanosine 3′ to the edited A. This would explain the strong preference for a guanosine moiety 3′ to the edited adenosine that is found in a great majority of substrates selectively edited by ADAR2 (Bass, 2002; Lehmann and Bass, 2000; Li et al., 2009; Riedmann et al., 2008) and more recently in long double-stranded RNA (Eggington and Bass, personal communication). This sequence preference disappears when the dsRBMs are deleted from ADAR2 (Eggington and Bass, personal communication) further supporting that this sequence requirement is due to dsRBM binding. Finally, in interacting with the guanosine 3′ to the edited adenosine and to the nucleotide that base-pairs with the editing site, dsRBM2 not only brings the deaminase domain in close proximity to the editing site, but also does not prevent access of the adenosine to the deaminase domain. When this precise positioning is impaired, specific editing is nearly abolished (see the effect of the S258A, H259A mutant) which emphasizes the functional importance of sequence-specific recognition of RNA by dsRBMs for A-to-I editing.

The sequence-specific contacts that we observed with the dsRBMs ADAR2 are interesting when comparing sequence alignments of several dsRBM structures that have been determined (Figure S6). This alignment reveals a surprisingly high variability in the length and amino acid sequence composition of the two regions of the dsRBMs mediating the sequence-specific interactions with the RNA, namely the helix α1 and the β1−β2 loop. This strongly suggests that dsRBMs are likely to have different binding specificity in agreement with reports indicating that dsRBMs from different proteins are not functionally interchangeable (Liu et al., 2000; Parker et al., 2008). Similar to ADAR2, many dsRBM-containing proteins involved in miRNA and siRNA processing and function are likely to bind RNA in a sequence-specific manner, that would modulate their target selection and mechanism of action. For example, DICER was shown to compete with ADARs for the same RNA substrates (Kawahara et al., 2007; Yang et al., 2006). Interestingly, ADARs modulate the processing of miRNA precursors not only by A-to-I modifications that alter the secondary structure of pri-miRNA (Kawahara et al., 2007; Tonkin and Bass, 2003; Yang et al., 2006), but also simply by RNA-binding alone to pri-miRNAs, as recently shown with catalytically inactive ADARs (Heale et al., 2009). This latter function for ADARs, as regulators of pri-miRNA processing, closely resemble that found for single-stranded sequence-specific RNA-binding proteins such as Lin28, hnRNP A1 or KSRP (Guil and Caceres, 2007; Heo et al., 2008, 2009; Michlewski et al., 2008; Newman et al., 2008; Trabucchi et al., 2009). Furthermore, RNAi activity has been shown to coincide with siRNA sequence motifs (Katoh and Suzuki, 2007). Altogether it is becoming clear that sequence-specific recognition mediated by dsRBMs is functionally important for dsRBM containing proteins. We have demonstrated here with ADAR2 how such sequence-specific recognition is mediated in dsRBMs and how this is relevant for RNA editing. Future work will be required to elucidate the variations in dsRNA-binding specificity and their functional relevance for numerous other members of the dsRBM-containing protein family.

EXPERIMENTAL PROCEDURES

Preparation of Proteins

Details on cloning, expression and purification of the ADAR2 dsRBM1, ADAR2 dsRBM2, and ADAR2 dsRBM12 constructs have been described previously (Stefl et al., 2005b, 2006).

NMR Spectroscopy

All NMR spectra were acquired at 310 K. Spectra were recorded at 500, 600, and 900 MHz Bruker spectrometers. All spectra were processed with XWINNMR or Topspin1.3/2.0 (Bruker BioSpin) and analyzed with Sparky 3.0 (Goddard T.G. and Kellner D.G., University of California, San Francisco). The ^{1}H, ^{13}C and ^{15}N chemical shifts of the protein in complex, were assigned by standard methods (Sattler et al., 1999). The ^{1}H-^{15}N HSQC and ^{1}H-^{13}C HSQC spectra of dsRBM1 and dsRBM2 in free and bound forms are shown in Figure S1 and Figure S2. All distance restraints were derived from 3D ^{15}N,^{13}C-edited NOESYs and 2D ^{1}H-^{1}H NOESY (t_m = 150 ms) collected at 900 MHz. RNA exchangeable proton resonances were assigned using ^{1}H-^{1}H NOESY spectrum (t_m = 200 ms) at 278 K. Nonexchangeable proton resonances were assigned using ^{1}H-^{1}H, NOESY, ^{1}H-^{1}H TOCSY, ^{1}H-^{13}C HSQC, 3D ^{13}C-edited NOESY, 2D ^{1}H-^{1}H double-half-filtered NOESY (t_m = 150 ms) (Peterson et al., 2004) and 3D ^{13}C F_1-edited, F_3-filtered NOESY-HSQC spectrum (t_m = 150 ms) (Zwahlen et al., 1997) in 99.99% $^{2}H_2O$ (v/v). The NOEs were semiquantitatively classified based on their intensities in the 2D and 3D NOESY spectra. Hydrogen bond distance restraints were used for base-pairs, when the imino-protons were observed experimentally. The assignments of intermolecular NOEs were based on 3D ^{13}C F_1-edited, F_3-filtered NOESY-HSQC spectrum (t_m = 150 ms), 2D ^{1}H-^{1}H F_1-^{13}C-filtered F_2-^{13}C-edited NOESY (t_m = 150 ms) on the protein-RNA complexes with either the protein or the RNA ^{13}C-^{15}N labeled. In case of dsRBM2–GluR-2 R/G LSL RNA complex, we observed an extra set of five weaker intermolecular nOes, which were discarded from structure calculation. These intermolecular restraints cannot be explained with the presented structure of dsRBM2-GluR-2 R/G LSL RNA complex. They originate from a minor conformation in which the protein is shifted up by one base pair toward the UUCG tetraloop.

Structure Calculation and Refinement

Distance constraints for the proteins bound to RNA where generated by the ATNOS/CANDID package (Herrmann et al., 2002). The accuracy of the list of automatically generated distance constraints was manually checked. Distance constraints for the free and bound RNAs as well as for the intermolecular NOEs were assigned manually. Preliminary structures of the free RNA and the protein-RNA complexes were obtained by a simulated annealing protocol in CYANA (Guntert et al., 1997; Herrmann et al., 2002). To impose better convergence of the ensemble, an artificial torsion angles for the canonical dsRNA regions were used as described previously (Oberstrass et al., 2006). Additional angle restraints to maintain proper local geometries were used (Tsui et al., 2000). The final refinement of all structures was performed using a 20 ps simulated annealing protocol in AMBER (Case et al., 2002) as described in the Supplemental Information. From 40 refined structures, the twenty conformers with the lowest AMBER energy were selected to form the final ensemble of structures. Structural quality was assessed using PROCHECK (Laskowski et al., 1996). Figures were prepared with MOLMOL (Koradi et al., 1996) and Pymol (DeLano, 2002).

Fluorescence Anisotropy

Fluorescence anisotropy was measured on a FluoroMax-4 spectrofluorometer (Horiba Jobin-Yvon, USA) equipped with a thermostated cell holder and an automatic titrator. All measurements were conducted in 50mM sodium phosphate buffer (pH 7.0) and 100mM NaCl at 10°C. To avoid any effects caused by 5′-end labeling of RNAs, the experiments were designed as a competition assay. At first, a reference measurement was carried out in which 1400 μl of 10nM fluorescein labeled wild-type RNA was titrated by the protein. Then, the same titration experiment was repeated in the presence of 500nM unlabeled RNA (either wild-type or mutants; Vasiljeva et al., 2008). Total volume of protein added to each reaction was 33 μl. The fitting was performed using DynaFit software (Kuzmic, 1996, 2006). Initially, the K_d for the reference protein–labeled RNA complex was determined. The obtained K_d value was then used as a fixed parameter when fitting the competition data. A 1:1 binding stoichiometry was assumed in all cases. The data were normalized for visualization purposes.

Quantitative Analysis of In Vitro RNA Editing

For in vitro editing reactions, a 116 nt RNA encoding a portion of the mouse GluR-2 pre-mRNA with the complete R/G duplex was transcribed in vitro (Stefl et al., 2006) and incubated with wild-type or mutant ADAR2 proteins derived from nuclear extracts obtained from transiently transfected HEK293 cells (Sansam et al., 2003). Equivalent amounts of wild-type and mutant ADAR2 protein, as determined by Western blotting, were incubated with 40 ng of the R/G transcript at 30°C for 20 min. These incubation conditions were determined empirically by performing time-course analyses with wild-type ADAR2 protein to ensure that the assay was in the linear range (*data not shown*). The reaction was stopped and the R/G transcript isolated by direct addition of TRI Reagent (Molecular Research Center) at the end of the incubation period. For quantification of RNA editing, the in vitro reaction product was reverse transcribed using AMV Reverse Transcriptase (Promega) and an antisense primer (5′-CGGCCAATCGTACGTACCTCCGGCCGAATTCTACAAACC GTTAAGAGTCTTA-3′) with a unique 5′-extension (underlined). The resulting amplicon was diluted 1:1000 and 1 μl was subsequently amplified by PCR using sense (5′-CCGGGAGCTCATCGCCACACCTAAAGGATCC-3′) and antisense (5′-CGGCCAATCGTACGTACCTCC-3′) primers corresponding to GluR-2 and the unique 5′-extension sequences, respectively. PCR amplicons were purified using the Wizard SV PCR and Gel Cleanup System (Promega) and digested with Mse I (New England Biolabs) to generate 100 and 70 bp products representing edited and nonedited transcripts, respectively. The resulting digestion products were resolved on a 4% Agarose gel and editing efficiency was quantified by phosphorimager analysis (GE Healthcare).

In vitro editing reactions using GluR-2 R/G mutant RNAs were performed as described above with equivalent amounts of wild-type ADAR2 protein derived from nuclear extracts obtained from transiently transfected HEK293 cells (Sansam et al., 2003). Wild-type and mutant transcripts were trace labeled with [α-^{32}P]-UTP and concentrations of in vitro transcribed RNAs were determined using a Perkin-Elmer Tri-Carb 2800TR scintillation spectrometer based upon the calculated specific activity for each transcript.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures and six figures and can be found with this article online at doi:10.1016/j.cell.2010.09.026.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation (Nr. 3100A0-118118) and the SNF-NCCR structural biology to F.H.T.A and the National Institutes of Health (R01 NS33323) to R.B.E. R.S. is supported by the Ministry of Education of the Czech Republic (MSM0021622413, Ingo LA08008), GACR (204/08/1212, 305/10/1490), GAAV (IAA401630903), HHMI/EMBO start-up grant, and HFSP Career Development Award. M.Z. and C.H. are supported by GACR (204/08/H054) and by the Ministry of Education of the Czech Republic (MSM0021622415). M.Z. is in receipt of a Brno City Scholarship for Talented Ph.D. Students. The coordinates of the structures of GluR-2 LSL RNA, ADAR2 dsRBM1 bound to GluR-2 USL RNA, ADAR2 dsRBM2 bound to GluR-2 LSL RNA and ADAR2 dsRBM12 bound to GluR-2 have been deposited in the Protein Data Bank with accession codes 2l2j, 2l3c, 2l2k, and 2l3j, respectively.

Received: September 21, 2009
Revised: May 26, 2010
Accepted: August 30, 2010
Published: October 14, 2010

REFERENCES

Aruscavage, P.J., and Bass, B.L. (2000). A phylogenetic analysis reveals an unusual sequence conservation within introns involved in RNA editing. RNA *6*, 257–269.

Bass, B.L. (2002). RNA editing by adenosine deaminases that act on RNA. Annu. Rev. Biochem. *71*, 817–846.

Blaszczyk, J., Gan, J., Tropea, J.E., Court, D.L., Waugh, D.S., and Ji, X. (2004). Noncatalytic assembly of ribonuclease III with double-stranded RNA. Structure *12*, 457–466.

Case, D.A., Pearlman, D.A., Caldwell, J.W., Cheatham, T.E., III, Wang, J., Ross, W.S., Simmerling, C.L., Darden, T.A., Merz, K.M., Stanton, R.V., et al. (2002). AMBER 7 (San Francisco: University of California).

Chilibeck, K.A., Wu, T., Liang, C., Schellenberg, M.J., Gesner, E.M., Lynch, J.M., and MacMillan, A.M. (2006). FRET analysis of in vivo dimerization by RNA-editing enzymes. J. Biol. Chem. *281*, 16530–16535.

Cho, D.S., Yang, W., Lee, J.T., Shiekhattar, R., Murray, J.M., and Nishikura, K. (2003). Requirement of dimerization for RNA editing activity of adenosine deaminases acting on RNA. J. Biol. Chem. *278*, 17093–17102.

DeLano, W.L. (2002). The PyMOL Molecular Graphics System (Palo Alto, CA, USA: DeLano Scientific).

Dreyfuss, G., Kim, V.N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. Nat. Rev. Mol. Cell Biol. *3*, 195–205.

Gallo, A., Keegan, L.P., Ring, G.M., and O'Connell, M.A. (2003). An ADAR that edits transcripts encoding ion channel subunits functions as a dimer. EMBO J. *22*, 3421–3430.

Gan, J., Shaw, G., Tropea, J.E., Waugh, D.S., Court, D.L., and Ji, X. (2008). A stepwise model for double-stranded RNA processing by ribonuclease III. Mol. Microbiol. *67*, 143–154.

Gan, J., Tropea, J.E., Austin, B.P., Court, D.L., Waugh, D.S., and Ji, X. (2006). Structural insight into the mechanism of double-stranded RNA processing by ribonuclease III. Cell *124*, 355–366.

Guil, S., and Caceres, J.F. (2007). The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. Nat. Struct. Mol. Biol. *14*, 591–596.

Guntert, P., Mumenthaler, C., and Wuthrich, K. (1997). Torsion angle dynamics for NMR structure calculation with the new program DYANA. J. Mol. Biol. *273*, 283–298.

Heale, B.S., Keegan, L.P., McGurk, L., Michlewski, G., Brindle, J., Stanton, C.M., Caceres, J.F., and O'Connell, M.A. (2009). Editing independent effects of ADARs on the miRNA/siRNA pathways (EMBO J.).

Heo, I., Joo, C., Cho, J., Ha, M., Han, J., and Kim, V.N. (2008). Lin28 mediates the terminal uridylation of let-7 precursor MicroRNA. Mol. Cell *32*, 276–284.

Heo, I., Joo, C., Kim, Y.K., Ha, M., Yoon, M.J., Cho, J., Yeom, K.H., Han, J., and Kim, V.N. (2009). TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation. Cell *138*, 696–708.

Herrmann, T., Guntert, P., and 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.

Higuchi, M., Maas, S., Single, F.N., Hartner, J., Rozov, A., Burnashev, N., Feldmeyer, D., Sprengel, R., and Seeburg, P.H. (2000). Point mutation in an AMPA

receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. Nature *406*, 78–81.

Higuchi, M., Single, F.N., Kohler, M., Sommer, B., Sprengel, R., and Seeburg, P.H. (1993). RNA editing of AMPA receptor subunit GluR-B: a base-paired intron-exon structure determines position and efficiency. Cell *75*, 1361–1370.

Katoh, T., and Suzuki, T. (2007). Specific residues at every third position of siRNA shape its efficient RNAi activity. Nucleic Acids Res. *35*, e27.

Kawahara, Y., Zinshteyn, B., Sethupathy, P., Iizasa, H., Hatzigeorgiou, A.G., and Nishikura, K. (2007). Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. Science *315*, 1137–1140.

Klein, D.J., Moore, P.B., and Steitz, T.A. (2004). The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. J. Mol. Biol. *340*, 141–177.

Koradi, R., Billeter, M., and Wuthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph *14*, 51–55, 29-32.

Kuzmic, P. (1996). Program DYNAFIT for the analysis of enzyme kinetic data: application to HIV proteinase. Anal. Biochem. *237*, 260–273.

Kuzmic, P. (2006). A generalized numerical approach to rapid-equilibrium enzyme kinetics: application to 17beta-HSD. Mol. Cell. Endocrinol. *248*, 172–181.

Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R., and Thornton, J.M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J. Biomol. NMR *8*, 477–486.

Legault, P., and Pardi, A. (1994). In-Situ Probing of Adenine Protonation in Rna by C-13 Nmr. J. Am. Chem. Soc. *116*, 8390–8391.

Lehmann, K.A., and Bass, B.L. (2000). Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. Biochemistry *39*, 12875–12884.

Li, J.B., Levanon, E.Y., Yoon, J.K., Aach, J., Xie, B., Leproust, E., Zhang, K., Gao, Y., and Church, G.M. (2009). Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science *324*, 1210–1213.

Liu, Y., Lei, M., and Samuel, C.E. (2000). Chimeric double-stranded RNA-specific adenosine deaminase ADAR1 proteins reveal functional selectivity of double-stranded RNA-binding domains from ADAR1 and protein kinase PKR. Proc. Natl. Acad. Sci. USA *97*, 12541–12546.

Lomeli, H., Mosbacher, J., Melcher, T., Hoger, T., Geiger, J.R., Kuner, T., Monyer, H., Higuchi, M., Bach, A., and Seeburg, P.H. (1994). Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. Science *266*, 1709–1713.

Macbeth, M.R., Lingam, A.T., and Bass, B.L. (2004). Evidence for auto-inhibition by the N terminus of hADAR2 and activation by dsRNA binding. RNA *10*, 1563–1571.

Macbeth, M.R., Schubert, H.L., Vandemark, A.P., Lingam, A.T., Hill, C.P., and Bass, B.L. (2005). Inositol hexakisphosphate is bound in the ADAR2 core and required for RNA editing. Science *309*, 1534–1539.

McMillan, N.A., Carpick, B.W., Hollis, B., Toone, W.M., Zamanian-Daryoush, M., and Williams, B.R. (1995). Mutational analysis of the double-stranded RNA (dsRNA) binding domain of the dsRNA-activated protein kinase, PKR. J. Biol. Chem. *270*, 2601–2606.

Melcher, T., Maas, S., Herb, A., Sprengel, R., Seeburg, P.H., and Higuchi, M. (1996). A mammalian RNA editing enzyme. Nature *379*, 460–464.

Michlewski, G., Guil, S., Semple, C.A., and Caceres, J.F. (2008). Posttranscriptional regulation of miRNAs harboring conserved terminal loops. Mol. Cell *32*, 383–393.

Newman, M.A., Thomson, J.M., and Hammond, S.M. (2008). Lin-28 interaction with the Let-7 precursor loop mediates regulated microRNA processing. RNA *14*, 1539–1549.

Nishikura, K. (2006). Editor meets silencer: crosstalk between RNA editing and RNA interference. Nat. Rev. Mol. Cell Biol. *7*, 919–931.

Oberstrass, F.C., Lee, A., Stefl, R., Janis, M., Chanfreau, G., and Allain, F.H. (2006). Shape-specific recognition in the structure of the Vts1p SAM domain with RNA. Nat. Struct. Mol. Biol. *13*, 160–167.

Ohman, M., Kallman, A.M., and Bass, B.L. (2000). In vitro analysis of the binding of ADAR2 to the pre-mRNA encoding the GluR-B R/G site. RNA *6*, 687–697.

Palladino, M.J., Keegan, L.P., O'Connell, M.A., and Reenan, R.A. (2000). A-to-I pre-mRNA editing in Drosophila is primarily involved in adult nervous system function and integrity. Cell *102*, 437–449.

Parker, G.S., Maity, T.S., and Bass, B.L. (2008). dsRNA binding properties of RDE-4 and TRBP reflect their distinct roles in RNAi. J. Mol. Biol. *384*, 967–979.

Peterson, R.D., Theimer, C.A., Wu, H., and 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.

Ramos, A., Grunert, S., Adams, J., Micklem, D.R., Proctor, M.R., Freund, S., Bycroft, M., St Johnston, D., and Varani, G. (2000). RNA recognition by a Staufen double-stranded RNA-binding domain. EMBO J. *19*, 997–1009.

Riedmann, E.M., Schopoff, S., Hartner, J.C., and Jantsch, M.F. (2008). Specificity of ADAR-mediated RNA editing in newly identified targets. RNA *14*, 1110–1118.

Rould, M.A., Perona, J.J., Soll, D., and Steitz, T.A. (1989). Structure of E. coli glutaminyl-tRNA synthetase complexed with tRNA(Gln) and ATP at 2.8 A resolution. Science *246*, 1135–1142.

Ryter, J.M., and Schultz, S.C. (1998). Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. EMBO J. *17*, 7505–7513.

Sansam, C.L., Wells, K.S., and Emeson, R.B. (2003). Modulation of RNA editing by functional nucleolar sequestration of ADAR2. Proc. Natl. Acad. Sci. USA *100*, 14018–14023.

Sattler, M., Schleucher, J., and Griesinger, C. (1999). Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Prog. Nucl. Magn. Reson. Spectrosc. *34*, 93–158.

Stefl, R., and Allain, F.H. (2005). A novel RNA pentaloop fold involved in targeting ADAR2. RNA *11*, 592–597.

Stefl, R., Skrisovska, L., and Allain, F.H. (2005a). RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. EMBO Rep. *6*, 33–38.

Stefl, R., Skrisovska, L., Xu, M., Emeson, R.B., and Allain, F.H. (2005b). Resonance assignments of the double-stranded RNA-binding domains of adenosine deaminase acting on RNA 2 (ADAR2). J. Biomol. NMR *31*, 71–72.

Stefl, R., Xu, M., Skrisovska, L., Emeson, R.B., and Allain, F.H. (2006). Structure and specific RNA binding of ADAR2 double-stranded RNA binding motifs. Structure *14*, 345–355.

Tian, B., Bevilacqua, P.C., Diegelman-Parente, A., and Mathews, M.B. (2004). The double-stranded-RNA-binding motif: interference and much more. Nat. Rev. Mol. Cell Biol. *5*, 1013–1023.

Tonkin, L.A., and Bass, B.L. (2003). Mutations in RNAi rescue aberrant chemotaxis of ADAR mutants. Science *302*, 1725.

Trabucchi, M., Briata, P., Garcia-Mayoral, M., Haase, A.D., Filipowicz, W., Ramos, A., Gherzi, R., and Rosenfeld, M.G. (2009). The RNA-binding protein KSRP promotes the biogenesis of a subset of microRNAs. Nature *459*, 1010–1014.

Tsui, V., Zhu, L., Huang, T.H., Wright, P.E., and Case, D.A. (2000). Assessment of zinc finger orientations by residual dipolar coupling constants. J. Biomol. NMR *16*, 9–21.

Valente, L., and Nishikura, K. (2005). ADAR gene family and A-to-I RNA editing: diverse roles in posttranscriptional gene regulation. Prog. Nucleic Acid Res. Mol. Biol. *79*, 299–338.

Valente, L., and Nishikura, K. (2007). RNA binding-independent dimerization of adenosine deaminases acting on RNA and dominant negative effects of nonfunctional subunits on dimer functions. J. Biol. Chem. *282*, 16054–16061.

Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S., and Meinhart, A. (2008). The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. Nat. Struct. Mol. Biol. *15*, 795–804.

Wu, H., Henras, A., Chanfreau, G., and Feigon, J. (2004). Structural basis for recognition of the AGNN tetraloop RNA fold by the double-stranded RNA-binding domain of Rnt1p RNase III. Proc. Natl. Acad. Sci. USA *101*, 8307–8312.

Xu, M., Wells, K.S., and Emeson, R.B. (2006). Substrate-dependent contribution of double-stranded RNA-binding motifs to ADAR2 function. Mol. Biol. Cell *17*, 3211–3220.

Yang, W., Chendrimada, T.P., Wang, Q., Higuchi, M., Seeburg, P.H., Shiekhattar, R., and Nishikura, K. (2006). Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. Nat. Struct. Mol. Biol. *13*, 13–21.

Zwahlen, C., Legault, P., Vincent, S.J.F., Greenblatt, J., Konrat, R., and Kay, L.E. (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.

Biomol NMR Assign
DOI 10.1007/s12104-010-9257-2

ARTICLE

^{1}H, ^{13}C, and ^{15}N chemical shift assignments of ZCCHC9

Maria Sanudo · Martin Jacko · Christiane Rammelt · Stepanka Vanacova · Richard Stefl

Received: 28 May 2010 / Accepted: 2 August 2010

Abstract ZCCHC9 is a human nuclear protein with sequence homology to yeast Air1p/Air2p proteins which are RNA-binding subunits of the Trf4/Air2/Mtr4 polyadenylation (TRAMP) complex involved in nuclear RNA quality control and degradation in yeast. The ZCCHC9 protein contains four retroviral-type zinc knuckle motifs. Here, we report the NMR spectral assignment of the zinc knuckle region of ZCCHC9. These data will allow performing NMR structural and RNA-binding studies of ZCCHC9 with the aim to investigate its role in the RNA quality control in human.

Keywords ZCCHC9 · Zinc knuckle · CCHC · RNA degradation

Maria Sanudo and Martin Jacko these authors contributed equally to this work.

M. Sanudo · M. Jacko · S. Vanacova (✉) · R. Stefl (✉)
National Centre for Biomolecular Research, Faculty of Science, Masaryk University, 62500 Brno, Czechia
e-mail: vanacova@chemi.muni.cz

R. Stefl
e-mail: stefl@chemi.muni.cz

C. Rammelt
Biozentrum, University of Basel, Klingelbergstrasse 50/70, 4057 Basel, Switzerland

Present Address:
C. Rammelt
University of Halle, Institute for Biochemistry, Kurt-Mothes-Strasse 3, 06120 Halle, Germany

Biological context

RNA processing in eukaryotes has to undergo a strict RNA quality control. In yeast nucleus, aberrant or short-lived RNAs are degraded by the TRAMP-exosome pathway. TRAMP is a poly(A) polymerase complex consisting of the Trf4 or Trf5 non-canonical polymerases, the Air1 or Air2 retroviral-type zinc knuckle proteins as potential RNA-binding subunits and the Mtr4 RNA helicase (Vanacova and Stefl 2007). The complex acts on diverse classes of RNAs (San Paolo et al. 2009; Vanacova and Stefl 2007). It can add short poly(A) tails to 3′end of the RNA molecules to recruit the nuclear exosome nucleolytic activities (Vanacova and Stefl 2007). However, in vivo, poly(A) polymerase activity is dispensable for the degradation of most TRAMP4 RNA targets (San Paolo et al. 2009). The minimal TRAMP poly(A) polymerase consists of Trf4/Trf5 and Air1 or Air2 heteromeric complexes (Vanacova and Stefl 2007). It has been proposed that Air1p and Air2p, respectively support RNA recognition and binding while Trf4/5 proteins provide the polyadenylation activity (Vanacova and Stefl 2007). Air1/2 contain 4–5 CCHC zinc knuckle motifs. Zinc knuckle motifs of other proteins are known to contact RNA molecules (D'Souza and Summers 2004).

Human genome encodes for at least 22 zinc knuckle proteins of the CCHC-type (identified by HUGO Gene Nomenclature Committee), in most cases rather poorly characterized. Our BLAST search identified three closest homologues of Air1/2p; the ZCCHC7, ZCCHC3, ZCCHC9 proteins. No relevant functional information have been reported on ZCCHC7 nor ZCCHC3 to date and only little is known about ZCCHC9. ZCCHC9 (zinc finger, CCHC domain containing 9; alias DKFZp761J139) is a nucleolar protein (Zhou et al. 2008). Based on mRNA expression

Published online: 14 August 2010

level profiling, it shows the highest expression levels in mouse brain, particularly cortex, and testicles. Genome-wide expression profiling study indicated ZCCHC9 up-regulation in the human neck cancers cells both HPV (human papillomaviruses) positive or negative and cervical cancer cells relative to normal cells (Pyeon et al. 2007). Overexpression of ZCCHC9 in HEK293 cells lead to the suppression of the transactivation by NF-kappa B and SRE, therefore ZCCHC9 has been implicated to be involved in the MAPK signaling cascade. However, the mechanism of ZCCHC9 function remains elusive.

ZCCHC9 contains 4 retroviral-type zinc knuckle motifs with consensus C–X_2–C–X_4–H–X_4-C that is found mainly in the nucleocapsid protein of retroviruses and also in eukaryotic proteins involved in RNA or ssDNA binding (D'Souza and Summers 2005). To gain structural and functional insights into the mechanism of ZCCHC9 MAPK regulation and the involvement of ZCCHC9 in RNA metabolism, we have initiated an NMR study of ZCCHC9. Here, we report ^{1}H, ^{13}C, and ^{15}N resonance assignments of ZCCHC9.

Methods and experiments

The coding sequence corresponding to the zinc knuckle region of the ZCCHC9 gene (accession number BC014841, Ref. Seq. ID NM_032280) from *Homo sapiens* (318 bp) was amplified by PCR and cloned into a pET30-G protein B1 fusion vector via BamHI and XhoI restriction sites (Zhou et al. 2001). A non-cleavable solubility-enhancement tag, GB1, was fused to the ZCCHC9 protein to enhance its solubility limit (Zhou et al. 2001), yielding a 175 amino acid construct (GB1 (58 aa) + ZCCHC9 (106 aa) + linker (5 aa) + hexahistidine tag). This tag is often used for NMR and X-ray studies of poorly behaving proteins (Zhou et al. 2001). The resulting N-terminal GB1 and C-terminal 6× histidine-tagged construct was verified by DNA sequencing. The protein was overexpressed in *E.coli* BL21-CodonPlus(DE3)-RIPL (Stratagene) at 37°C in M9 minimal medium, supplemented with 50 μM $ZnSO_4$. For isotope labeling, the medium was supplemented with $^{15}NH_4Cl$ and U-$^{13}C_6$-glucose. Cells were grown at 37°C to OD_{600} ~0.5 and induced with 400 μM isopropyl-β-D-thiogalactoside (IPTG) for 6 h at 37°C. Cells were harvested by centrifugation (4,000 rpm, 10 min, 4°C), resuspended in lysis buffer (20 mM Tris pH 7.0; 400 mM NaCl; 100 μM $ZnSO_4$; 10% glycerol; 25% sacharose; 2 mM $MgCl_2$; 5 mM DTT; 500 μM PMSF; 2 μg/ml leupeptin; 1 μg/ml pepstatin) and disrupted by sonication. An equal volume of buffer A (20 mM Tris pH 7.0, 400 mM NaCl, 100 μM $ZnSO_4$, 10% glycerol, 1% Nonidet P40, 40 mM imidazole, 5 mM DTT, 500 μM PMSF, 2 μg/ml leupeptin, 1 μg/ml pepstatin) was added to the lysate. The cell debris was cleared by centrifugation (14 000 rpm, 45 min, 4°C). The supernatant was next incubated for 45 min at 4°C with Ni-NTA agarose beads (QIAGEN), which were pre-equilibrated with a buffer B (20 mM Tris pH 7.0, 400 mM NaCl, 20 mM imidazole). The beads were washed with 10 volumes of wash buffer B and the protein was eluted in buffer EB (20 mM Tris pH 7.0, 400 mM NaCl, 100 μM $ZnSO_4$, 10% glycerol, 5 mM DTT, 1 mM PMSF, 400 mM imidazole). The purified protein was dialyzed to buffer C (20 mM Tris pH 7.0, 200 mM NaCl, 100 μM $ZnSO_4$, 14 mM 2-mercaptoethanol) and concentrated to ~1.5 mM in 550 μl of buffer C.

All NMR experiments were recorded at 293 K on a Bruker AVANCE 600 MHz spectrometer equipped with a cryogenic triple resonance probe (Bruker BioSpin). The NMR spectra were processed using the TopSpin 2.1 (Bruker BioSpin) and NMRPipe (Delaglio et al. 1995). The chemical shifts of backbone and side-chains were assigned manually using Sparky software (Goddard T.G. and Kellner D.G., University of California, San Francisco). For backbone resonance assignment, the following standard triple resonance experiments were carried out: HNCA, HNCACO, HNCACB (Sattler et al. 1999). For side-chain resonance assignment, 2D ^{1}H-^{13}C HMQC, 3D ^{1}H-^{13}C NOESY-HSQC, ^{1}H-^{15}N NOESY-HSQC (both NOESY with mixing time of 150 ms), 3D HNHA, 3D HCCH-TOCSY and 2D HB(CB)(CGCD)HD, were recorded (Sattler et al. 1999).

Extent of assignments and data deposition

The 2D ^{1}H-^{15}N HSQC spectrum of GB1-ZCCHC9 is shown in Fig. 1 with indicated assignments. With standard sequential assignment procedure, near complete assignments of the backbone were obtained. For the entire protein construct, 98% of backbone amide protons and ^{15}N nuclei, 98% of the Cα and Cβ were assigned; prolines and M1 not counted. The chemical shift deviations of Cα and Hα of the assigned backbone resonances of GB1-ZCCHC9 from the sequence-dependent random coil values (data not shown) agree well with the canonical secondary structure elements of the GB1 domain ($\beta1\beta2\alpha1\beta3\beta4$; Gronenborn et al. 1991) and show no pattern for the ZCCHC9 region. The latter agrees well with the absence of secondary structure elements in the topology of zinc knuckles (D'Souza and Summers 2004). The histidines of CC**H**C motif of zinc knuckles (H71, H98, H127, and H154) are involved in coordination of zinc ions as the N-H correlations of these residues span the same region in the 2D ^{1}H-^{15}N HSQC (Fig. 1) that is characteristic for zinc fingers and knuckles. The histidine C2 chemical shifts that are sensitive probes to

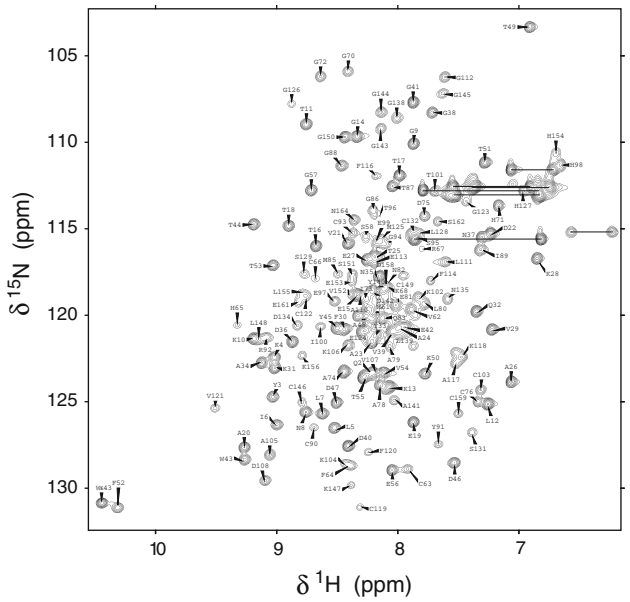


Fig. 1 2D [^{1}H-^{15}N]-HSQC spectrum of 1.5 mM uniformly [^{15}N,^{13}C]-labeled ZCCHC9 in 20 mM Tris buffer (pH 7.0), 200 mM NaCl, 100 μM $ZnSO_4$, 14 mM 2-mercaptoethanol, (90% H_2O/10% D_2O). The spectrum was acquired at 293 K on a Bruker Avance 600 MHz spectrometer. The assignments are *labeled* by the one-letter code of amino acids accompanied by a sequence number. The side-chain resonances of asparagine and glutamine are connected by *horizontal lines*

monitor the histidine protonation state, indicate that the histidines are not protonated under the conditions used in our study (pH 7). Further, the presence of folded zinc knuckles was confirmed by the addition of EDTA in excess of zinc ions that resulted in a loss of dispersion of many resonances in the 2D ^{1}H-^{15}N HSQC (data not shown), and consequently, in significant precipitation of the protein. All assignment for ^{1}H, ^{15}N and ^{13}C backbone and side-chain chemical shifts of ZCCHC9 have been deposited at the BMRB (http://www.bmrb.wisc.edu) and can be accessed under the accession number 16958.

Acknowledgments This work was supported by Wellcome Trust International Senior Fellowship 084316/Z/07/Z (to S.V.), EMBO Installation Grant Project No. 1642 (to S.V.), European Molecular Biology Organization/Howard Hughes Medical Institute Start-up Grant No. 1263 (to R.S.), and Human Frontier Science Program Career Development Award No. 0049/2006 (to R.S.), Ministry of Education of the Czech Republic (MSM0021622413, Ingo LA08008), GACR (204/08/1212 and 305/10/1490), GAAV (IAA401630903). MS is supported by the EC FP-7 (grant no. 205872).

References

D'Souza V, Summers MF (2004) Structural basis for packaging the dimeric genome of Moloney murine leukaemia virus. Nature 431: 586–590

D'Souza V, Summers MF (2005) How retroviruses select their genomes. Nat Rev Microbiol 3:643–655

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

Gronenborn AM, Filpula DR, Essig NZ, Achari A, Whitlow M, Wingfield PT, Clore GM (1991) A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein G. Science 253:657–661

Pyeon D, Newton MA, Lambert PF, den Boon JA, Sengupta S, Marsit CJ, Woodworth CD, Connor JP, Haugen TH, Smith EM et al (2007) Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers. Cancer Res 67:4605–4619

San Paolo S, Vanacova S, Schenk L, Scherrer T, Blank D, Keller W, Gerber AP (2009) Distinct roles of non-canonical poly(A) polymerases in RNA metabolism. PLoS Genet 5:e1000555

Sattler M, Schleucher J, Griesinger C (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Prog Nucl Magn Reson Spectrosc 34:93–158

Vanacova S, Stefl R (2007) The exosome and RNA quality control in the nucleus. EMBO Rep 8:651–657

Zhou A, Zhou J, Yang L, Liu M, Li H, Xu S, Han M, Zhang J (2008) A nuclear localized protein ZCCHC9 is expressed in cerebral cortex and suppresses the MAPK signal pathway. J Genet Genomics 35:467–472

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

Biomol NMR Assign (2010) 4:119–121
DOI 10.1007/s12104-010-9221-1

ARTICLE

^{1}H, ^{13}C, and ^{15}N chemical shift assignments for the RNA recognition motif of Nab3

Roberto Pergoli · Karel Kubicek · Fruzsina Hobor · Josef Pasulka · Richard Stefl

Received: 13 November 2009 / Accepted: 10 March 2010 / Published online: 23 March 2010

Abstract Nuclear polyadenylated RNA-binding (Nab)3 protein is an RNA-binding protein that is involved in the poly(A) independent termination pathway. Here, we report the NMR spectral assignments of RNA-recognition motif (RRM) of Nab3. The assignment will allow performing NMR structural and RNA-binding studies of Nab3 with the aim to investigate its role in the poly(A) independent termination pathway.

Keywords Nab3 · RNA-binding motif · RRM · Transcription termination

Biological context

RNA polymerase II transcripts that include small nuclear and small nucleoar RNAs, and a class of intergenic and anti-sense RNAs are terminated via poly(A)-independent termination pathway (Arigo et al. 2006; Thiebaut et al. 2006). Termination of these transcripts requires the nuclear pre-mRNA down-regulation (Nrd)1 and the nuclear polyadenylated RNA-binding (Nab)3 proteins, the RNA helicase Sen1 (Steinmetz et al. 2001). Nrd1 contains a conserved CTD-interaction domain that preferentially binds C-terminal domain (CTD) of RNA polymerase II phosphorylated at Ser5 (Gudipati et al. 2008; Vasiljeva et al. 2008). In addition, a sequence-specific RNA binding of both Nrd1 and Nab3 (via their RNA recognition motifs (RRMs)) is critical for this termination pathway. Further, the exosome, a complex of 10–12 exoribonucleolytic and RNA-binding proteins (Vanacova and Stefl 2007) and the Trf4–Air2–Mtr4 polyadenylation (TRAMP) complex (Vanacova and Stefl 2007) are involved in the 3′-end processing of these transcripts (Vanacova and Stefl 2007).

Recent studies identified an RNA sequence motif, UCUU, that is recognized by Nab3 and is considered as one of the signals that direct termination and exosome-TRAMP trimming/degradation of nonpolyadenylated transcripts (Carroll et al. 2007; Steinmetz and Brow 1998). It was shown that the RNA-recognition motif (RRM) of Nab3 binds to the UCUU sequence. The RRM is the most common RNA-binding motif. It has a typical $\beta\alpha\beta\beta\alpha\beta$ topology that forms a four-stranded β-sheet packed against two α-helices (Stefl et al. 2005). The solved structures of RRM-containing proteins in complex with RNA reveal the complexity of protein–RNA recognition mediated by the RRM, which often involves not only protein–RNA interactions but also RNA–RNA and protein–protein interactions. (Stefl et al. 2005; Maris et al. 2005). In terms of sequence identity, the RRM of Nab3 is closest to the RRM of HNRPC (37% sequence identity). However, these two RRMs bind to different RNA targets. Interestingly, Nab3 RRM binds the same RNA sequence (UCUU) as the RRMs of Polypyrimidine tract-binding protein (PTB) despite the fact they have extremely low sequence identity (3%).

To gain structural and mechanistic insights into the poly(A)-independent termination pathway, we have initiated an NMR study of the RRM of Nab3. Here, we report ^{1}H, ^{13}C, and ^{15}N resonance assignment of Nab3 RRM.

Roberto Pergoli, Karel Kubicek contributed equally to this work.

R. Pergoli · F. Hobor · J. Pasulka · R. Stefl (✉)
National Centre for Biomolecular Research, Faculty of Science, Masaryk University, 62500 Brno, Czechia
e-mail: rchrdstfl@gmail.com

K. Kubicek
Department of Condensed Matter Physics, Faculty of Science, Masaryk University, 62500 Brno, Czechia

Methods and experiments

The coding sequence corresponding to the RRM of the Nab3 gene from *Saccharomyces cerevisiae* (312 bp) was amplified by polymerase chain reaction (PCR), and cloned into a pET22b expression vector (Novagen) via NdeI and XhoI restriction sites. The resulting C-terminal 6× histidine-tagged construct was verified by DNA sequencing. The protein was overexpressed in *E.coli* BL21-Codon Plus (DE3)-RIPL, transformed with the pET22b-RRM Nab3 construct at 37°C in M9 minimal medium. For isotope labeling, the medium was supplemented with $^{15}NH_4Cl$ and U-$^{13}C_6$-glucose. Cells were grown at 37°C to OD_{600} ~1 and induced with 1 mM isopropyl β-D-thiogalactoside (IPTG). Cells were harvested by centrifugation (6,000 rpm, 10 min), resuspended in lysis buffer (50 mM Na_2HPO_4, 300 mM NaCl, 10 mM β-mercaptoethanol, pH 8), and disrupted by sonication. The cell debris was cleared by centrifugation and supernatant was applied to Ni-NTA affinity column (QIAGEN) and further purified on a Superdex 75 gel filtration column (GE Healthcare). The purified protein was concentrated to 2.5 mM in 550 µl of 50 mM sodium phosphate (pH 8.0), containing 300 mM NaCl and 10 mM β-mercaptoethanol.

All NMR experiments were conducted at 30°C on a Bruker AVANCE 600 and 900 MHz spectrometers equipped with a cryogenic triple resonance probes (Bruker BioSpin). For backbone resonance assignment, the following standard triple resonance experiments were carried out: HNCA, HNCACO, HNCACB, CBCACONH (Sattler et al. 1999). For side-chain resonance assignment, 2D ^{1}H-^{13}C HMQC, 3D ^{1}H-^{13}C NOESY-HSQC, ^{1}H-^{15}N NOESY-HSQC (both NOESY with mixing time of 120 ms), 3D HNHA, 3D HCCH-TOCSY and 2D HB(CB)(CGCD)HD, were recorded (Sattler et al. 1999). NMR spectra were processed using the TopSpin 2.1 (Bruker BioSpin) and NMRPipe (Delaglio et al. 1995). The chemical shifts of backbone and side-chains were assigned manually using Sparky software (Goddard and Kneller 2004).

Extent of assignments and data deposition

Following a standard sequential assignment procedure, near complete assignments could be obtained for all observable resonances. The 2D ^{1}H-^{15}N HSQC spectrum of Nab3 RRM is shown in Fig. 1 with indicated assignment. For the structured domain, 98% of backbone amide protons and ^{15}N nuclei (prolines not counted), 89% of the Cα and 91% of the Cβ, and 76% of ^{1}H, ^{13}C, and ^{15}N side-chain resonances were assigned. The chemical shift deviations of Cα and Hα of the assigned backbone resonances of Nab3 RRM from the sequence-dependent random coil values are shown in Fig. 2. These data agree well with the canonical secondary structure elements of the RRM domain family ($\beta1\alpha1\beta2\beta3\alpha2\beta4$; Maris et al. 2005; Stefl et al. 2005), and indicate no additional topological extension, that often occurs in the RRM domain family (Maris et al. 2005; Stefl

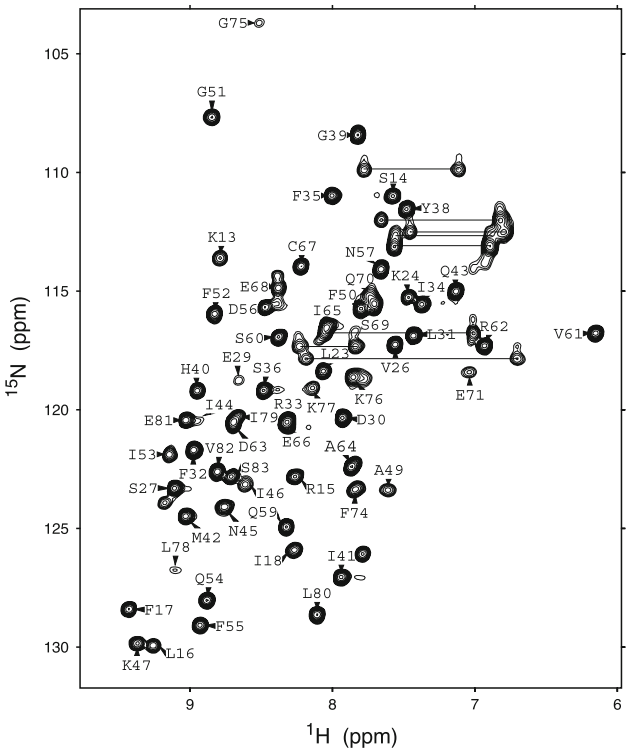


Fig. 1 2D [^{1}H-^{15}N]-HSQC spectrum of 2.5 mM uniformly [^{15}N,^{13}C]-labeled Nab3 RRM in 50 mM sodium phosphate buffer (pH 8.0), 300 mM NaCl, 10 mM β-mercaptoethanol, (90% H2O/10% D2O). The spectrum was acquired at 303 K on a Bruker Avance 600 MHz spectrometer. The assignments are labeled by the one-letter code of amino acids accompanied by a sequence number. The side-chain resonances of asparagine and glutamine are connected by horizontal lines

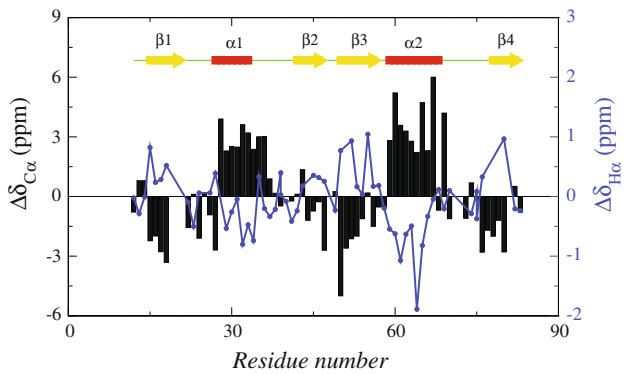


Fig. 2 Hα and Cα chemical shift deviations of Nab3 RRM from the sequence-dependent random coil values along with the secondary structure. α-helices and β-strands are indicated by *red boxes* and *yellow arrows*, respectively, at the top of the diagram

et al. 2005). All assignment for ^{1}H, ^{15}N and ^{13}C backbone and side-chain chemical shifts of Nab3 RRM have been deposited at the BMRB (http://www.bmrb.wisc.edu) and can be accessed under the accession number 16567.

Acknowledgments This work was supported by the Ministry of Education of the Czech Republic (MSM0021622413, Ingo LA08008), GACR (204/08/1212 and 305/10/1490), GAAV (IAA401630903), HHMI/EMBO start-up grant, and HFSP Career Development Award. NOESY spectra were obtained at the CERM NMR facility supported by the EU-NMR program (RII3-026145). FH is in receipt of a Brno City Municipality Scholarship for Talented Ph.D. Students. RP is supported by the EC FP-7 (grant no. 205872).

References

Arigo JT, Eyler DE, Carroll KL, Corden JL (2006) Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. Mol Cell 23:841–851

Carroll KL, Ghirlando R, Ames JM, Corden JL (2007) Interaction of yeast RNA-binding proteins Nrd1 and Nab3 with RNA polymerase II terminator elements. RNA 13:361–373

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

Goddard TD, Kneller DG (2004) SPARKY 3. University of California, San Francisco

Gudipati RK, Villa T, Boulay J, Libri D (2008) Phosphorylation of the RNA polymerase II C-terminal domain dictates transcription termination choice. Nat Struct Mol Biol 15:786–794

Maris C, Dominguez C, Allain FH (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J 272:2118–2131

Sattler M, Schleucher J, Griesinger C (1999) Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. Prog Nucl Magn Reson Spectrosc 34:93–158

Stefl R, Skrisovska L, Allain FH (2005) RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. EMBO Rep 6:33–38

Steinmetz EJ, Brow DA (1998) Control of pre-mRNA accumulation by the essential yeast protein Nrd1 requires high-affinity transcript binding and a domain implicated in RNA polymerase II association. Proc Natl Acad Sci U S A 95:6699–6704

Steinmetz EJ, Conrad NK, Brow DA, Corden JL (2001) RNA-binding protein Nrd1 directs poly(A)-independent 3′-end formation of RNA polymerase II transcripts. Nature 413:327–331

Thiebaut M, Kisseleva-Romanova E, Rougemaille M, Boulay J, Libri D (2006) Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance. Mol Cell 23:853–864

Vanacova S, Stefl R (2007) The exosome and RNA quality control in the nucleus. EMBO Rep 8:651–657

Vasiljeva L, Kim M, Mutschler H, Buratowski S, Meinhart A (2008) The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. Nat Struct Mol Biol 15:795–804

EMBO
reports

review

The exosome and RNA quality control in the nucleus

Stepanka Vanacova[1] *& Richard Stefl*[2+]

[1]University of Basel, Basel, Switzerland and [2]Masaryk University, Brno, Czech Republic

To control the quality of RNA biogenesis in the nucleus, cells use sophisticated molecular machines. These machines recognize and degrade not only RNA trimmings—the leftovers of RNA processing—but also incorrectly processed RNAs that contain defects. By using this mechanism, cells ensure that only high-quality RNAs are engaged in protein synthesis and other cellular processes. The exosome—a complex of several exoribonucleolytic and RNA-binding proteins—is the central 3'-end RNA degradation and processing factor in this surveillance apparatus. The exosome operates with auxiliary factors that stimulate its activity and recruit its RNA substrates in the crowded cellular environment. In this review, we discuss recent structural and functional data related to the nuclear quality-control apparatus, including the long-awaited structure of the human exosome and its activity.

Keywords: RNA decay; exoribonuclease; exosome; structure; nuclear polyadenylation; TRAMP

EMBO *reports* (2007) **8,** 651–657. doi:10.1038/sj.embor.7401005

Introduction

Cellular RNAs are diverse in length and shape, and their functions range from simple messengers to regulators and enzymes involved in gene expression. In eukaryotes, RNAs are produced by one of three RNA polymerases (Pol I–III), which are subsequently processed to their mature form and their trimmings recycled. This concerted RNA biogenesis, which is mediated by many RNA-binding proteins, ribonucleases and other enzymes, is overseen by RNA quality-control mechanisms that ensure only correctly processed RNAs are exported to the cytoplasm, where another set of cellular factors modulate the rate of RNA degradation (Garneau *et al*, 2007).

The exosome is the main RNA degrader in the nucleus. It was initially identified as a component in the processing and maturation of ribosomal RNA precursors but it is in fact involved in the processing and degradation of most nuclear RNAs (reviewed by Houseley *et al*, 2006). The exosome localizes to important sites of RNA biogenesis, such as actively transcribed genes (Andrulis *et al*, 2002; Hieronymus *et al*, 2004) or the nucleolus (Dez *et al*, 2006), and is engaged in RNA surveillance (Dez *et al*, 2006; Torchet *et al*, 2002) possibly through its interactions with other RNA-binding proteins. Although the role of the exosome was described more than 10 years ago, it is only recently that structural and functional studies have provided information on the biochemistry and the molecular organization of this large complex. Furthermore, an additional RNA quality-control pathway involving the exosome and a newly described complex, known as the Trf4 or Trf4–Air2–Mtr4 polyadenylation (TRAMP) complex (Kadaba *et al*, 2006; LaCava *et al*, 2005; Vanacova *et al*, 2005; Wyers *et al*, 2005), has been discovered. Here, we highlight the recent data on the structural and biochemical differences between the human and yeast exosomes, and summarize the current knowledge of the pathways that control the quality of RNA in the eukaryotic nucleus. Although the new studies have provided a tremendous amount of data in the field of nuclear RNA surveillance, there are still a few contradictory issues, which is typical of any young discipline.

Function of the exosome in the nucleus

The exosome is required for the processing and degradation of pre-ribosomal RNAs, pre-small nuclear/small nucleolar RNAs and pre-transfer RNAs (Allmang *et al*, 1999; Hilleren *et al*, 2001; Houseley *et al*, 2006; Kadaba *et al*, 2004; Kuai *et al*, 2004; Mitchell *et al*, 1997; Torchet *et al*, 2002; van Hoof *et al*, 2000) and is also involved in nuclear degradation of aberrant pre-messenger RNAs that result from mutations in various 3'-end processing, splicing and export factors (Bousquet-Antonelli *et al*, 2000; Das *et al*, 2003; Libri *et al*, 2002; Torchet *et al*, 2002). It also degrades messenger RNAs with mutations in their coding sequence in a process known as degradation of RNA in the nucleus (Das *et al*, 2006). Exosome-mediated RNA surveillance is linked directly to and might regulate other processes in the cell. Surveillance of different types of RNAs occurs at distinct nuclear, nucleolar or subnucleolar regions (reviewed by Fasken & Corbett, 2005; Houseley *et al*, 2006; Jensen *et al*, 2003). In yeast, the exosome components interact with transcription elongation factors and are recruited directly to transcriptionally induced genes (Hieronymus *et al*, 2004; Hilleren *et al*, 2001; Vasiljeva & Buratowski, 2006). In this way, the exosome co-transcriptionally monitors the messenger ribonucleoprotein state of nascent transcripts and retains messenger RNAs that are aberrant or otherwise export-incompetent (Fasken & Corbett, 2005; Hilleren *et al*, 2001; Thomsen *et al*, 2003). In yeast mutants that have defective exosome subunits, such RNAs are released from transcription sites and can either be exported to the cytoplasm

[1]Biozentrum, University of Basel, Klingelbergstrasse 50/70, CH-4057 Basel, Switzerland
[2]National Centre for Biomolecular Research, Masaryk University, Kamenice 5/A4, CZ-62500 Brno, Czech Republic
[+]Corresponding author. Tel: +420 5 49 49 24 36; Fax: +420 5 49 49 25 56;
E-mail: stefl@chemi.muni.cz

Submitted 26 February 2007; accepted 4 May 2007

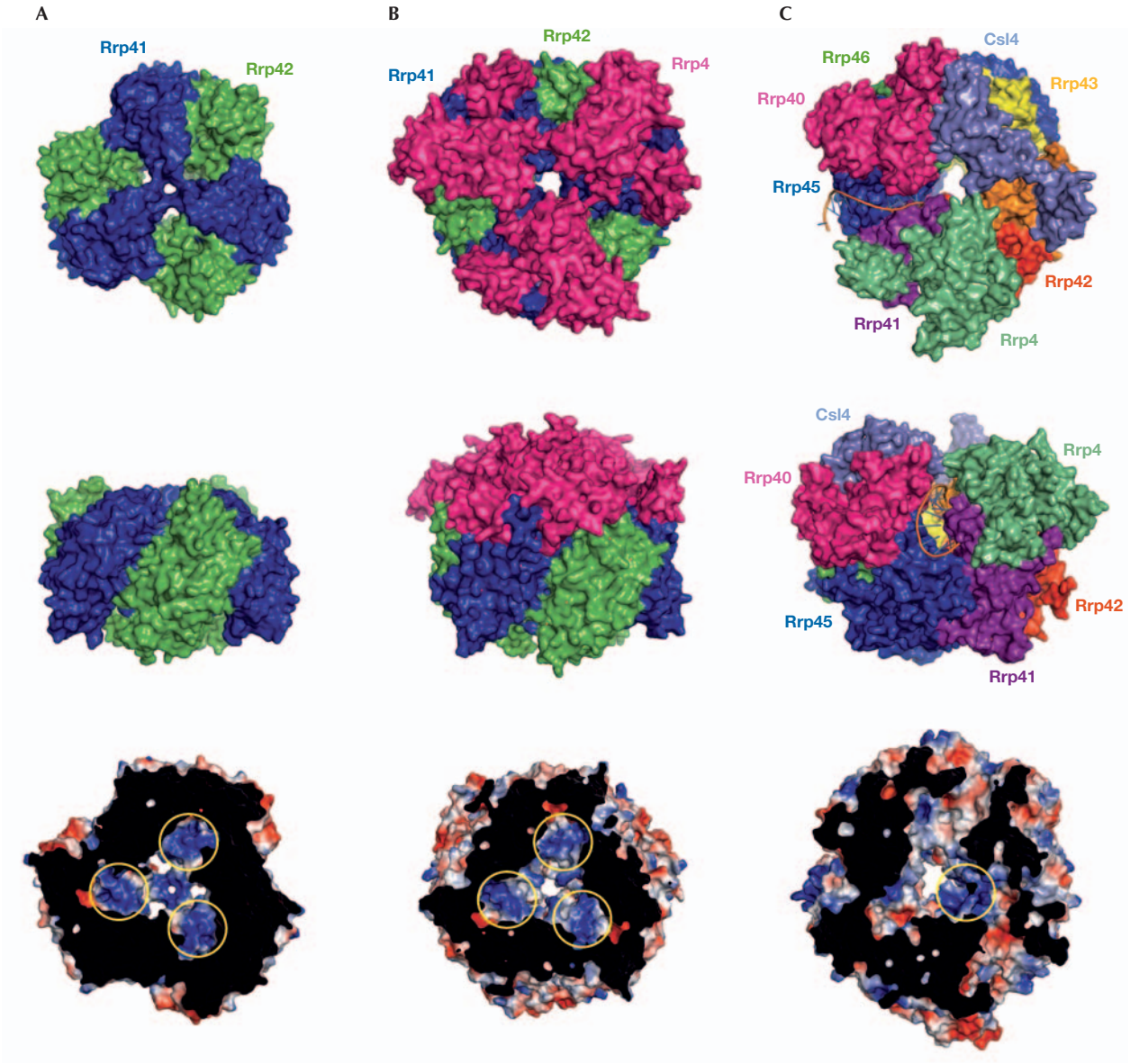


Fig 1 | Molecular architecture of the exosomes. (**A**) Surface representation of the archaeal exosome (Buttner *et al*, 2005; Lorentzen *et al*, 2005); (**B**) the archaeal exosome with the 'cap' formed by Rrp4 (Buttner *et al*, 2005; Lorentzen *et al*, 2005); and (**C**) the eukaryotic exosome (Dziembowski *et al*, 2007; Liu *et al*, 2006). Top views (top row), side views (middle row) and bottom views (bottom row). (**C**) A model of an RNA substrate highlights a putative path of RNA into the exosome through the cleft that is specific to the human exosome (Dziembowski *et al*, 2007; Liu *et al*, 2006). In the bottom row, a slice view on the electrostatic surface potentials (blue, positive; red, negative) shows the RNA-binding pockets in the vicinity of the catalytic sites as indicated by the yellow circles. Graphics were prepared with PyMOL (http://pymol.sourceforge.net). Csl, cep1 synthetic lethality; Rrp, ribosomal RNA processing factor.

or are sequestered to specific subnucleolar foci to be degraded (Carneiro *et al*, 2007; Dez *et al*, 2006; Hilleren *et al*, 2001).

The use of mutants defective in RNA degradation in combination with microarray analysis led to the identification of cryptic unstable transcripts with no obvious or, as yet, identified function (Davis & Ares, 2006; Wyers *et al*, 2005). Notably, the use of tiling microarrays has recently uncovered transcription from many previously unannotated regions of the yeast and the mammalian genomes (David *et al*, 2006; Davis & Ares, 2006; Wyers *et al*, 2005). All these apparently cryptic transcripts are transcribed by the Pol II machinery, although it is unknown why only some of them are highly prone to degradation. It was proposed that the efficient degradation of cryptic unstable transcripts limits the genomic noise resulting from the transcription of inappropriate promoter-like regions (Wyers *et al*, 2005). However, some of the RNAs might be functional or, alternatively, the regulation of their transcription and stability might affect the rate of transcription of the surrounding genes through a transcription-interference mechanism (Davis & Ares, 2006).

Molecular architecture of exosomes

The exosome is conserved from yeast to humans, and is also found in several archaea. In addition, the core of the exosome has a structural homologue in bacteria—the polynucleotide phosphorylase—which is part of the degradosome (Symmons *et al*, 2000). The overall molecular architecture of the exosome has evolved from an ancient ring-like fold of the bacterial phosphorolytic exoribonuclease RNase PH, which is a hexameric ring of three homodimers (Choi *et al*, 2004).

In archaea, the exosome core forms a hexameric ring structure with a central hole involving two types of RNase PH-like subunit, ribosomal RNA processing factor (Rrp)41 and Rrp42 (Buttner *et al*, 2005; Lorentzen *et al*, 2005). The ring is assembled such that the two subunits interact to form a trimer of Rrp41–Rrp42 heterodimers (Fig 1). The three Rrp41 exosome subunits have phosphorolytic activity, whereas the three Rrp42 subunits are inactive. The catalytic sites are located inside the chamber of the ring-like exosome core structure (Fig 1). The Rrp42 non-catalytic subunits have a structural role as they provide a platform for the ring formation and assist in the binding of RNA substrates (Lorentzen & Conti, 2005). In addition, they are involved in binding to Rrp4 or cep1 synthetic lethality (Csl)4, which are RNA-binding proteins with amino-terminal, K-homology and S1 domains (Rrp4) or zinc-ribbon domains (Csl4; Table 1; Buttner *et al*, 2005), and to other protein factors (Walter *et al*, 2006). Three copies of Rrp4 and/or Csl4 associate on top of the exosome core structure and form an extension of the exosomal entry tunnel (Fig 1). Recent structural studies rationalized that such an architectural arrangement allows the regulation of RNA entry into the processing chamber (Lorentzen *et al*, 2007). Conti and colleagues also revealed how the exosome core binds to RNA in the vicinity of the catalytic site (Lorentzen *et al*, 2007; Lorentzen & Conti, 2005). The RNA substrates are recognized by a network of base-non-specific and ribose-specific interactions between the sugar-phosphate backbone and the predominantly arginine side-chains of the Rrp41–Rrp42 interface, which forms an electropositive binding pocket (Fig 1).

In contrast to archaea, the eukaryotic exosome uses six different subunits to form the ring structure. However, these subunits can be classified as Rrp41-like—Rrp41, Rrp46, and Mtr3—and Rrp42-like—Rrp45 (polymyositis/scleroderma (PM/Scl)-75 human), Rrp43 (opa-interacting protein (OIP)2 human), and Rrp42—based on their similarities to their archaeal counterparts. Interestingly, the eukaryotic six-member ring does not assemble into a stable structure in the absence of Rrp4, Rrp40 and Csl4 (Liu *et al*, 2006), indicating that the eukaryotic exosome core is composed of nine subunits. The human exosome is asymmetrical and shows a more complex architecture compared with its archaeal counterpart (Fig 1). Each Rrp41-like–Rrp42-like heterodimer binds only the correct heterodimer partner and does not form trimers of the same heterodimers. It is highly likely that the eukaryotic exosome has evolved to include different subunits that are necessary to bind various auxiliary factors involved in RNA degradation and processing. The human exosome core has only one catalytic site in the Rrp41–Rrp45 heterodimer (Liu *et al*, 2006), a characteristic that was previously suspected (Lorentzen & Conti, 2005). The human Rrp41–Rrp45 heterodimer in isolation also has RNA-degradation activity that results in a degradation pattern identical to the pattern of the complete exosome. This could be a result of the asymmetrical arrangement in which the RNA entry tunnel of the human exosome is significantly widened by a large cleft between Rrp4 and Rrp40 subunits. This cleft is located exactly over the Rrp41–Rrp45 heterodimer and could act as an entry site for the RNA (Fig 1). In this scenario, the Rrp41–Rrp45 heterodimer would use its surface within the cleft to bind to RNA. This would explain the virtually identical RNA degradation patterns of the Rrp41–Rrp45 heterodimer and the human exosome. Consistently, the RNase PH domain of human Rrp45 in isolation binds to U-rich and AU-rich sequences (Anderson *et al*, 2006). To understand how RNA is threaded into the human exosome, the structure of the exosome in complex with RNA will need to be solved.

Interestingly, the structural studies by Conti, Sattler and colleagues show that the N-terminal region of *Saccharomyces cerevisiae* Rrp40 is unstructured in its free form, whereas the S1 and K-homology domains are essentially identical to human RRP40 in the exosome core structure (Oddone *et al*, 2007). This indicates that the exosome core formation includes an induced-fit mechanism in which certain domains undergo a conformational change on finding the 'correct' binding partner. These authors also show that the weak RNA-binding affinity of individual subunits becomes significantly attenuated in the context of the assembled exosome.

In eukaryotes, the nuclear exosome associates with the auxiliary factor Rrp6, which participates in both RNA processing and quality control. The structure of *S. cerevisiae* Rrp6 has a conserved RNase D core with a flanking helicase and RNase D carboxy-terminal domain, and an N-terminal domain that is proposed to mediate the interaction with the exosome core (Midtgaard *et al*, 2006).

Clearly, further structures of the eukaryotic exosome core that bind to additional protein factors—particularly the TRAMP complex—and to RNA substrates need to be solved to decipher how this quality-control machinery operates.

Table 1 | Subunits of the eukaryotic and archaeal exosomes

Human	Yeast	Archaea	Domains
Core			
Rrp41*	Rrp41	Rrp41*	RNase PH
Rrp45	Rrp45	—	RNase PH
Mtr3	Mtr3	—	RNase PH
Rrp42	Rrp42	Rrp42	RNase PH
Rrp46	Rrp46	—	RNase PH
Rrp43	Rrp43	—	RNase PH
Csl4	Csl4	—	NT, S1, zinc-ribbon
Rrp4	Rrp4	—	NT, S1, KH
Rrp40	Rrp40	—	NT, S1, KH
Associated proteins			
—	—	Csl4	NT, S1, zinc-ribbon
—	—	Rrp4	NT, S1, KH
Rrp6*	Rrp6*	—	RNase D
—	Rrp44* (Dis3)	—	RNase R
—	Rrp47	—	

*denotes active subunits. Note that for activity, the Rrp41 subunit requires Rrp42 and Rrp45 subunits in archaea and human, respectively. Csl, cep1 synthetic lethality; KH, K-homology; Mtr, mRNA transport; NT, N-terminal; PH, phosphate-dependent; Rrp, ribosomal RNA processing factor.

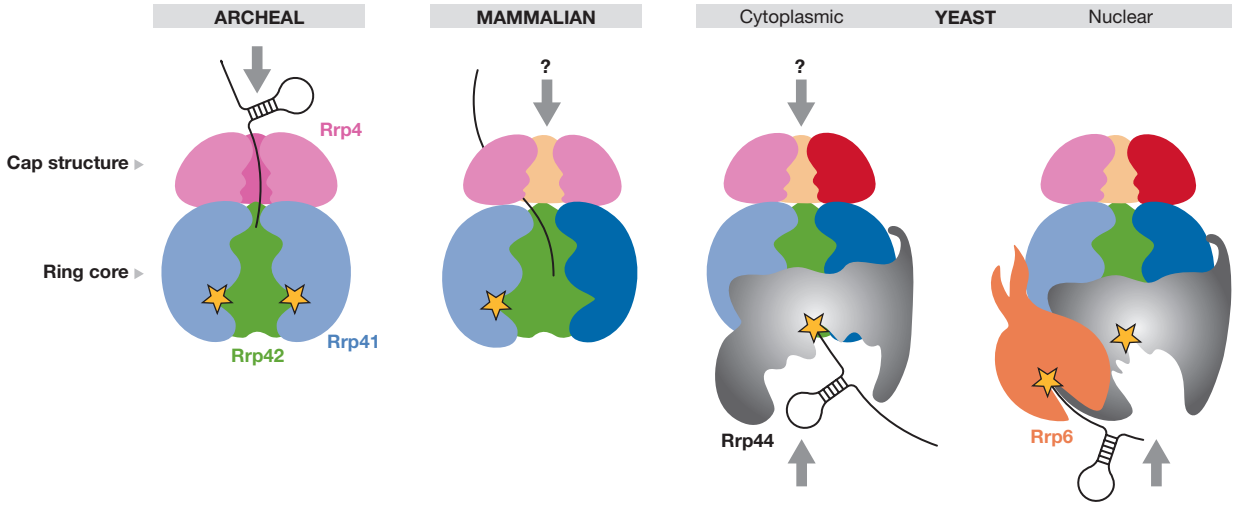


Fig 2 | Schematic diagram of the exosome topologies and RNA threading. The arrows indicate the confirmed or hypothetical (indicated by a question mark) pathways of an RNA substrate in various forms of the exosomes (Dziembowski *et al*, 2007; Liu *et al*, 2006; Lorentzen *et al*, 2007). Yellow stars indicate catalytically active sites. The predicted positions of the yeast hydrolytic exonucleases Rrp44 and Rrp6 are indicated. Rrp, ribosomal RNA processing factor.

Degradation mechanism of exosomes

The exosome core associates with additional protein factors in a compartment- and organism-specific manner. In yeast, the exosome core has an additional stable subunit, the hydrolytic exonuclease Rrp44 (also known as Dis3; Mitchell *et al*, 1997), and in the nucleus it binds to two more subunits: the hydrolytic ribonuclease Rrp6, and Rrp47 (Mitchell *et al*, 2003). By contrast, in humans, Rrp44 was not detected in the affinity-purified exosomes and the Rrp6 homologue (PM/Scl-100) was also found in the cytoplasmic fractions (Chen *et al*, 2001).

The reasons for the differences in the composition of the human and yeast exosomes have become apparent from recent studies (Fig 2). It was shown that the human core exosome has one catalytically active site, within the Rrp41–Rrp45 dimer (Dziembowski *et al*, 2007; Liu *et al*, 2006), whereas the yeast nine-subunit core has no catalytically active site (Dziembowski *et al*, 2007; Liu *et al*, 2006). Instead, the hydrolytic nuclease Rrp44 is the crucial enzyme for RNA degradation in yeast (Dziembowski *et al*, 2007; Liu *et al*, 2006). Therefore, eukaryotic exosomes have human-like and yeast-like subtypes with phosphorolytic and hydrolytic activities, respectively. The former subtype degrades RNA inside the exosome chamber, whereas the latter subtype uses the hydrolytic enzyme that is most likely anchored on the surface of the exosome core assembly (Hernandez *et al*, 2006). It is not known why yeast switched to the hydrolytic mode of RNA degradation; however, it is still possible that under specific conditions and/or on specific RNA substrates the yeast Rrp41 also becomes active. This is a question that needs to be investigated in the future. Similarly, the contribution of the human homologues of the hydrolytic nucleases, RRP6 and RRP44, to the exosome activities *in vivo* must be assessed further.

The recombinant yeast Rrp44 is highly processive, even on structured RNAs; however, its activity is significantly reduced on binding to the exosome core (Liu *et al*, 2006). This is perhaps to prevent high activity levels that could interfere with the stability of many functional RNAs in the cell. At present, it is not clear whether Rrp44 undergoes a conformational change or whether the active site is partly masked on binding to the exosome core. The observation that Rrp44 is responsible for yeast exosome activity suggests that the substrate RNA might not pass through the central channel of the ring-like structure. Further structural and functional analyses are required to determine the exact path of the RNA substrate.

Rrp6, the second nucleus-specific hydrolytic exonuclease in yeast, has distributive exonucleolytic activity on unstructured and poly(A)-extended RNAs, and displays only weak activity on structured substrates (Liu *et al*, 2006). Its activity is not affected on assembly with the exosome core and Rrp44. Although *Rrp6* is a non-essential gene, its deletion results in many RNA-processing defects (Allmang *et al*, 1999; Davis & Ares, 2006; Egecioglu *et al*, 2006; Kadaba *et al*, 2004; Kuai *et al*, 2004; Liu *et al*, 2006; van Hoof *et al*, 2000; Wyers *et al*, 2005). This suggests that although partly overlapping in function, Rrp44 and Rrp6 have different substrate specificities *in vivo*. The function of human RRP6 has not been analysed in detail and the extent to which it is specific for nuclear RNA metabolism is unclear.

Other players in nuclear RNA surveillance

Although the reconstituted exosomes efficiently degrade RNAs *in vitro* (Dziembowski *et al*, 2007; Liu *et al*, 2006), *in vivo* RNAs often form complex secondary and tertiary conformations, acquire nucleotide modifications and/or assemble into specific ribonucleoprotein particles (RNPs) that need to be protected from nucleases. Accessory factors are then required to help the exosome recognize aberrant RNAs or RNPs that need to be discarded or further processed (reviewed by Olesen *et al*, 2005). Two accessory protein complexes were shown to stimulate the exosome in the yeast nucleus. First is the polyadenylation complex called TRAMP4 and/or TRAMP5 depending on whether it contains the poly(A) polymerase Trf4 or Trf5 (Houseley & Tollervey, 2006; Kadaba *et al*, 2004, 2006; LaCava *et al*, 2005; Vanacova *et al*, 2005; Wyers *et al*, 2005). TRAMP

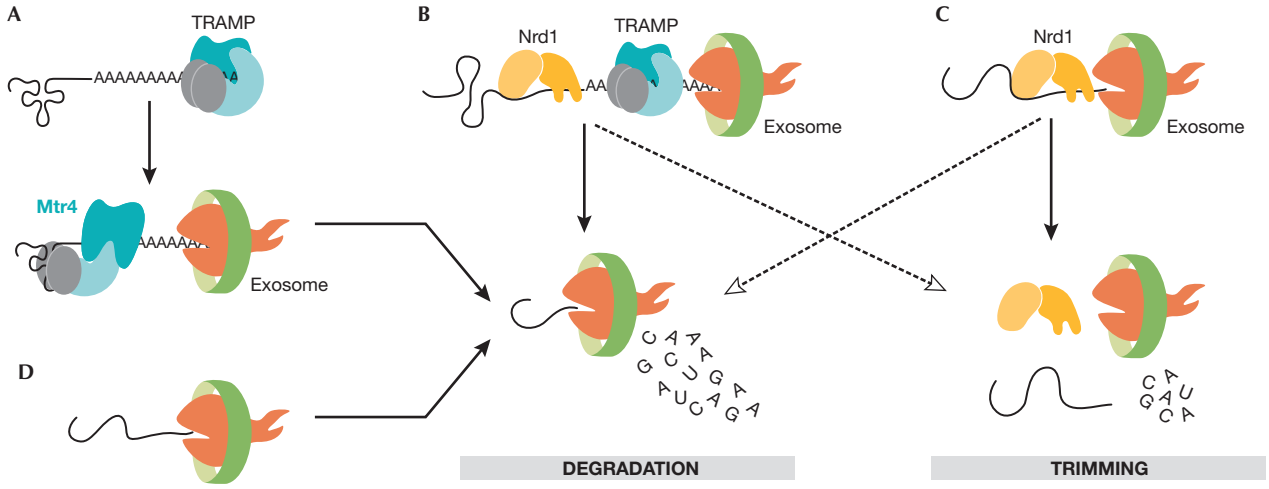


Fig 3 | Exonucleolytic activity of the exosome is stimulated by accessory protein complexes *in vitro* and *in vivo*. (**A**) The Trf4–Air2–Mtr4 polyadenylation (TRAMP) complex tags aberrant RNAs with short stretches of oligo(A)s, which initiates RNA digestion by the exosome (LaCava *et al*, 2005; Vanacova *et al*, 2005). (**B**) Mtr4 helicase of the TRAMP complex unwinds the structured parts of RNAs. The TRAMP complex associates with the Nrd1 complex that binds to short sequence elements on a subset of nuclear RNAs (Vasiljeva & Buratowski, 2006). The interaction between the specific RNA recognition mediated by the Nrd1 complex and the polyadenylation activity mediated by the TRAMP complex acts as the initiation step for RNA degradation by the exosome. (**C**)The Nrd1 complex can stimulate exosome activity on RNAs with the Nrd1 complex-specific binding sites (Vasiljeva & Buratowski, 2006). This often leads to partial digestion of the RNA (trimming), but can also cause RNA degradation. (**D**) The exosome destroys the leftovers of RNA processing, such as the products of endonucleolytic cleavage, apparently by itself. Air, arginine methyltransferase-interacting RING finger protein; Mtr, mRNA transport; Nrd, nuclear pre-mRNA down-regulation; Trf, topoisomerase one-related function.

adds short poly(A) tails to aberrant or unstable transcripts, forming a favourable substrate for the exosome (Fig 3). The identity of the exonuclease within the yeast nuclear exosome that degrades these polyadenylated targets is still unknown. Rrp6 shows higher activity on unstructured poly(A)-extended molecules and Rrp44 can degrade more structured RNAs (Liu *et al*, 2006); therefore, it is possible that Rrp6 initiates digestion and, by inducing some conformational changes, it hands the RNA over to Rrp44. The helicase Mtr4 might help to unwind the structured portions of the RNAs when Rrp44 is in its less active conformation. Mtr4 might also act as a scaffolding protein because it interacts with both the TRAMP and the exosome complexes (de la Cruz *et al*, 1998; LaCava *et al*, 2005; Vanacova *et al*, 2005; Wyers *et al*, 2005). It is not known how TRAMP recognizes aberrant RNAs. One possibility is that it monitors the conformation status of RNAs. For example, it can specifically polyadenylate and target transfer RNAs that are incorrectly folded for degradation (Kadaba *et al*, 2004; Vanacova *et al*, 2005). Further structural and functional studies are required to understand fully the mechanism of substrate recognition.

Both the exosome and the TRAMP4 complexes interact with an additional RNA-binding complex called the nuclear pre-mRNA down-regulation (Nrd)1 complex (Vasiljeva & Buratowski, 2006), which consists of the RNA helicase Sen1, and the proteins Nrd1 and nuclear polyadenylated RNA-binding (Nab)3 that recognize specific sequence motifs on RNAs (Carroll *et al*, 2004; Steinmetz & Brow, 1998). This complex is required for transcription termination of small nuclear RNA and small nucleolar RNA genes (Conrad *et al*, 2000; Steinmetz & Brow, 1996, 1998; Steinmetz *et al*, 2001). *In vitro*, it can directly stimulate exosome degradation of substrates with Nrd1- and Nab3-binding motifs (Vasiljeva & Buratowski, 2006). *In vivo*, it probably helps to bring the exosome to specific RNA substrates. However, it is not clear whether the Nrd1 complex acts by itself or whether it requires the association with TRAMP4 to stimulate the exosome (Vasiljeva & Buratowski, 2006).

Concluding remarks and open questions

The recent structural information on the human exosome core supports the existence of a common basis for RNA-degradation machineries in prokaryotes, eukaryotes and archaea. It is also striking how the RNA-degradation pathway mediated by the exosome resembles the features of protein degradation by proteasomes, a concept first envisioned by van Hoof & Parker (1999). Despite the increasing amount of data on exosomes, there are still many questions to be answered. How does the quality-control machinery distinguish its RNA substrates from other RNA molecules? What allows the exosome to switch from the degradation mode to the processing mode? How exactly is the exosome stimulated by the TRAMP and Nrd1 complexes, and what other protein factors are required for its proper function *in vivo*? Many homologues of Trf4-like proteins exist in higher eukaryotes; therefore, it will be interesting to see whether similar poly(A)-mediated degradation pathways operate in the nuclei of metazoa. To conclude, there is no doubt that in the near future we will witness considerable efforts to unravel the details of the RNA surveillance apparatus—a universal machinery with the fundamental quest to seek and destroy damaged molecules.

ACKNOWLEDGEMENTS

We apologize to those authors whose work could not be cited owing to space constraints. We thank W. Keller for support, and C. Rammelt, M.-J. Schmidt, D. Ladle and L. Krejci for critical reading of the manuscript and for helpful comments. The authors are supported by the Ministry of Education of the Czech Republic (MSM0021622413), the University of Basel and the Swiss National Science Fund. R.S. is in receipt of a European Molecular Biology Organization/Howard Hughes Medical Institute Start-up Grant and a Human Frontier Science Program Career Development Award.

REFERENCES

Allmang C, Kufel J, Chanfreau G, Mitchell P, Petfalski E, Tollervey D (1999) Functions of the exosome in rRNA, snoRNA and snRNA synthesis. *EMBO J* **18:** 5399–5410

Anderson JR, Mukherjee D, Muthukumaraswamy K, Moraes KC, Wilusz CJ, Wilusz J (2006) Sequence-specific RNA binding mediated by the RNase PH domain of components of the exosome. *RNA* **12:** 1810–1816

Andrulis ED, Werner J, Nazarian A, Erdjument-Bromage H, Tempst P, Lis JT (2002) The RNA processing exosome is linked to elongating RNA polymerase II in *Drosophila*. *Nature* **420:** 837–841

Bousquet-Antonelli C, Presutti C, Tollervey D (2000) Identification of a regulated pathway for nuclear pre-mRNA turnover. *Cell* **102:** 765–775

Buttner K, Wenig K, Hopfner KP (2005) Structural framework for the mechanism of archaeal exosomes in RNA processing. *Mol Cell* **20:** 461–471

Carneiro T, Carvalho C, Braga J, Rino J, Milligan L, Tollervey D, Carmo-Fonseca M (2007) Depletion of the yeast nuclear exosome subunit Rrp6 results in accumulation of polyadenylated RNAs in a discrete domain within the nucleolus. *Mol Cell Biol* **27:** 4157–4165

Carroll KL, Pradhan DA, Granek JA, Clarke ND, Corden JL (2004) Identification of *cis* elements directing termination of yeast nonpolyadenylated snoRNA transcripts. *Mol Cell Biol* **24:** 6241–6252

Chen CY, Gherzi R, Ong SE, Chan EL, Raijmakers R, Pruijn GJ, Stoecklin G, Moroni C, Mann M, Karin M (2001) AU binding proteins recruit the exosome to degrade ARE-containing mRNAs. *Cell* **107:** 451–464

Choi JM, Park EY, Kim JH, Chang SK, Cho Y (2004) Probing the functional importance of the hexameric ring structure of RNase PH. *J Biol Chem* **279:** 755–764

Conrad NK, Wilson SM, Steinmetz EJ, Patturajan M, Brow DA, Swanson MS, Corden JL (2000) A yeast heterogeneous nuclear ribonucleoprotein complex associated with RNA polymerase II. *Genetics* **154:** 557–571

Das B, Butler JS, Sherman F (2003) Degradation of normal mRNA in the nucleus of *Saccharomyces cerevisiae*. *Mol Cell Biol* **23:** 5502–5515

Das B, Das S, Sherman F (2006) Mutant LYS2 mRNAs retained and degraded in the nucleus of *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* **103:** 10871–10876

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM (2006) A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci USA* **103:** 5320–5325

Davis CA, Ares M, Jr (2006) Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* **103:** 3262–3267

de la Cruz J, Kressler D, Tollervey D, Linder P (1998) Dob1p (Mtr4p) is a putative ATP-dependent RNA helicase required for the 3' end formation of 5.8S rRNA in *Saccharomyces cerevisiae*. *EMBO J* **17:** 1128–1140

Dez C, Houseley J, Tollervey D (2006) Surveillance of nuclear-restricted pre-ribosomes within a subnucleolar region of *Saccharomyces cerevisiae*. *EMBO J* **25:** 1534–1546

Dziembowski A, Lorentzen E, Conti E, Seraphin B (2007) A single subunit, Dis3, is essentially responsible for yeast exosome core activity. *Nat Struct Mol Biol* **14:** 15–22

Egecioglu DE, Henras AK, Chanfreau GF (2006) Contributions of Trf4p- and Trf5p-dependent polyadenylation to the processing and degradative functions of the yeast nuclear exosome. *RNA* **12:** 26–32

Fasken MB, Corbett AH (2005) Process or perish: quality control in mRNA biogenesis. *Nat Struct Mol Biol* **12:** 482–488

Garneau NL, Wilusz J, Wilusz CJ (2007) The highways and byways of mRNA decay. *Nat Rev Mol Cell Biol* **8:** 113–126

Hernandez H, Dziembowski A, Taverner T, Seraphin B, Robinson CV (2006) Subunit architecture of multimeric complexes isolated directly from cells. *EMBO Rep* **7:** 605–610

Hieronymus H, Yu MC, Silver PA (2004) Genome-wide mRNA surveillance is coupled to mRNA export. *Genes Dev* **18:** 2652–2662

Hilleren P, McCarthy T, Rosbash M, Parker R, Jensen TH (2001) Quality control of mRNA 3'-end processing is linked to the nuclear exosome. *Nature* **413:** 538–542

Houseley J, Tollervey D (2006) Yeast Trf5p is a nuclear poly(A) polymerase. *EMBO Rep* **7:** 205–211

Houseley J, LaCava J, Tollervey D (2006) RNA-quality control by the exosome. *Nat Rev Mol Cell Biol* **7:** 529–539

Jensen TH, Dower K, Libri D, Rosbash M (2003) Early formation of mRNP: license for export or quality control? *Mol Cell* **11:** 1129–1138

Kadaba S, Krueger A, Trice T, Krecic AM, Hinnebusch AG, Anderson J (2004) Nuclear surveillance and degradation of hypomodified initiator tRNAMet in *S. cerevisiae*. *Genes Dev* **18:** 1227–1240

Kadaba S, Wang X, Anderson JT (2006) Nuclear RNA surveillance in *Saccharomyces cerevisiae*: Trf4p-dependent polyadenylation of nascent hypomethylated tRNA and an aberrant form of 5S rRNA. *RNA* **12:** 508–521

Kuai L, Fang F, Butler JS, Sherman F (2004) Polyadenylation of rRNA in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* **101:** 8581–8586

LaCava J, Houseley J, Saveanu C, Petfalski E, Thompson E, Jacquier A, Tollervey D (2005) RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell* **121:** 713–724

Libri D, Dower K, Boulay J, Thomsen R, Rosbash M, Jensen TH (2002) Interactions between mRNA export commitment, 3'-end quality control, and nuclear degradation. *Mol Cell Biol* **22:** 8254–8266

Liu Q, Greimann JC, Lima CD (2006) Reconstitution, activities, and structure of the eukaryotic RNA exosome. *Cell* **127:** 1223–1237

Lorentzen E, Conti E (2005) Structural basis of 3' end RNA recognition and exoribonucleolytic cleavage by an exosome RNase PH core. *Mol Cell* **20:** 473–481

Lorentzen E, Walter P, Fribourg S, Evguenieva-Hackenberg E, Klug G, Conti E (2005) The archaeal exosome core is a hexameric ring structure with three catalytic subunits. *Nat Struct Mol Biol* **12:** 575–581

Lorentzen E, Dziembowski A, Lindner D, Seraphin B, Conti E (2007) RNA channelling by the archaeal exosome. *EMBO Rep* **8:** 470–476

Midtgaard SF, Assenholt J, Jonstrup AT, Van LB, Jensen TH, Brodersen DE (2006) Structure of the nuclear exosome component Rrp6p reveals an interplay between the active site and the HRDC domain. *Proc Natl Acad Sci USA* **103:** 11898–11903

Mitchell P, Petfalski E, Shevchenko A, Mann M, Tollervey D (1997) The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'-->5' exoribonucleases. *Cell* **91:** 457–466

Mitchell P, Petfalski E, Houalla R, Podtelejnikov A, Mann M, Tollervey D (2003) Rrp47p is an exosome-associated protein required for the 3' processing of stable RNAs. *Mol Cell Biol* **23:** 6982–6992

Oddone A, Lorentzen E, Basquin J, Gasch A, Rybin V, Conti E, Sattler M (2007) Structural and biochemical characterization of the yeast exosome component Rrp40. *EMBO Rep* **8:** 63–69

Olesen JR, Libri D, Jensen TH (2005) A link between transcription and mRNP quality in *Saccharomyces cerevisiae*. *RNA Biol* **2:** 45–48

Steinmetz EJ, Brow DA (1996) Repression of gene expression by an exogenous sequence element acting in concert with a heterogeneous nuclear ribonucleoprotein-like protein, Nrd1, and the putative helicase Sen1. *Mol Cell Biol* **16:** 6993–7003

Steinmetz EJ, Brow DA (1998) Control of pre-mRNA accumulation by the essential yeast protein Nrd1 requires high-affinity transcript binding and a domain implicated in RNA polymerase II association. *Proc Natl Acad Sci USA* **95:** 6699–6704

Steinmetz EJ, Conrad NK, Brow DA, Corden JL (2001) RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts. *Nature* **413:** 327–331

Symmons MF, Jones GH, Luisi BF (2000) A duplicated fold is the structural basis for polynucleotide phosphorylase catalytic activity, processivity, and regulation. *Structure* **8:** 1215–1226

Thomsen R, Libri D, Boulay J, Rosbash M, Jensen TH (2003) Localization of nuclear retained mRNAs in *Saccharomyces cerevisiae*. *RNA* **9:** 1049–1057

Torchet C, Bousquet-Antonelli C, Milligan L, Thompson E, Kufel J, Tollervey D (2002) Processing of 3'-extended read-through transcripts by the exosome can generate functional mRNAs. *Mol Cell* **9:** 1285–1296

van Hoof A, Parker R (1999) The exosome: a proteasome for RNA? *Cell* **99:** 347–350

van Hoof A, Lennertz P, Parker R (2000) Yeast exosome mutants accumulate 3'-extended polyadenylated forms of U4 small nuclear RNA and small nucleolar RNAs. *Mol Cell Biol* **20:** 441–452

Vanacova S, Wolf J, Martin G, Blank D, Dettwiler S, Friedlein A, Langen H, Keith G, Keller W (2005) A new yeast poly(A) polymerase complex involved in RNA quality control. *PLoS Biol* **3:** e189

Vasiljeva L, Buratowski S (2006) Nrd1 interacts with the nuclear exosome for 3' processing of RNA polymerase II transcripts. *Mol Cell* **21:** 239–248

Walter P, Klein F, Lorentzen E, Ilchmann A, Klug G, Evguenieva-Hackenberg E (2006) Characterization of native and reconstituted exosome complexes from the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Mol Microbiol* **62:** 1076–1089

Wyers F *et al* (2005) Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121:** 725–737

Stepanka Vanacova ***Richard Stefl***

EMBO reports

scientific report

The testis-specific human protein RBMY recognizes RNA through a novel mode of interaction

Lenka Skrisovska[1], Cyril F. Bourgeois[2,3], Richard Stefl[1], Sushma-Nagaraja Grellscheid[4], Liliane Kister[2,3], Philipp Wenter[5], David J. Elliott[4], James Stevenin[2,3+] & Frédéric H.-T. Allain[1++]

[1]Institute of Molecular Biology and Biophysics, ETH Zurich, Swiss Federal Institute of Technology, Zurich, Switzerland, [2]Institut de Génétique et de Biologie Moléculaire et Cellulaire, Inserm, U596 & CNRS UMR7104, Illkirch, France, [3]Université L. Pasteur, Strasbourg, France, [4]Institute of Human Genetics, University of Newcastle upon Tyne, Newcastle upon Tyne, UK, and [5]Laboratory of Nucleic Acid Chemistry-EPFL, Lausanne, Switzerland

The RBMY (RNA-binding motif gene on Y chromosome) protein encoded by the human Y chromosome is important for normal sperm development. Although its precise molecular RNA targets are unknown at present, it is suggested that human RBMY (hRBMY) participates in splicing in the testis. Using systematic evolution of ligands by exponential enrichment, we found that RNA stem–loops capped by a $C^A/_UCAA$ pentaloop are high-affinity binding targets for hRBMY. Subsequent nuclear magnetic resonance structural determination of the hRBMY RNA recognition motif (RRM) in complex with a high-affinity target showed two distinct modes of RNA recognition. First, the RRM β-sheet surface binds to the RNA loop in a sequence-specific fashion. Second, the β2–β3 loop of the hRBMY inserts into the major groove of the RNA stem. The first binding mode might be conserved in the paralogous protein heterogeneous nuclear RNP G, whereas the second mode of binding is found only in hRBMY. This structural difference could be at the origin of the function of RBMY in spermatogenesis.

Keywords: alternative splicing; NMR; protein–nucleic acid recognition; spermatogenesis; SELEX

EMBO reports (2007) **8**, 372–379. doi:10.1038/sj.embor.7400910

INTRODUCTION

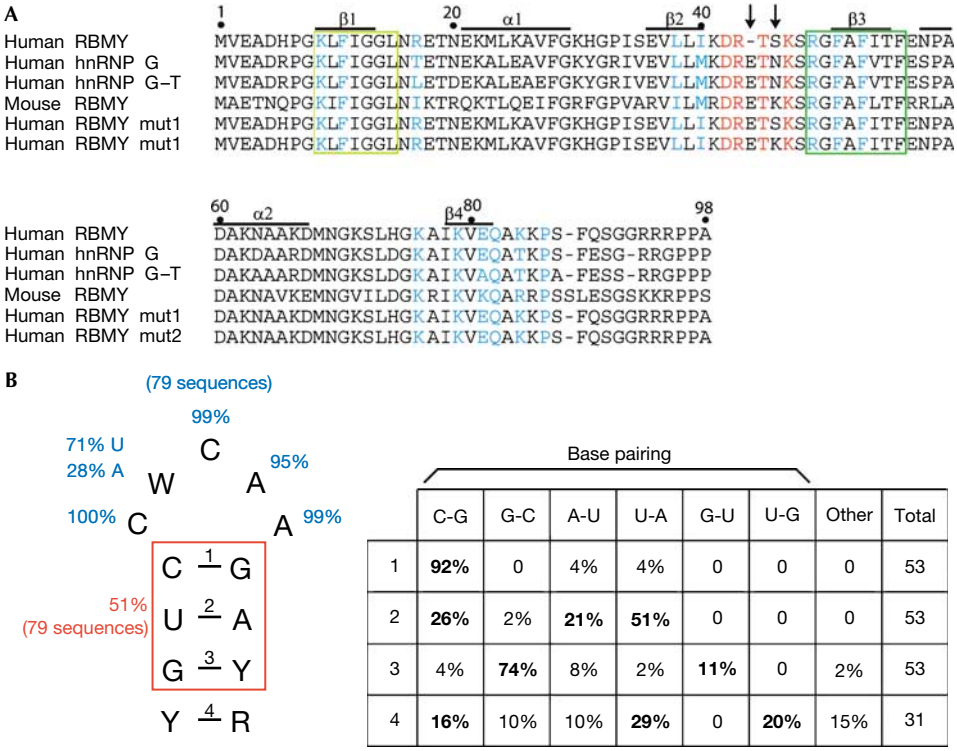
Testes have a highly specialized gene expression program, which is necessary to carry out the complex differentiation of spermatogonia into mature spermatozoa. In particular, alternative splicing of specific pre-mRNAs is prevalent in testis, although little is known about how these events are regulated at the molecular level (Venables, 2002). Among the potential specific regulators of alternative splicing in testis, the family of human *RBMY* (*hRBMY*; RNA-binding motif gene on Y chromosome) genes was identified as a candidate for the azoospermia factor (AZF; Elliott, 2004). Human RBMY is expressed specifically in the nuclei of adult male germ cells throughout all transcriptionally active stages of spermatogenesis, and deletion of the functional copies of *RBMY* is associated with an arrest of meiotic division I during spermatogenesis (Elliott *et al*, 1997). Human RBMY has an amino-terminal RNA recognition motif (RRM) and a carboxy-terminal domain composed of four repetitions of a Ser-Arg-Gly-Tyr tetrapeptide motif (SRGY box; Ma *et al*, 1993). *RBMY* is found on the Y chromosome of all mammals (Mahadevaiah *et al*, 1998). The mouse RBMY (mRBMY) contains an RRM with 74% similarity to hRBMY, followed by only one SRGY box. RBMY-deficient mice do not show the same phenotype as in humans; they have abnormal sperm development but are not sterile (Mahadevaiah *et al*, 1998).

Human and mouse *RBMY* have a chromosome X-located paralogue (*RBMX*), which encodes the widely expressed heterogeneous nuclear ribonucleoprotein (RNP) G (Delbridge *et al*, 1999). Human hnRNP G contains an N-terminal RRM with 88% similarity to hRBMY, followed by only one SRGY box. There is also a third human retrogene (*hnRNP G-T*) that belongs to the same family, which is expressed only in the testis. hnRNP G-T contains an N-terminal RRM with 84% similarity to hRBMY but no SRGY box (Elliott *et al*, 2000b).

Although nothing is known about the RNA targets of these proteins, several studies suggest a role in the regulation of RNA processing (Venables *et al*, 1999, 2000; Elliott *et al*, 2000a). Human RBMY was shown to interact with Sam68 and the closely related T-STAR protein (signal transduction and RNA binding), which are considered to be molecular transducers between cell signalling and splicing regulation (Elliott, 2004). Furthermore, hRBMY can affect splicing through its interaction with SRp20

[1]Institute of Molecular Biology and Biophysics, ETH Zurich, Swiss Federal Institute of Technology, 8093 Zurich, Switzerland
[2]Institut de Génétique et de Biologie Moléculaire et Cellulaire, and Inserm, U596 and CNRS UMR7104, 67400 Illkirch, France
[3]Université L. Pasteur, 67000 Strasbourg, France
[4]Institute of Human Genetics, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 3BZ, UK
[5]Laboratory of Nucleic Acid Chemistry-EPFL, 1015 Lausanne, Switzerland
[+]Corresponding author. Tel.: +33 (0)3 88 65 33 61; Fax: +33 (0)3 88 65 32 01; E-mail: stevenin@titus.u-strasbg.fr
[++]Corresponding author. Tel.: +41 1 633 3940; Fax: +41 1 633 1294; E-mail: allain@mol.biol.ethz.ch

Received 9 November 2006; revised 20 December 2006; accepted 2 January 2007; published online 23 February 2007



	C-G	G-C	A-U	U-A	G-U	U-G	Other	Total
1	**92%**	0	4%	4%	0	0	0	53
2	**26%**	2%	**21%**	**51%**	0	0	0	53
3	4%	**74%**	8%	2%	**11%**	0	2%	53
4	**16%**	10%	10%	**29%**	0	**20%**	15%	31

Fig 1 | Amino-acid sequence of the heterogeneous nuclear RNP G protein family and RNA-binding SELEX consensus sequence obtained for human RBMY. (**A**) Sequence alignment of human RBMY (hRBMY) RRM, and related orthologues and paralogues. The RNP2 and RNP1 motifs are indicated by light green and dark green boxes, respectively. Amino acids of hRBMY found in contact with the RNA loop and stem are shown in blue and red, respectively. The arrows show the two non-conserved mutated amino acids of β2–β3 loop. (**B**) Consensus sequence obtained after SELEX. All analysed sequences (supplementary Fig S1 online) can fold into a stable stem–loop structure with a highly conserved CWCAA loop. The table shows the statistical predominance for specific base-pairing in the stem, including non-canonical base pairs (G-U or U-G). The red box shows the core consensus sequence of the stem that we observed in 51% of sequences. RBMY, RNA-binding motif gene on Y chromosome; SELEX, systematic evolution of ligands by exponential enrichment; Y, C or U; W, U or A; R, G or C.

or Tra2β (Elliott *et al*, 2000a, Venables *et al*, 2000), which belong to the family of SR-rich pre-mRNA splicing regulators (Bourgeois *et al*, 2004).

To understand better the biological function of this family of proteins, we carried out systematic evolution of the ligand by exponential enrichment (SELEX) with hRBMY RRM. We found that the RRM binds with high specificity to stem–loop RNAs containing a $C^A/_UCAA$ consensus sequence in the loop. We determined the solution structure of hRBMY RRM in complex with one of its RNA targets. The structure shows that the hRBMY RRM not only recognizes the loop in a sequence-specific manner, but also the shape of the RNA as the β2–β3 loop of the RRM is inserted into the major groove of the RNA stem.

RESULTS

Identification of human RBMY RNA targets by SELEX

Human RBMY contains an RRM with unknown properties; therefore, we used the SELEX approach to determine its RNA-binding specificity (Cavaloc *et al*, 1999). The glutathione *S*-transferase (GST) fusion protein included the 108 N-terminal residues of hRBMY containing the RRM. After six cycles of selection, we obtained fully coherent results as the same consensus motif was obtained with two distinct 20-nucleotide (nt) randomized matrices (supplementary Fig S1 online). Interestingly, the selected motifs (Fig 1) consist of a hairpin, including an invariable pentaloop ($C^A/_UCAA$) and a stem of variable length (between 4 and 11 perfect base pairs) in which a C-G base-pair is predominantly adjacent to the loop. Altogether, 51% of the selected clones had a GUC–loop–GAY structure. However, the selection of other stem sequences indicates that hRBMY can recognize stems with different base pair compositions. Electrophoretic mobility shift assays (EMSA) confirmed that the interaction occurs with high affinity ($K_d \approx 10^{-9}$ M; see below).

Structure of RBMY RRM in complex with S1A RNA

To understand the molecular basis of the recognition by hRBMY RRM for the selected RNA stem–loops, we investigated the structure of one complex using nuclear magnetic resonance (NMR; supplementary Fig S2 online). The chosen RNA ('S1A') was derived from the S1 sequence. S1A is 21 nt-long and contains eight base pairs capped by a CACAA loop (supplementary Fig S2A online). The stem contains the most common SELEX sequence, GUC–loop–GAC.

The structure of the complex was determined with very high precision because a structure calculation was used with a very high number of nuclear Overhauser effect (NOE) constraints

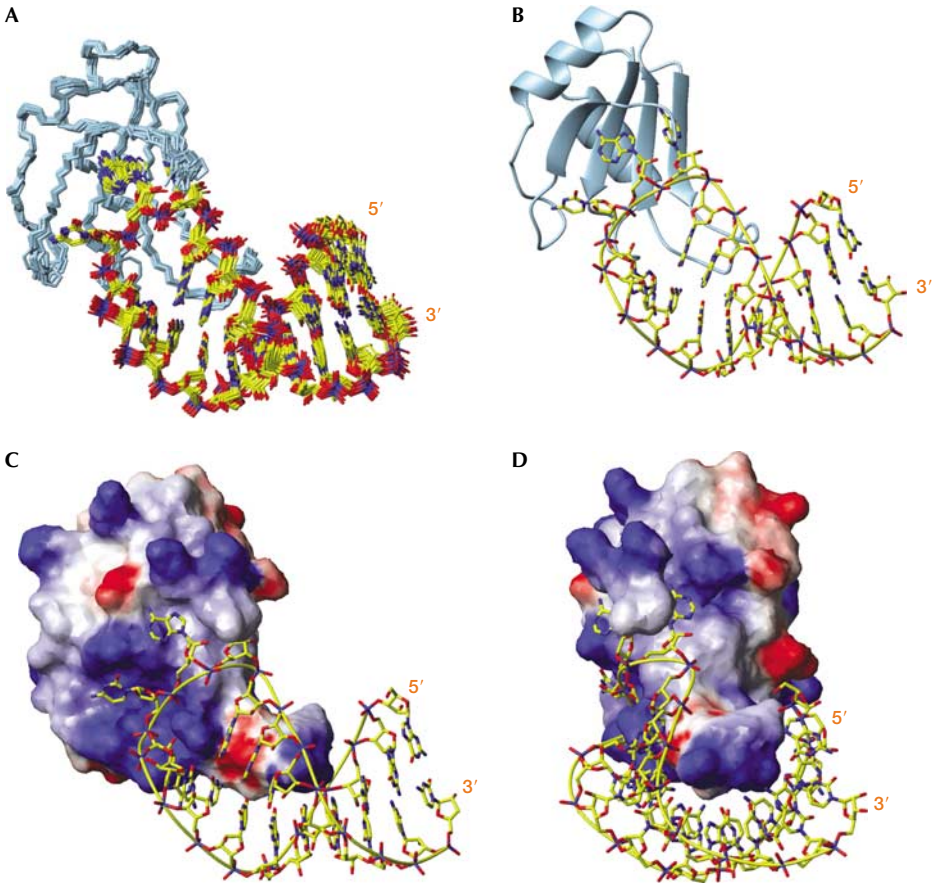


Fig 2 | Overall view of the complex between the human RBMY RRM and S1A RNA. (**A**) The 17 lowest energy structures of the complex are superimposed. (**B**) View of the lowest energy structure of the ensemble. (**C**) Surface representation of the protein in the complex and (**D**) rotated by 90°. RBMY, RNA-binding motif gene on Y chromosome; RRM, RNA recognition motif.

(1,879 including 124 intermolecular), supplemented by 54 angle constraints from residual dipolar couplings in both the RNA and the protein (supplementary Table SI online; Fig 2A). Human RBMY RRM adopts the expected βαββαβ topology, with two α-helices packed against one side of the four-stranded β-sheet (Maris *et al*, 2005).

Human RBMY RRM specifically recognizes a CAA triplet

C11, *A12* and *A13* (nucleotides are written in italics to distinguish them from amino acids) protrude from the loop and are spread on the RRM β-sheet surface (Fig 2B). The three nucleotides are stabilized by contact with the protein main chain and several side chains but not by intra-RNA interactions (Fig 3). *A12* adopts an unusual *syn* conformation and the sugar puckers of *C11* and *A13* are both C2′-*endo*. There are specific contacts with *A12* and *A13*, which discriminate for adenines at these positions. *A13* is recognized by K84 main-chain oxygen and by K9 (β1) and L38 (β2) side chains (Fig 3C). Similarly, *A12* is specifically recognized by the K84 main-chain amide (Fig 3B), the Q82 (β4) main-chain oxygen and the E81 (β4) side chain (K84 amide experiences a large chemical shift change on complex formation; supplementary Fig S2B online). *C11* is recognized by a contact with the K79 (β4) side chain (Fig 3A). From the structure, it appears that a *U* could also be accommodated at this position.

The β2–β3 loop inserts into the RNA major groove

An unexpected structural feature of the complex is the interaction of the RRM β2–β3 loop with the major groove of the RNA helix from the second base pair to the last base pair of the stem (Figs 2C,D,4A). The seven amino acids of the β2–β3 loop (D42–R48) form a β-hairpin that inserts itself into the deep major groove of the RNA helix. Most of the intermolecular interactions are non-sequence-specific in nature and involve side-chain and main-chain contact with phosphate groups of the RNA (Fig 4A). All five side chains from D42 to D46 are involved in the recognition of the RNA major groove. The R43 side chain forms salt bridges with *A3* and *C4* phosphate, whereas its main-chain amide forms contacts with *A15* phosphate on the opposite strand of the helix. The K46 side chain forms a salt bridge with *C16* phosphate (Fig 4A). The R43 and K46 side chains, located on opposite strands of the β2–β3 loop, cross each other to contact on the phosphate oxygen with the different strands of the helix (Fig 4B). Contacts by T44 to *G2* phosphate, and possibly by S45, further stabilize the interaction. Finally, the side chain of D42 forms two hydrogen

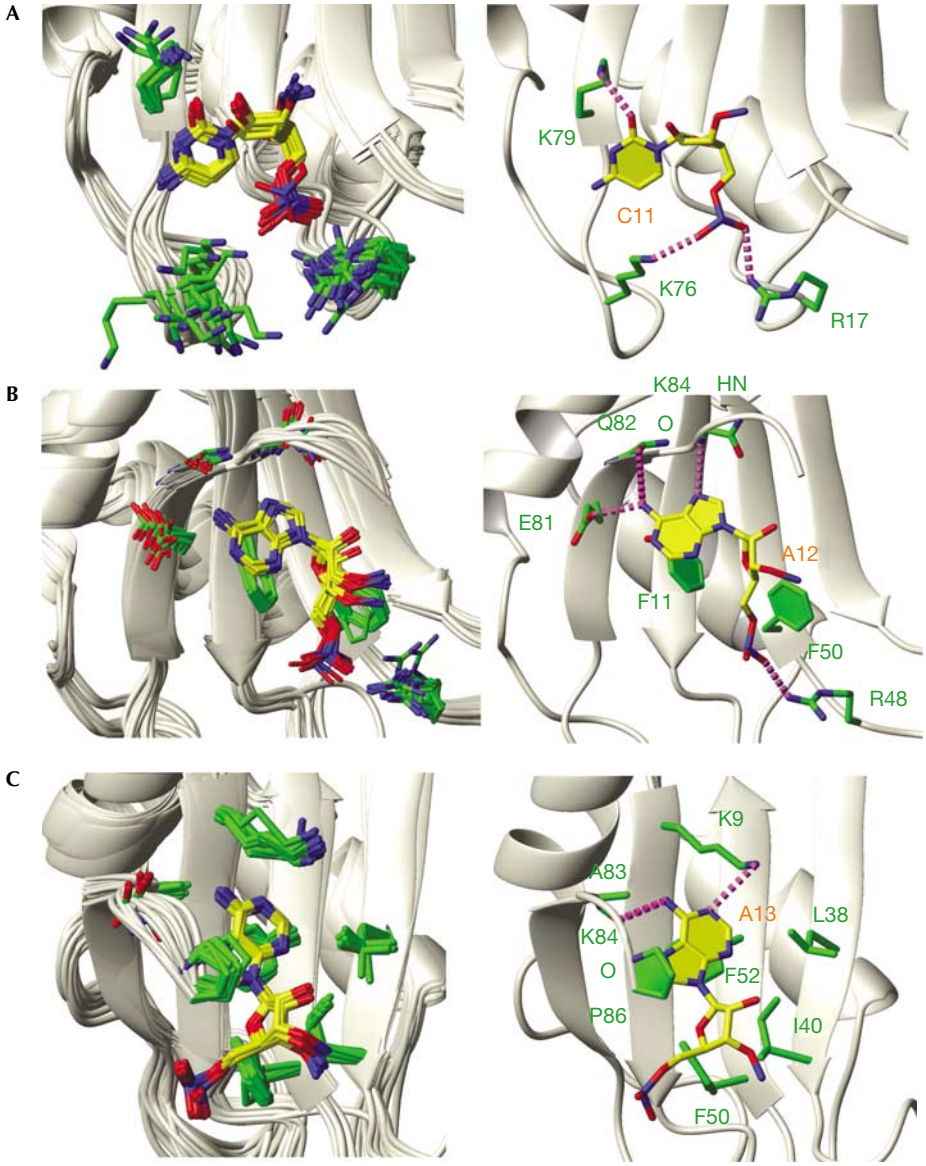


Fig 3 | Close-up of the interactions between the RRM β-sheet and *C11*, *A12* and *A13* of the RNA. (**A**) The ensemble of complex structures are superimposed around *C11* (left) and a view of the most representative structure (right). (**B**) As in (**A**), but around *A12*. (**C**) As in (**A**), but around *A13*. Protein side chains are shown in green and RNA residues are shown in yellow. Potential hydrogen bonds are represented by purple dotted lines. RRM, RNA recognition motif.

bonds, one with the main-chain amide of R48 and the other with *A15* amino (Fig 4A).

RBMY RRM stabilizes the 5′ end of the pentaloop

C9 and *A10* extend the RNA helical stack, as *C9* stacks over *C8*, and *A10* over *C9*. In addition, *C9* forms contacts with *C8* phosphate (Fig 4C) and the R17 side chain (β1–α1 loop) stacks over *A10* and forms contacts with *C11* phosphate. *C11* phosphate is also in contact with K76 (α2–β4 loop, Fig 4C). Human RBMY binding selectivity for *C9* and *A10* seems to be indirect; *C9* might be preferred to other nucleotides to prevent the formation of a base pair with *A13* in the free RNA and *A10* might be preferred to a *G* as a *G* amino would clash sterically with the protein backbone. Finally, R48 further stabilizes the complex by forming contacts with all three RNA elements (Fig 4C): the stem (*G14*), the CAA triplet (phosphate of *A12*), and *C9* or *A10*.

Mutagenesis studies confirm the recognition mode

We carried out EMSA experiments using RNA sequences that are representative of the two sets of sequences identified by SELEX. Sequences 2, 4 and 6 from set I (data not shown), S1 and S2 from set II, as well as S1A, all showed efficient binding with the GST-fused hRBMY RRM (Fig 5A). The apparent K_d was between 0.6 and 0.9 nM for S1 and S2, respectively. Mutating *A12* or *A13* to a

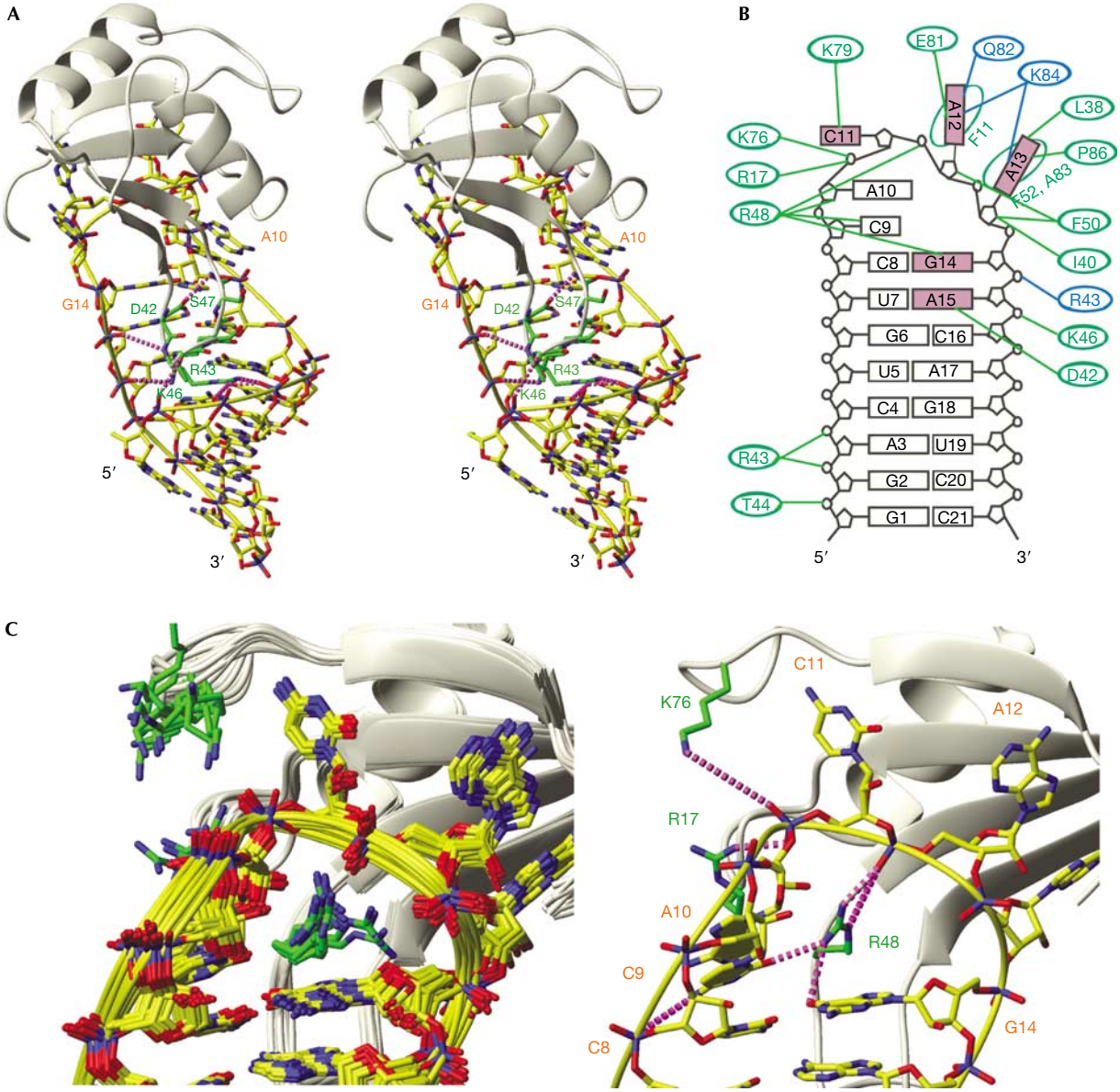


Fig 4 | View of the protein–RNA interactions between RRM loops and the RNA, and schematic diagram of the interactions. (**A**) Stereoviews of the contact between the RRM β2−β3 loop and the RNA stem. (**B**) Schematic diagram of interactions. The amino-acid residues in contact with the RNA side chain and main chain are shown in green and blue, respectively. The nucleotides shown in purple are sequence specifically recognized by human RBMY RRM. (**C**) The ensemble of complex structures are superimposed around *C9*, *A10* and *C11* (left), and a view of the most representative structure. RBMY, RNA-binding motif gene on Y chromosome; RRM, RNA recognition motif.

G resulted in a complete loss of binding, showing that the identity of these residues is crucial for high affinity (Fig 5B). By contrast, the replacement of *C11* by a *U* (*C11U*) was better tolerated as it resulted in an approximately fourfold decrease in affinity. This agrees well with the structure, as the identity of *C11* is not recognized as strongly as that of *A12* or *A13*. Changing the size of the pentaloop also has a strong effect, as the removal of the A/U residue (*A10*) resulted in very weak binding. Similarly, the insertion of an additional *A* (+*A13*) resulted in a complete loss of binding (Fig 5B). These experiments confirm that a loop size of five nucleotides is optimal for hRBMY binding. Finally, we tested whether the hydrogen bond between D42 and *A15* was crucial by replacing *U7-A15* by a *C7-G15* base pair (Fig 5B). Surprisingly, this change in the sequence did not alter the affinity. It is possible that a slight rearrangement could take place, allowing D42 to interact with *C7* in the mutant RNA instead of *A15* in the wild type. This mutation indicates that the contact mediated by D42 is not sequence specific. Thus, the interactions mediated by the β2–β3 loop would constitute a shape-specific recognition of the RNA major groove.

The structure shows that the β2–β3 loop fits into the major groove of the RNA, and that there are steric constraints associated

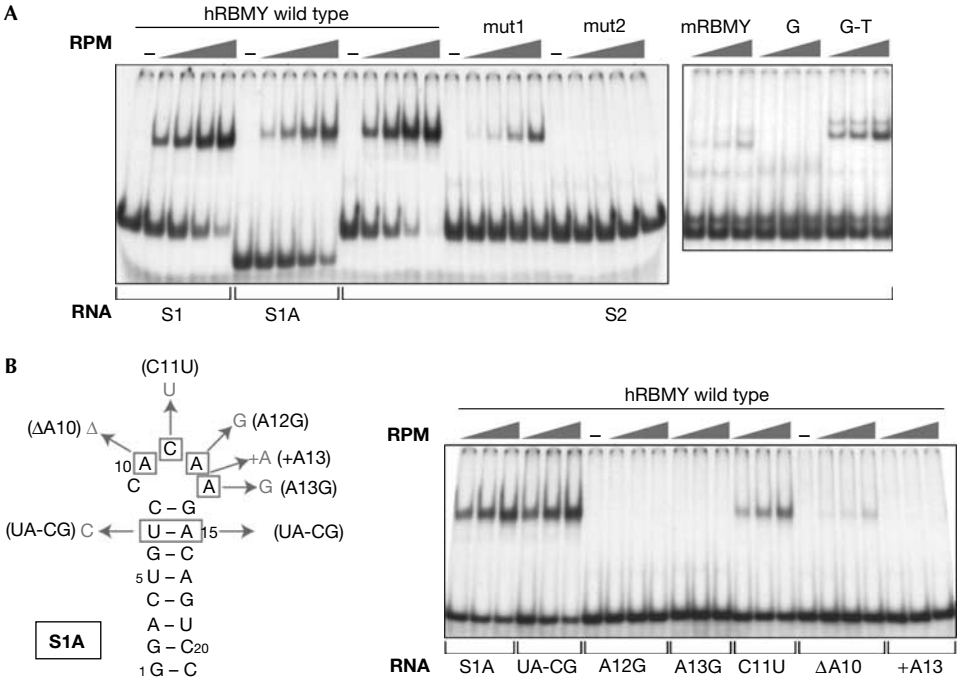


Fig 5 | SELEX RNA sequences are recognized specifically by human RBMY RRM. (**A**) The β2−β3 loop of human RBMY (hRBMY) RRM defines the specificity of interaction with its RNA targets. Increasing concentrations of the different RRMs (0.625, 1.25, 2.5 and 5 nM for wild-type hRBMY, 1.25, 2.5, 5 and 10 nM for mutant hRBMY, 2.5, 5 and 10 nM for mRBMY and heteronuclear RNP G and G-T) were incubated with RNA and the complexes were separated by EMSA. (**B**) The sequence of the RNA pentaloop is a crucial determinant of the interaction with the hRBMY RRM. The different mutations introduced in the S1A RNA are shown in grey boxes. For EMSA, the same concentrations of hRBMY RRM (2.5, 5 and 10 nM) were added to a constant concentration of wild-type or mutant RNA, and resulting complexes were analysed as in (**A**). EMSA, electrophoretic mobility shift assay; RBMY, RNA-binding motif gene on Y chromosome; RRM, RNA recognition motif; SELEX, systematic evolution of ligands by exponential enrichment.

with this insertion. Interestingly, the RRM of the mRBMY—as well as those of the human hnRNP G and G-T—shows two changes in the β2–β3 loop compared with that of hRBMY (Fig 1A): the three RRMs contain an additional E between R43 and T44, and S45 of hRBMY is replaced by K in mRBMY or N in hnRNP G and G-T. To study the role of these residues, two hRBMY RRM mutants (mut1 and mut2; Fig 1A) were made and their affinity was compared with the wild-type RRM (Fig 5A, left panel). Strikingly, the simple insertion of an E in the β2–β3 loop resulted in more than a tenfold decrease in binding. Furthermore, when the additional S to K mutation was carried out—to make a mouse-like β2–β3 loop—no detectable interaction was observed (Fig 5A). This confirms the crucial role of the β2–β3 loop of the hRBMY RRM for binding RNA stem–loops. We also showed that the three RRMs of mRBMY, human hnRNP G and G-T could not efficiently bind to the S2 stem–loop (Fig 5A, right panel).

DISCUSSION

A dual mode of RNA recognition by human RBMY

Using SELEX, we have identified an unusual RNA binding consensus sequence for hRBMY RRM. The high-affinity sites are RNA stem–loops with a $C^{A}/_{U}CAA$ loop and a GUC–loop–GAY consensus in the last three base pairs of the stem (Fig 1B). The structure of the RRM complexed with a stem–loop containing a CACAA loop showed that the recognition of the RNA is both sequence- and shape-specific (Stefl *et al*, 2005). The structure explains how *C9, A12, A13* and *G14* are sequence-specifically recognized by the RRM (Figs 3,4), confirming and explaining the SELEX consensus in the pentaloop and at the first base pair. However, from a structural point of view, it remains unclear why an $^{A}/_{U}$ and a C are preferred in the second and third positions, respectively of the loop in the SELEX sequence. It is possible that other base types in these two positions would modify the accessibility of the last pentaloop triplet or induce a folding of RNA that might prevent protein binding or lower the binding affinity. More surprisingly, we found that the RNA stem is recognized by the RRM β2–β3 loop, which forms a β-hairpin and is inserted into the major groove of the RNA helix. This recognition is shape specific, as it is the complementary shape and charge between the β2–β3 loop and the RNA major groove that dictate this intermolecular interaction (Fig 2). When this β2–β3 loop is elongated or mutated, RNA binding is weakened or lost, confirming the importance of this interaction for complex formation. These two modes of binding result from an 'induced fit' of one of the binding partners. In the sequence-specific mode, the unstructured RNA pentaloop folds after binding on the rigid β-sheet and becomes ordered. In the shape-recognition mode, the flexible β2–β3 loop inserts into the major groove of the rigid RNA stem and becomes ordered (L.S. & F.A., unpublished data). This mode of recognition is unprecedented among RRMs, once again

confirming the remarkable plasticity of this RNA recognition motif (see also the supplementary information online).

RNA recognition by other hnRNP-G family members

We have shown that the RRMs of human hnRNP G, G-T or mRBMY are unable to bind with high affinity to the hRBMY-specific hairpin structure. This is in good agreement with our structural data, when considering the role of the β2–β3 loop in strengthening the interaction between hRBMY and the RNA (Figs 4,5). However, as K9, L38 and E81 are conserved between hnRNP G and hRBMY, it is possible that hnRNP G RRM can similarly recognize CAA in a sequence-specific manner. Interestingly, one RNA sequence identified as a potential target for hnRNP G contains a CAA triplet (Nasim *et al*, 2003). By contrast, the replacement of E81 in hnRNP G-T and mRBMY (by A or K, respectively) is likely to impair binding to CAA. So far, little data are available concerning the RNA-recognition properties of RBMY and related hnRNP proteins. A previous study indicated that hRBMY and hnRNP G are nonspecific RNA-binding proteins (Hofmann & Wirth, 2002). We show here that this might not be the case, but further analyses are necessary to understand better the properties of all these related RRMs.

Finally, the particular RNA-binding properties of hRBMY could have an important role in the function of the protein in human testes. The infertility caused by deletions in this RRM-encoding gene (Elliott *et al*, 1997) indicate that crucial RNA processing pathways are disrupted. Protein–protein interactions and sub-nuclear localization experiments suggest a role for RBMY in splicing, although its precise function is still not clear. Testis is one of the tissues in which alternative splicing is largely used (Xu *et al*, 2002; Yeo *et al*, 2004), and is needed to establish the exclusive pattern of gene expression that occurs throughout the different stages of spermatogenesis (Venables, 2002). A recent study carried out with 52 different tissues and more than 10,000 genes showed that testis has the highest rate of divergence in alternative splicing events between human and mouse (Kan *et al*, 2005). Our findings might explain why the mRBMY seems to have a function different from that of the hRBMY (Szot *et al*, 2003), as both proteins might have different RNA targets. However, we cannot rule out that the natural hRBMY target sequences differ significantly from the stem–loop motifs we identified by SELEX. As a first step towards the identification of the biological RNA targets of hRBMY *in vivo*, we screened an alternative exon database using an algorithm based on the conservation of the stem–loop structure we characterized by SELEX (supplementary information online). This screening showed that putative hRBMY targets indeed exist within or in the vicinity of exons alternatively spliced in the testis (supplementary Table S2 online). Further work is necessary to analyse whether these RNA sequences are evolutionarily conserved in mammals and whether they are functionally relevant.

METHODS

SELEX and EMSA. SELEX was carried out as described previously (Cavaloc *et al*, 1999), with only minor modifications using the GST-fused RRM of hRBMY (amino acids 1–108; for details see the supplementary information online).

For EMSA, we used the GST-fused RRMs of hRBMY, mRBMY, hnRNP G or hnRNP G-T, as well as the two mutants of hRBMY. [^{32}P]RNA was transcribed *in vitro* with T7 RNA polymerase and incubated with the different proteins as described previously (Cavaloc *et al*, 1999).

Cloning, expression and purification of hRBMY RRM for NMR. The N-terminal RRM (amino acids 1–108) of the hRBMY was subcloned in pET30a+ (Invitrogen Corp., Carlsbad, CA, USA). The construction included a two-residue linker (L-E) between the RRM and the 6 × His tag. For ^{15}N and ^{15}N,^{13}C-labelling, *Escherichia coli* BL21(DE3)pLysS was used. Expression was carried out in M9 media containing [^{15}N]NH_4Cl and [^{13}C]glucose. Human RBMY RRM was purified by Ni affinity and cation exchange chromatography. The protein solution was concentrated to 1 mM, as measured by UV spectroscopy at 205 nm.

RNA transcription and complex formation for NMR analysis. S1A RNA was prepared *in vitro* using T7 RNA polymerase and purified by anion exchange chromatography. RNA samples were dissolved in 25 mM NaH_2PO_4/NaOH at pH 7.0.

The hRBMY RRM–S1A complex was studied at 1 mM at a 1:1 ratio of protein and RNA in 25 mM NaH_2PO_4/NaOH and 25 mM NaCl buffer (pH 7.0; see the supplementary information online for the resonance assignment strategy).

Structure determination. In total, 124 intermolecular NOEs between the RRM and S1A were assigned. The RDCs of hRBMY RRM–S1A were obtained using the Pf1 phages as an aligning medium.

The structure determination was carried out as described previously (Oberstrass *et al*, 2006). The 17 final conformers with the lowest total energy or with the lowest alignment tensor energy were selected to form the final ensemble of conformers.

Structural data. All restraints used in structure determination and the 17 final structures have been deposited at the Protein Data Bank under the accession code 2FY1.

Supplementary information is available at *EMBO reports* online (http://www.emboreports.org).

ACKNOWLEDGEMENTS

This investigation was supported by the Swiss National Science Foundation—National Center of Competence in Research (SNF-NCCR) Structural Biology, by the Roche Research Fund for Biology at the Eidgenössische Technische Hochschule Zürich (ETH) Zürich and by the ETH Zürich (TH- Fonds Nr. 0-20960-01) to F.H.T.A., and by grants from Inserm, Centre National de la Recherche Scientifique (CNRS) the European Union Network of Excellence on Alternative Splicing (EURASNET, 6th Framework Program) and the Association pour la Recherche sur le Cancer to J.S., and an Human Frontier Science Program (HSFP) postdoctoral fellowship to R.S. F.H.T.A. is an EMBO Young Investigator.

REFERENCES

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

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

Delbridge ML, Lingenfelter PA, Disteche CM, Graves JA (1999) The candidate spermatogenesis gene RBMY has a homologue on the human X chromosome. *Nat Genet* **22:** 223–224

Elliott DJ (2004) The role of potential splicing factors including RBMY, RBMX, hnRNPG-T and STAR proteins in spermatogenesis. *Int J Androl* **27:** 328–334

Elliott DJ, Bourgeois CF, Klink A, Stevenin J, Cooke HJ (2000a) A mammalian germ cell-specific RNA-binding protein interacts with ubiquitously expressed proteins involved in splice site selection. *Proc Natl Acad Sci USA* **97:** 5717–5722

Elliott DJ, Millar MR, Oghene K, Ross A, Kiesewetter F, Pryor J, McIntyre M, Hargreave TB, Saunders PTK, Vogt PH, Chandley AC, Cooke H (1997) Expression of RBM in the nuclei of human germ cells is dependent on a critical region of the Y chromosome long arm. *Proc Natl Acad Sci USA* **94:** 3848–3853

Elliott DJ, Venables JP, Newton CS, Lawson D, Boyle S, Eperon IC, Cooke HJ (2000b) An evolutionarily conserved germ cell-specific hnRNP is encoded by a retrotransposed gene. *Hum Mol Genet* **9:** 2117–2124

Hofmann Y, Wirth B (2002) hnRNP-G promotes exon 7 inclusion of survival motor neuron (SMN) via direct interaction with Htra2-β1. *Hum Mol Genet* **11:** 2037–2049

Kan Z, Garrett-Engele PW, Johnson JM, Castle JC (2005) Evolutionarily conserved and diverged alternative splicing events show different expression and functional profiles. *Nucleic Acids Res* **33:** 5659–5666

Ma K *et al* (1993) A Y chromosome gene family with RNA-binding protein homology: candidates for the azoospermia factor AZF controlling human spermatogenesis. *Cell* **75:** 1287–1295

Mahadevaiah SK *et al* (1998) Mouse homologues of the human AZF candidate gene RBM are expressed in spermatogonia and spermatids, and map to a Y chromosome deletion interval associated with a high incidence of sperm abnormalities. *Hum Mol Genet* **7:** 715–727

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

Nasim MT, Chernova TK, Chowdhury HM, Yue B-G, Eperon IC (2003) HnRNP G and Tra2β: opposite effects on splicing matched by antagonism in RNA binding. *Hum Mol Genet* **12:** 1337–1348

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

Stefl R, Skrisovska L, Allain FHT (2005) RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. *EMBO Rep* **6:** 33–38

Szot M, Grigoriev V, Mahadevaiah SK, Ojarikre OA, Tour A, Von Glasenapp E, Rattigan A, Turner JMA, Elliott DJ, Burgoyne PS (2003) Does Rbmy have a role in sperm development in mice? *Cytogenetic and Genome Research* **103:** 330–336

Venables JP (2002) Alternative splicing in the testes. *Curr Opin Genet Dev* **12:** 615–619

Venables JP, Elliott DJ, Makarova OV, Makarov EM, Cooke HJ, Eperon IC (2000) RBMY, a probable human spermatogene. *Hum Mol Genet* **9:** 685–694

Venables JP, Vernet C, Chew SL, Elliott DJ, Cowmeadow RB, Wu J, Cooke HJ, Artzt K, Eperon IC (1999) T-STAR/ETOILE: a novel relative of SAM68 that interacts with an RNA-binding protein implicated in spermatogenesis. *Hum Mol Genet* **8:** 959–969

Xu Q, Modrek B, Lee C (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res* **30:** 3754–3766

Yeo G, Holste D, Kreiman G, Burge CB (2004) Variation in alternative splicing across human tissues. *Genome Biol* **5:** R74

nature structural & molecular biology

Shape-specific recognition in the structure of the Vts1p SAM domain with RNA

Florian C Oberstrass[1], Albert Lee[2], Richard Stefl[1], Michael Janis[2], Guillaume Chanfreau[2] & Frédéric H-T Allain[1]

Although the abundant sterile alpha motif (SAM) domain was originally classified as a protein-protein interaction domain, it has recently been shown that certain SAM domains have the ability to bind RNA, defining a new type of post-transcriptional gene regulator. To further understand the function of SAM-RNA recognition, we determined the solution structures of the SAM domain of the *Saccharomyces cerevisiae* Vts1p (Vts1p-SAM) and the Smaug response element (SRE) stem-loop RNA as a complex and in isolation. The structures show that Vts1p-SAM recognizes predominantly the shape of the SRE rather than its sequence, with the exception of a G located at the tip of the pentaloop. Using microarray gene profiling, we identified several genes in *S. cerevisiae* that seem to be regulated by Vts1p and contain one or more copies of the SRE.

The SAM domain is one of the most abundant eukaryotic protein motifs mediating protein-protein interactions that are involved in transcription regulation and signal-transduction cascades[1,2]. For example, the SAM domain of the transcription activator ETS-1 provides a docking site for the mitogen-activated protein (MAP) kinase ERK2, which phosphorylates ETS-1 (ref. 3). The SAM domain of TEL serves as a self-association motif that can either activate kinase catalytic domains or repress transcription[4]. However, there is increasing evidence that SAM domains can also bind nucleic acids. The RuvABC complex processes the Holliday junction intermediate in prokaryotic cells, with the SAM domain of RuvA binding B-DNA[5]. More recently, a subclass of SAM domains have been identified with RNA binding abilities[6,7]. It has been shown that the SAM domains of *S. cerevisiae* Vts1p[8] (Vts1p-SAM) and of its homolog Smaug from *Drosophila melanogaster* specifically bind an RNA stem-loop with high affinity[6]. This Smaug recognition element (SRE) consists of an RNA stem-loop containing a pentaloop with a CNGGN sequence, where the Ns can be any nucleotides. Disruption of the stem or single mutations in three essential loop nucleotides have large effects on RNA binding by the SAM domain of Smaug or Vts1p[6].

In *D. melanogaster*, Smaug recognizes the SRE present in the 3′ untranslated region (UTR) of *nos* transcripts[9,10] and controls embryonic axis formation by inhibiting translation via CUP and eIF4E interaction with non–posterior-localized *nos* messenger RNA[11]. Disruption of SRE recognition by Smaug leads to translation of *nos* mRNA that is not properly localized, in turn resulting in lethal anterior body-patterning defects[9]. In addition, Smaug seems to be involved in another distinct gene-expression regulation pathway. In early *D. melanogaster* embryos, Smaug recruits the CCR4–POP2–NOT deadenylase complex to trigger localization of maternal transcripts[12]. A similar mechanism has been observed in *S. cerevisiae*. The yeast homolog of Smaug, Vts1p, can induce the degradation of a green fluorescent protein (GFP) reporter mRNA harboring SREs in its 3′ UTR, thereby lowering its expression level. Mutations in the SRE pentaloop that disrupt binding of Vts1p abolish this effect[6]. Furthermore, this SRE-dependent repression of the GFP reporter construct is abolished in Δ*CCR4* cells. As the CCR4–POP2–NOT complex has been shown to have deadenylase activity[13,14], a mechanism has been proposed in which Vts1p influences stability of mRNA transcripts by regulating poly(A) tail length in a CCR4-dependent manner[6]. However, in *S. cerevisiae*, no natural gene targets for Vts1p have so far been identified.

To gain a better knowledge of the mechanism by which SAM domains recognize RNA, we have determined the solution NMR structure of the *S. cerevisiae* Vts1p-SAM domain and the 23-nucleotide SRE stem-loop RNA, containing a CUGGC pentaloop, both in their free states and in complex. These three structures reveal the molecular basis of SAM-SRE recognition. Both the protein and the RNA are conformationally preorganized in the free forms, and recognition occurs by a rigid body fit. The SAM domain seems to recognize primarily the shape of the RNA rather than its sequence; only the G in position 3 in the loop seems to be specifically recognized. In addition, we set out to further delineate the function of Vts1p in yeast. Using microarray technology, we identified several genes in *S. cerevisiae* that are strongly upregulated in the *vts1*Δ strain. Characterization of these genes by bioinformatics showed that many contain one to several copies of the SRE.

[1]Institute for Molecular Biology and Biophysics, ETH Zürich, CH-8093 Zurich, Switzerland. [2]Department of Chemistry and Biochemistry, University of California at Los Angeles, Box 951569, 607 Charles E. Young Drive East, Los Angeles, California 90095-1569, USA. Correspondence should be addressed to F.H.-T.A. (allain@mol.biol.ethz.ch).

Received 30 August 2005; accepted 15 November 2005; published online 22 January 2006; doi:10.1038/nsmb1038

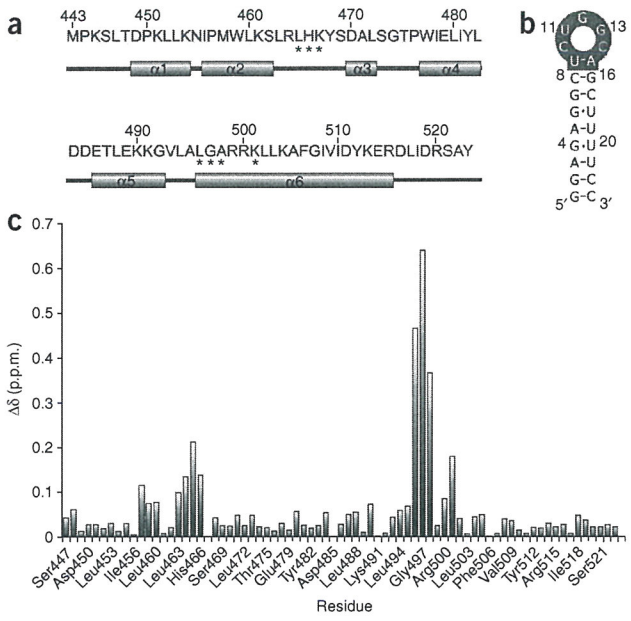


Figure 1 Domain structure of Vts1p-SAM, secondary structure of the SRE RNA and chemical shift difference mapping upon formation of the complex. (**a**) Primary sequences of the Vts1p-SAM constructs, with secondary structure elements shown. Residues with the largest chemical shift changes are indicated by stars. (**b**) Secondary structure of the SRE RNA used in our study. (**c**) Plot of the chemical shift difference between amide groups of the free and of the bound form of Vts1p-SAM ($\Delta\delta = [\Delta\delta_{HN}^2 + (\Delta\delta_N^2/R_{scale})]^{1/2}$, $R_{scale} = 6.5$)[44].

RESULTS

Structure of the free Vts1p SAM domain

The *S. cerevisiae* Vts1p-SAM (residues 443–523, **Fig. 1**) in its free form gave rise to a well-dispersed NMR spectrum with good linewidth (**Supplementary Fig. 1** online), indicating that the protein is folded also in the absence of RNA. The solution structure of the free protein was determined using 2,366 NOE-derived distance constraints, leading to a precise structure with an r.m.s. deviation of 0.32 Å for the protein backbone (**Supplementary Fig. 2** online). Vts1p-SAM adopts a globular fold made of six helices that are packed by a hydrophobic core similar to the one found in other SAM domains, such as those of EphB2, Byr2 and p73 (**Fig. 2a**). On its surface, Vts1p-SAM displays a very large electropositive region formed by three lysines (Lys461, Lys467 and Lys501), two arginines (Arg464 and Arg500) and one histidine (His466). This positively charged patch might be important for RNA binding (**Fig. 2b**). The large electropositive region surrounds a small hydrophobic cavity (Leu463, Leu465, Tyr468, Val493, Ala495, Ala498) located between helices 2 and 3 and the N-terminal part of helix 6. This cavity is about the size of a purine base. Overall, the structure of the SAM domain of Vts1p resembles the previously determined structure of that of Smaug[7]. The major difference resides in helix 6 of the Smaug SAM domain, which extends and contacts the PHAT domain, a protein domain required for the stabilization of the SAM domain of Smaug[7].

Structure of the free SRE RNA

The SRE used in this study is a 23-nucleotide RNA stem-loop containing a CUGGC pentaloop (**Fig. 1** and **Supplementary Fig. 1**) that has previously been shown to bind Vts1p-SAM with high affinity[6]. The sequence of the stem is identical to the native one, except that it starts with two G-C base pairs to generate a high level of transcription. The structure of the SRE RNA is well defined (r.m.s. deviation for all heavy atoms of 1.00 Å), with a highly structured loop (**Supplementary Fig. 2**). The pentaloop adopts a rather unusual fold. The first and fourth residues in the loop, C10 and G13, form a Watson-Crick base pair, extending the stem (**Fig. 2a,c**). C14 is flipped out and seems rather flexible (**Fig. 2c** and **Supplementary Fig. 2**). The base of the third residue in the loop, G12, stacks on the C10-G13 base pair, capping the stem. G12 is further stabilized in this position by a hydrogen bond between its N7 and the 2′ hydroxyl of C10. The U11 base lies in the minor groove of the stem, where it is stabilized by hydrophobic contacts between U11's H5 and H6 and the sugar moiety of C10. The hairpin turn is achieved by only two nucleotides, U11 and G12 (**Fig. 2c**). Thus, the SRE pentaloop resembles a tetraloop with a topology that is similar but not identical to a CUYG tetraloop. One difference is that the SRE pentaloop has an additional nucleotide (C14). However, C14 seems to be flexible and not to contribute to the fold of the RNA. Further differences between the SRE and the CUYG tetraloop are that at position 3 in the SRE loop there is a guanine (G12) instead of a pyrimidine and that in the SRE, the loop is closed by a U-A base pair instead of the C-G base pair that usually closes CUYG tetraloops to increase their stability (reviewed in ref. 15). Thus, our structure seems to show an extension of the CUYG tetraloop fold.

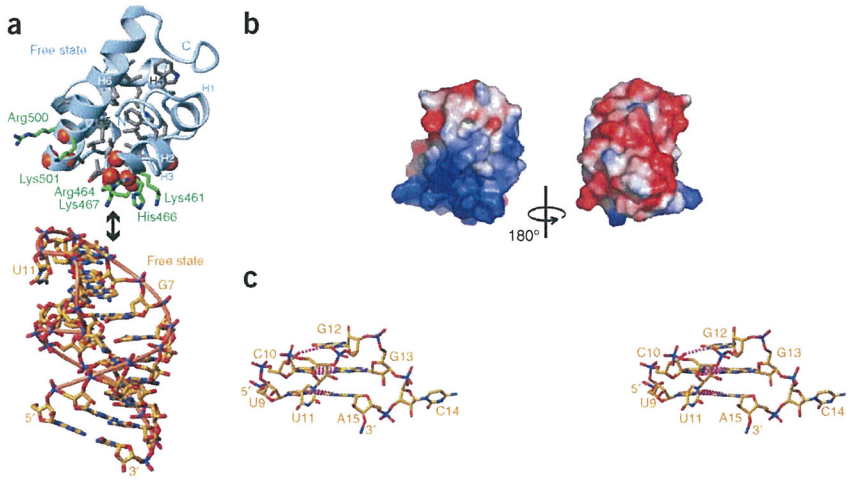


Figure 2 Structure of the free Vts1p-SAM domain and of the free SRE RNA. (**a**) Top, the most representative conformer of the free Vts1p SAM domain. Red balls, backbone amides undergoing large chemical shift perturbation upon RNA binding; green, protein side chains interacting with the RNA; gray, hydrophobic core of Vts1p-SAM. Bottom, the free SRE. Arrow, rigid body interaction. (**b**) Solvent-accessible surface representation of Vts1p-SAM colored by electrostatic potential (blue, positive; red, negative). Orientation at left is as in **a**. (**c**) Stereo view of the pentaloop and the closing base pair of the free SRE RNA (U9 to A15). Magenta dashed lines, hydrogen bonds.

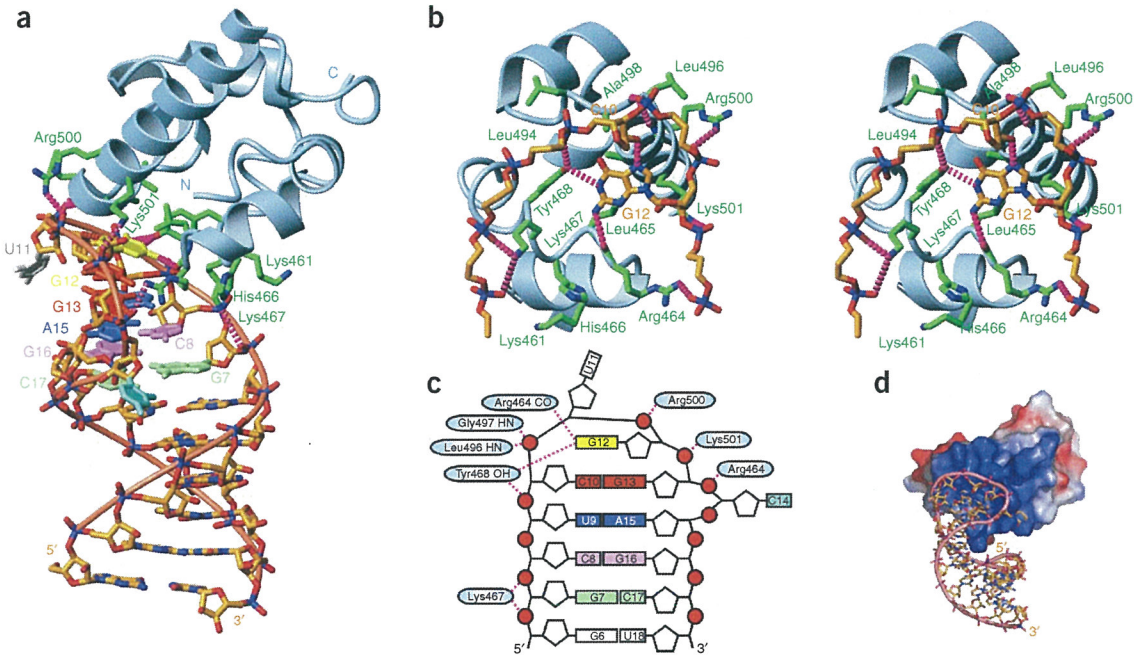


Figure 3 Solution structure of Vts1p-SAM domain bound to SRE RNA and molecular basis of the recognition. (**a**) Vts1p-SAM–SRE complex. Green, side chains important for recognition; magenta dashed lines, possible hydrogen bonds. (**b**) Stereo view of the most representative structure showing all the interactions important for recognition, colored as in **a**. (**c**) Schematic representation of the intermolecular interactions, colored as in **a**. (**d**) Surface representation of the protein in complex. Vts1p-SAM is colored by electrostatic potential (blue, positive; red, negative) and the SRE RNA is shown in stick representation.

NMR structure of the Vts1p-SAM–SRE complex

Upon RNA binding, large chemical shift perturbations in the amide resonances of Vts1p-SAM were observed. They cluster in two regions: Leu463–Lys467 and Leu496–Lys501 (**Fig. 1c** and **Supplementary Fig. 1**). These two segments of Vts1p-SAM interact with one another and form a small hydrophobic cavity (Leu463, Leu465, Tyr468, Val493, Ala495, Ala498) surrounded by positively charged residues (Lys461, Arg464, His466, Lys467, Arg500, Lys501) (**Fig. 3**). The structure of the complex confirms that these two regions of Vts1p-SAM participate in the recognition of the SRE RNA (**Fig. 3a**). Using 2,154 NOE-derived distance constraints (including 48 intermolecular NOEs) and 111 residual dipolar coupling (RDC) constraints originating from both the RNA and the protein, we determined the structure of Vts1p-SAM–SRE complex (r.m.s. deviation over all heavy atoms of 1.48 Å, **Supplementary Fig. 2**). The RDC constraints were essential to improve the mutual orientation of the protein and the RNA. To resolve ambiguous assignments of several resonances, we used two chemically synthesized RNAs that have specific nucleotide riboses labeled with ^{13}C. The highly simplified spectra obtained using these RNAs were essential to unambiguously identify a number of intermolecular NOEs (**Supplementary Fig. 3** online). The protein-RNA interface is well defined, allowing us to elucidate the molecular basis of Vts1p-SAM–SRE recognition (**Fig. 3b,c**).

Mechanism of Vts1p-SAM–SRE recognition

In the complex, Vts1p-SAM and the SRE RNA adopt the same structures as in their free forms. This suggests that binding occurs via a rigid body fit. Vts1p-SAM caps the pentaloop, binding in the major groove of the RNA (**Figs. 2** and **3**). G12, located on top of the RNA loop, has the highest number of interactions with the protein. G12 occupies a small hydrophobic cavity formed between Leu465 and Ala495, and its base identity is specifically recognized by two intermolecular hydrogen bonds. G12 amino and imino protons are hydrogen-bonded with the backbone carbonyl oxygen of Arg464 and the hydroxyl oxygen of Tyr468, respectively (**Fig. 3b,c**). The other protein-RNA interactions exclusively involve contacts to the RNA phosphate oxygens. Phosphate oxygens of G7 and C8 are contacted by Lys467, of C10 by the hydroxyl of Tyr468, of U11 by the backbone amides of Leu496 and Gly497, and of G12, G13 and C14 by the side chains of Arg500, Lys501 and Arg464, respectively (**Fig. 3b–d**). In addition, the dipole of helix 6 provides affinity, as the phosphate of U11 is placed just below its N terminus. The sum of all these interactions explains the high affinity determined for this complex ($K_d \approx 20$ nM)[6]. With most of the protein-RNA contacts involving the RNA sugar-phosphate backbone, whose structure is unchanged upon complex formation, the Vts1p-SAM–SRE recognition seems to be shape-specific rather than sequence-specific, with the exception of G12 recognition. In other words, the shape of the CUGGC pentaloop creates a unique electrostatic fingerprint that is recognized by the basic surface of Vts1p-SAM (**Fig. 3d**).

Assuming that rigid body interaction is required for recognition by Vts1p-SAM, the RNA in its free form must already be structurally preorganized to be bound by Vts1p. It has been demonstrated that a single mutation in C10, G12 or G13 drastically reduces affinity, whereas mutations of U11 or C14, or even deletion of C14 does not strongly affect binding[6]. Our structure confirms these observations. G12 identity is specifically recognized, and a single mutation of C10 or G13 would disrupt the Watson-Crick base pair that is formed in the loop and stabilizes the RNA fold. On the basis of our structure, an XNGY(N) loop capping a 4-base-pair stem seems to be the consensus

sequence required for high-affinity binding (where Ns are any nucleotides, X and Y form a Watson-Crick base pair and the parentheses indicate that a pentaloop or a tetraloop can be bound). To check the validity of this refined consensus sequence for Vts1p binding, we performed a binding study with SRE stem-loops containing a U10-A13, G10-C13 or A10-U13 base pair instead of the wild-type C10-G13. Vts1p binding to mutant and wild-type RNAs were analyzed by NMR spectroscopy and nondenaturing gel electrophoresis (**Supplementary Fig. 4** online). NMR titration experiments showed that Vts1p-SAM binds all three mutated stem-loop RNAs. The same set of amide resonances undergoes a large chemical shift change in both the wild-type SRE RNA and the three mutants, indicating that the same protein-recognition surface is involved in all four complexes (see **Supplementary Fig. 4** for the U10-A13 mutant RNA). The RNA and protein resonance linewidths of the complex with the SRE mutant U10-A13 are as sharp as the ones with the wild-type SRE (**Supplementary Fig. 4**), indicating that the affinities of the two complexes are very similar. However, for the other two SRE mutants, the NMR lines are much broader, indicating that the affinity constants for the latter two complexes are lower than that for the wild-type complex (data not shown). Qualitative mobility shift assays of all four complexes further confirm the conclusion obtained by NMR spectroscopy (**Supplementary Fig. 4**). A protein–RNA complex is formed with all four RNAs, but only complexes with wild-type and U10-A13 SREs show discrete bands, indicative of a stable complex. Altogether, these studies confirm the hypothesis, derived from the structure, that any Watson-Crick base pairs at positions 1 and 4 in the loop would permit VTS1p to bind. However, they also show that a pyrimidine and a purine at positions 1 and 4, respectively, are preferred. A possible explanation for this preference might be that with a purine in position 4 (A13 or G13), three purines in a row (G12, G13 or A13, and A15) are stacked. This all-purine stacking interaction would stabilize both the free RNA and the complex better than would a purine-pyrimidine-purine combination.

The structure of the complex is also very consistent with the results of mutagenesis studies previously performed on both Vts1p and Smaug. Many residues whose mutations greatly affect binding have side chains that interact with RNA in our complex[6,7]. The only exceptions are His466 and Lys461, whose equivalent residues in Smaug have been shown to contribute to RNA binding. These side chains are not in direct contact with the RNA in our complex structure; therefore, it is possible that they either affect the local geometry of the protein or participate in RNA binding via a water-mediated contact, as both are located within 5 Å of the RNA (**Fig. 3a**).

Microarray analysis of VTS1 deletion strain

To search for natural RNAs regulated by Vts1p, we performed microarray analysis on a yeast strain lacking Vts1p (*vts1Δ*) and on an isogenic wild-type strain. This analysis revealed that a subset of genes (29) were strongly upregulated in the absence of Vts1p, as expected if Vts1p were involved controlling the expression of these genes (**Supplementary Table 1** online). To discriminate between direct and indirect effects, we predicted the folds of the top mRNA candidates' open reading frames (ORFs) and flanking sequences using the Vienna package[16] and searched for stem-loops that would match our Vts1p-binding site consensus sequence. This analysis revealed that 6 of the 29 mRNAs from strongly upregulated genes in the *vts1Δ* strain contained a stem capped by a CNGGN pentaloop (**Fig. 4**). Moreover, when the consensus of the loop was extended to include tetraloops or pentaloops with a G in position 3 of the loop and any Watson-Crick base pair between the first and the fourth loop nucleotides, as

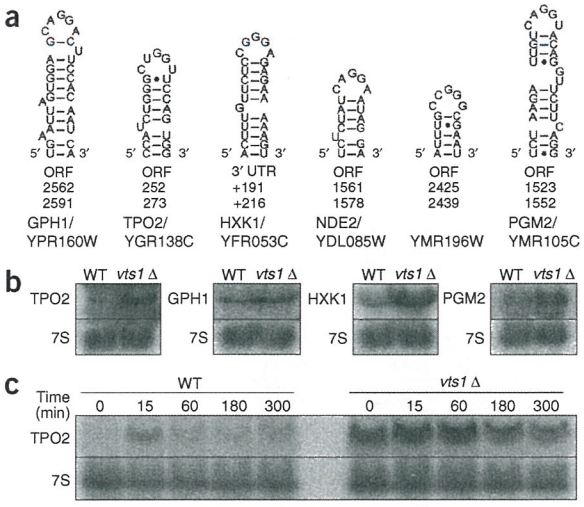


Figure 4 Genomic search for natural Vts1p substrates. (**a**) RNA stems capped by CNGG(N)-type pentaloops found in the mRNAs strongly upregulated in the *vts1Δ* strain compared to the wild-type strain. Indicated are gene names (if available) followed by *Saccharomyces* Genome Database accession codes. Biological functions of these transcripts are listed in **Supplementary Table 1**. (**b**) Northern analysis of transcripts containing CNGG(N)-type stem-loops. Shown are the levels of mRNAs expressed from isogenic wild-type and *vts1Δ* strains. The 7S noncoding RNA was used as a loading control. (**c**) Analysis of the expression of *TPO2* in wild-type and *vts1Δ* strains during a shift from stationary phase in minimal medium into rich medium. A sample was harvested in stationary phase in minimal medium (time 0) and subsequent samples were harvested at the indicated times after a shift to rich medium.

suggested by the structural analysis, we found six additional genes with a potential Vts1p-SAM–binding site and seven genes with multiple potential binding sites (**Supplementary Fig. 5** online).

Notably, a number of these structures were phylogenetically conserved in other yeast species such as *Saccharomyces paradoxus*, *Saccharomyces bayanus* and *Saccharomyces mikatae* (**Supplementary Fig. 6** online), although the sequences and the positions of the stems could slightly differ from those observed in *S. cerevisiae*. The phylogenetic conservation of some of these predicted RNA structures strongly suggest that they correspond to conserved regulatory signals bound by Vts1p-SAM. Other mRNAs upregulated in the *vts1Δ* strain did not show any apparent CNGGN pentaloops capping a short stem. The upregulation of these mRNAs may correspond to indirect effects, as some of these genes are frequently found to be upregulated in microarray studies (such as the heat shock protein HSP12).

Comparison of the frequencies at which CNGGN-type stem-loops are found in the mRNAs upregulated in the absence of Vts1p to the frequencies at which they are found genome-wide showed that these structures are enriched in the microarray set (see Methods and **Supplementary Table 2** online). The CNGGN pentaloop motif adjacent to a stem-loop of at least four canonical or G•U base pairs is seen four times more frequently in the genes upregulated in the absence of Vts1p than in the remaining genomic ORFs (**Supplementary Table 2**). This enrichment is statistically significant, considering the low probability of finding an equal number or more of CNGGN stem-loop structures in a random sample of predicted folds than the number in the microarray set ($P = 0.011344$, as determined by a hypergeometric distribution analysis comparing the ratio of CNGGN stem-loops occurring in predicted folds in the upregulated gene set

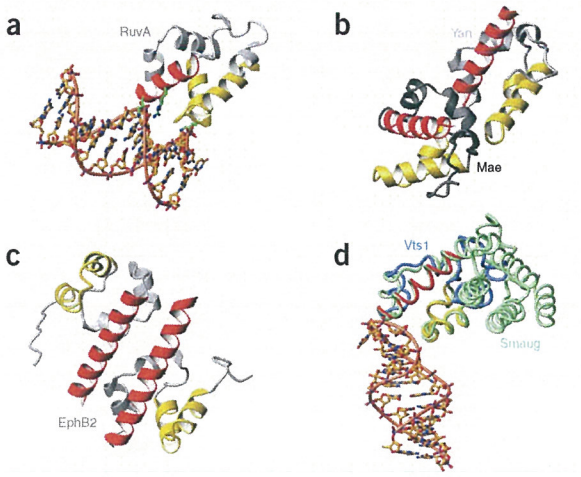


Figure 5 Comparison of the Vts1p-SAM–SRE complex with the structures of other SAM domain–containing proteins. (**a**) Crystal structure of the RuvA SAM-like domain in complex DNA[45] (PDB entry 1C7Y); red and yellow helices correspond to those in Vts1p-SAM shown in **d**. (**b**) Crystal structure of the heterodimer formed by the Yan and Mae SAM domains[17] (1SVO), colored light and dark gray, respectively. (**c**) Crystal structure of the homodimer formed by the EphB2 SAM domain[18] (1B4F). (**d**) Overlay of the Vts1p-SAM–SRE complex (blue, red and yellow) and the crystal structure of Smaug (light green; 1OXJ) solved in its free state[7].

relative to the genomic ORFs). Thus, the CNGGN-type stem-loop structures found in genes upregulated in the absence of Vts1p are likely to represent *bona fide* SRE structures.

To validate the microarray data, we first analyzed by northern blot the expression of the *TPO2*, *GPH1*, *HXK1* and *PGM2* mRNAs, which have predicted CNGGN-type stem-loops (the *NDE2* and *YMR196W* mRNAs could not be analyzed owing to low signals). This analysis confirmed that these four mRNAs are expressed at higher levels in the *vts1*Δ strain than in the wild-type strain (**Fig. 4b**). Next, we took advantage of the fact that expression of the *TPO2* mRNA is transiently induced for a period of 15 min after a shift from stationary phase to log phase (**Fig. 4c**), enabling us to follow the kinetics of its decay in wild-type and *vts1*Δ strains. The induction is followed by a rapid decrease in the expression of this mRNA, which drops to background levels 60 min after the shift in the wild-type strain (**Fig. 4c**). The timing of induction of *TPO2* was similar in the *vts1*Δ strain, but the kinetics of the decay were delayed compared to the wild-type strain: in contrast to wild-type cells, where background levels of *TPO2* were observed at 60 min, high levels of *TPO2* transcripts were still observed in the *vts1*Δ strain 60 min after the shift (**Fig. 4c**). These results show that although the same induction is observed upon return from stationary phase in wild-type and *vts1*Δ strains, the clearance of this mRNA after its induction is slower in the *vts1*Δ strain. This result suggests that the Vts1p protein is involved in the degradation of the *TPO2* mRNA.

DISCUSSION

So far, our structural knowledge of SAM domains has been limited to SAM-DNA and SAM-SAM interactions. RuvA contains a domain comprising four α-helices that resembles a SAM domain. In the structure of RuvA in complex with a DNA Holliday junction, this SAM-like domain interacts with the DNA phosphate backbone (**Fig. 5a**). The SAM-RNA and SAM-DNA recognition modes resemble each other in both being shape specific and using the same surface of the SAM domain. But the two recognition modes widely differ, as RuvA binds the minor groove of a perfectly double-stranded B-DNA, whereas Vts1p-SAM binds the major groove of an RNA pentaloop. To mediate SAM-SAM interactions, either the same protein surface is used as in the structure of the Yan-Mae heterodimer[17] (**Fig. 5b**) or a very different surface of the SAM domain is used to mediate the interaction, as in the structure of the EphB2 homodimer[18] (**Fig. 5c**).

Proteins containing a SAM domain that is homologous to those of Vts1p and Smaug have been found from *Candida albicans* to *Homo sapiens*. In these homologous proteins, the residues contributing to RNA binding are remarkably conserved from Trp459 to Tyr468 (Vst1p numbering) and in the N terminus of helix 6. In comparison, SAM domains that do not bind RNA have very different amino acid sequences in these two regions. For example, Gly498 and Ala499 are mostly replaced by bulkier residues in the non–RNA-binding SAM domains. Bulkier residues may stabilize the fold of the domain but would prevent RNA binding. By superimposing the structure of the SAM domain of Smaug with Vts1p-SAM in complex with RNA, we could see that most protein-RNA contacts with Vts1p could be reconstituted in this model (**Fig. 5d**). This comparison suggests that the RNA-binding SAM domain subfamily, its RNA targets and its RNA recognition mode have been conserved throughout evolution.

The shape recognition mode observed here in the Vts1p-SAM–SRE complex is similar to the ones mediated by double-stranded RNA–binding motifs (dsRBMs) or zinc fingers (reviewed recently[19]). The canonical dsRBMs bind any dsRNA, regardless of base composition, whereas certain dsRBMs like the dsRBM of Rnt1p (an RNase III homolog from budding yeast) specifically recognizes the shape of the conserved AGAA tetraloop[20]. Similarly, CCHH-type zinc fingers 4, 5 and 6 of transcription factor IIIA recognize the specific shape of part of the 5S RNA embedding loops A and E[21]. As for the SRE here, in these protein–RNA complexes the RNA target is conformationally preorganized in its free form and no or little conformational change occurs upon protein binding. This mode of RNA recognition contrasts with the sequence-specific mode of RNA recognition, which often requires a change in the conformation of the RNA upon binding, that is, an induced fit. This mode of RNA recognition is more often found[22], for example with the most common type of RNA recognition motif, the RNP domain[23].

Most genes found to be upregulated in the *vts1*Δ strain can be classified in two groups. One group of genes is associated with stress response and the other group with glucose metabolism and storage. Some of the genes of the former group are frequently found to be upregulated in microarray experiments, so they may not correspond to direct Vts1p targets. Notably, many genes from both groups are found to be upregulated during the diauxic shift[24], the switch from anaerobic growth to aerobic respiration upon depletion of glucose. This may imply that Vts1p is involved in regulating the diauxic shift at the post-transcriptional level. Our data also suggest that Vts1p is involved in the clearance of some mRNAs after a shift from stationary phase to log phase. A possible mechanism of action for Vts1p is that Vts1p would recruit the CCR4–POP2–NOT complex to SRE-containing transcripts to downregulate their expression, similarly to its *D. melanogaster* homolog Smaug[6,12]. This model would partly explain the role of the CCR4–POP2–NOT deadenylation complex in regulating the diauxic shift[25,26].

METHODS

Protein, RNA and complex preparation. The SAM domain of *S. cerevisiae* Vts1p DNA was amplified by PCR using the primer set 5′-GGAATTCCA TATGCCAAAGAGTTTGACAGACCCCAAGC-3′ and 5′-CCGCTCGAGTTAA

TAAGCAGATCTATCAATTAAATCACG-3′ and cloned into pET28a expression vector (Novagen) with an N-terminal His tag. The protein was overexpressed in *Escherichia coli* host cell BL21(DE3)+RIL (Stratagene). Cell cultures were induced at OD_{600} = 0.6 by 1 mM isopropyl β-D-thiogalactoside at 30 °C for 3 h in LB medium in the case of unlabeled protein or M9 medium containing $^{15}NH_4Cl$ as a nitrogen source and/or $^{13}C_6$-D-glucose as carbon source for labeled protein samples. For purification, Ni-NTA metal-affinity chromatography was used. SUPERase-In RNase inhibitor (Ambion) was added to all washing and elution buffers to eliminate RNase activity. After elution, the protein was kept at 4 °C to reduce precipitation, dialyzed against NMR-Buffer (50 mM NaCl, 20 mM NaH_2PO_4 (pH 6.5)) and concentrated to 1–2 mM. Identity and purity were verified by SDS-PAGE and MALDI mass spectrometry.

For all the studies, a 23-nt RNA stem-loop (5′-GGAGAGGCUCUGGCAG CUUUUCC-3′) with five different labeling schemes was used. The unlabeled RNA samples and the two ^{13}C,^{15}N-labeled RNA samples (with only G and C or with only A and U labeled) were produced by *in vitro* run-off transcription with T7 polymerase and purified by anion-exchange high-pressure liquid chromatography under denaturing conditions. Two RNA samples with ^{13}C labeling only in the sugars were chemically synthesized. In one sample, C8, C10, G12, C14 and G16 were labeled, and in the other sample, U9, U11, G13 and A15 were labeled. This labeling scheme was essential to obtain a precise structure of the RNA structure in both its free and bound forms. The RNA oligonucleotides were annealed at low salt and pH 6.5 by heating to 95 °C and snap-cooling on ice to favor a stem-loop conformation. Initially, the complex was prepared by titrating the RNA into the protein in three steps until a 1:1 stoichiometric ratio was reached. Later in the project, protein and RNA were mixed directly to form a 1:1 complex. Concentrations were determined by UV spectroscopy ($\varepsilon_{protein,280}$ = 16,500 M^{-1} cm^{-1}; $\varepsilon_{RNA,260}$ = 217,300 M^{-1} cm^{-1}). The RNAs used in the mutation studies were prepared by an identical procedure.

NMR spectroscopy. NMR spectra were acquired at 293 K for the free protein and at 303 K for the free RNA and the complex. Spectra we recorded on DRX-500, DRX-600 and Avance 900 Bruker spectrometers. All spectra were processed with XWINNMR (Bruker) and analyzed with Sparky 3.0 (T.G. Goddard and D.G. Kellner, University of California, San Francisco). The ^1H,^{13}C and ^{15}N chemical shifts of the protein, free and in complex, were assigned by standard methods[27]. All distance restraints were derived from 3D ^{15}N,^{13}C-edited NOESYs and 2D ^1H-^1H NOESY (τ_m = 150 ms) collected at 900 MHz. Slowly exchanging amide protons were identified from ^1H-^{15}N HSQC experiments after dissolving the lyophilized protein or the complex in ^2H_2O. CH and NH RDCs were measured from $^1J_{CH}$ differences in t_2-coupled HSQCs and $^1J_{NH}$ differences in 2D IPAP ^{15}N-^1H HSQCs, respectively, in the presence or absence of 18 mg ml^{-1} Pf1 filamentous phage[28]. Exchangeable proton resonances in the RNA were assigned using a 2D ^1H-^1H NOESY spectrum (τ_m = 200 ms) in 90% ^1H_2O/10% ^2H_2O (v/v) at 278 K. Nonexchangeable proton resonances were assigned using 2D ^1H-^1H NOESY, 2D ^1H-^1H TOCSY, ^1H-^{13}C HSQC, 3D ^{13}C-edited NOESY, 2D ^1H-^1H double-half-filtered NOESY (τ_m = 150 ms)[29] and a 3D ^{13}C F_1-edited, F_3-filtered NOESY-HSQC spectrum (τ_m = 150 ms)[30] in 99.99% ^2H_2O (v/v). The NOEs were semiquantitatively classified on the basis of their intensities in the 2D and 3D NOESY spectra. Hydrogen bonding distance restraints were used for base pairs when the imino proton was observed experimentally. The assignments of intermolecular NOEs were based on a 3D ^{13}C F_1-edited, F_3-filtered NOESY-HSQC spectrum (τ_m = 150 ms) and 2D ^1H-^1H F_1-^{13}C-filtered F_2-^{13}C-edited NOESY (τ_m = 150 ms) on the protein–RNA complexes, with either the protein ^{13}C,^{15}N labeled and the RNA unlabeled or the protein unlabeled and the RNA labeled.

Structure calculation and refinement. The CANDID/DYANA package[31,32] was used to calculate preliminary structures and a list of automatically assigned NOE distance constraints for the protein, free and in complex, started from random structures. The accuracy of the list of automatically generated distance constraints was manually checked. Preliminary structures of the free RNA and the protein–RNA complex were obtained by a simulated annealing protocol in

Table 1 NMR and refinement statistics for the unbound Vts1p-SAM and SRE and their complex

			Vts1 SAM–SRE complex	
	Vts1 SAM	SRE	Vts1 SAM	SRE
NMR distance and dihedral constraints				
Distance restraints				
Total NOE	2,366	455	1,680	426
Intra-residue	393	215	472	198
Inter-residue	1,973	240	1,208	228
Sequential ($\|i-j\|=1$)	607	187	398	182
Nonsequential ($\|i-j\|>1$)	1,352	28	779	21
Hydrogen bonds	14	25	31	25
Protein-RNA intermolecular			48	
Total dihedral angle restraints		91		91
RNA				
Sugar pucker		5		5
Backbone[a]		86		86
RDC restraints			69	42
Structure statistics[b]				
Violations (mean and s.d.)				
Number of distance restraint violations > 0.2 Å	4.63 ± 1.64	0.00 ± 0.00	0.65 ± 0.67	
Number of dihedral angle restraint violations > 5°	NA	0.05 ± 0.22	0.15 ± 0.37	
Max. dihedral angle restraint violation (°)	NA	0.97 ± 1.24	3.41 ± 2.93	
Max. distance constraint violation (Å)	0.27 ± 0.01	0.12 ± 0.01	0.23 ± 0.05	
Absolute RDC restraint violations			0.93 ± 0.74	
Deviations from idealized geometry[b]				
Bond lengths (Å)	0.011 ± 0.0001	0.012 ± 0.00008	0.011 ± 0.0004	
Bond angles (°)	2.682 ± 0.04	2.557 ± 0.024	2.298 ± 0.029	
Average pairwise r.m.s. deviation (Å)[b]				
Protein (447–523)				
Heavy	0.83 ± 0.11		1.00 ± 0.10	
Backbone	0.32 ± 0.07		0.36 ± 0.07	
RNA (1–13, 15–23)				
All RNA heavy		1.00 ± 0.24	1.00 ± 0.32	
Pentaloop (Δ14)		1.01 ± 0.29	0.45 ± 0.16	
Complex				
All complex heavy			1.48 ± 0.36	

NA, not applicable. [a]Based on A-form geometry derived from high-resolution crystal structures: α (270°–330°), β (150°–210°), γ (30°–90°), δ (50°–110°), ε (180°–240°) and ζ (260°–320°). These restraints were used only for the double-helical region. [b]Calculated for an ensemble of the 20 lowest-energy structures.

DYANA using manually assigned NOE distance constraints. To impose better convergence of the ensemble, some artificial torsion angles for RNA were used (**Table 1**, footnote a); these torsion restraints cover the A-form RNA double-helical range derived from the high-resolution crystal structures (Nucleic Acid Database accession codes AR0010 (R = 1.16 Å), AR0013 (R = 1.2 Å), AR0020 (R = 1.20 Å), AR0021 (R = 1.20 Å), ARB002 (R = 0.8 Å), ARB003 (R = 1.1 Å), ARB004 (R = 0.9 Å), ARB005 (R = 0.8 Å), ARF0108 (R = 1.8 Å) and ARH064 (R = 1.8 Å)).

All the conformers that served as starting structures for refinement in AMBER 7.0 (ref. 33) were generated in DYANA[32] using 200 random starting conformers. After 30,000 steps of simulated annealing in DYANA, the 40 best structures were selected on the basis of the lowest target function. In all AMBER calculations, the Cornell *et al.*[34] force field was used along with the generalized Born model[35] to mimic solvent. Both the free protein and the free RNA were refined using a protocol previously described[36]. A 10-ps simulated annealing protocol was used for the complex. The complex was refined using NOE-derived distances, torsion angles and RDCs as summarized in **Table 1**. Square-well penalty functions were used for all NMR restraints with the force constants 50 kcal mol^{-1} Å^{-2}, 200 kcal mol^{-1} rad^{-2} and 0.3 kcal mol^{-1} Hz^{-2} for NOEs, torsion angles and RDCs, respectively. A 0.2-fs time step was used for integration of Newton's motion equations to prevent atom overlap that could be enforced by RDC restraints. A 15-Å nonbonded cutoff was used. The relative weights of the valence-angle energy, torsion energy and 'improper' torsional terms were gradually increased during the simulated annealing to maintain the planarity of aromatic rings and proper local geometries. To prevent high violations of local geometries, additional angle restraints were used around the bonds where RDCs were accommodated. RDC restraints were collected for three different samples. Therefore, three sets of RDCs were used and each set was associated with a different alignment tensor. Initial estimates of alignment tensors were calculated as described previously[37]. The simulated annealing protocol was followed by a short energy minimization of 500 cycles (a combination of steepest-descents minimization followed by conjugate gradient technique). The 20 conformers with the lowest AMBER energy (free protein and free RNA) or with the lowest alignment-tensor energy (complex) were selected to form the final ensemble of structures (**Table 1**). Structural quality was analyzed using PROCHECK[38] (in the Ramachandran plot analyses of the Vts1-SAM domain free and in complex, respectively, 72.9% and 82.8% of all residues are in the most favored regions, 26.3% and 16.6% in the allowed regions and 0.8% and 0.6% in the disallowed regions). Figures were prepared with MOLMOL[39] and PyMOL[40].

Microarray analysis and RNA target screen. Isogenic wild-type and *vts1::KanMX* strains were purchased from Open Biosystems. Cells were grown in minimal medium, and RNA preparation and microarray analysis were performed as described in ref. 41 using Affymetrix microarrays. Three biological *vts1*Δ samples and two biological wild-type samples were analyzed from strains grown in synthetic dextrose medium. Northern analysis was performed as described in ref. 41 using PCR-generated random-primed probes or riboprobes. For the shift from stationary phase to log phase, wild-type and *vts1*Δ strains were grown to OD_{600} = 10–12 in synthetic dextrose medium and then shifted back to OD_{600} = 0.2 in YPD medium. Cells were collected at various time points and the corresponding RNAs were analyzed by northern blot for *TPO2* mRNA expression.

Bioinformatic analyses. *S. cerevisiae* genomic sequences for chromosomes I–XVI were culled from the *Saccharomyces* Genome Database (http://www.yeastgenome.org, PLN 14-AUG-2004). The GenBank accession numbers for the *S. cerevisiae* sequences reported in this paper are (in chromosome numerical order) NC_001133 through NC_001148. Draft sequence (seven-fold coverage, released 31 May, 2003) genomic sequences for three *Saccharomyces* species, *S. paradoxus*, *S. mikatae* and *S. bayanus* (phylogenetic distance ~20 million years) were culled from the Broad Institute's yeast comparative genomics project[42]. ORF comparison from *S. cerevisiae* yielded 4,492 annotated homologous genes in *S. bayanus*, 4,788 homologs in *S. paradoxus* and 4,525 homologs in *S. mikatae*. ORFs and flanking intergenic sequences from the top hits of the microarray results were searched for sequences that maintained an invariant G, allowing either a tetraloop or pentaloop structure and requiring canonical loop closure and an additional canonical Watson-Crick base pairing interaction between the –2 and +1 positions (numbered relative to the invariant G). Fifteen base sequences centered upon the invariant G were then subjected to theoretical folding of 100-mers using the Vienna RNA package[16]. Subfold nesting patterns in the longer sequence fold that maintained the tetra- or pentaloop motif adjoined by a minimum of four canonical base pairs (with one G•U base pair allowed) were inspected manually for adjacent features in the pattern suggested in refs. 7,43. Mapped ORFs of *S. cerevisiae* and the three other *Saccharomyces* species plus additional flanking intergenic sequences were then subjected to the same folding specifications. Location relative to the coding region and manually inspected regional structural features in putative folds were noted and incorporated into subsequent pattern matches.

The genomic ORF sequences representing well-annotated *S. cerevisiae* genes (culled from the *Saccharomyces* Genome Database; released 22 October, 2005) were trimmed to exclude dubious or overlapping ORF predictions and subjected to pattern-searching parameters and computational folding. A candidate list of regions with the potential to fold (determined through pattern searching) and the propensity to fold (determined using the Vienna RNA package[16]) were assembled and normalized by the total cumulative length of the sequences considered. In a comparison of remaining genomic elements, putative structures that maintained a CNGGN pentaloop were tabulated as a subset of all theoretical folds and their frequencies of occurrence were compared to the upregulated gene set.

Accession codes. Protein Data Bank: Coordinates have been deposited with accession codes 2ES6, 2ES5 and 2ESE for the free Vts1p, the free SRE-RNA and their complex, respectively. Gene Expression Omnibus: Microarray data have been deposited with accession code GSE3859.

Note: Supplementary information is available on the Nature Structural & Molecular Biology website.

ACKNOWLEDGMENTS

We are grateful to Y. Barral, Institute of Biochemistry, Swiss Federal Institute of Technology in Zurich, for providing the yeast genomic DNA for cloning and to S. Pitsch, L. Reymond and P. Wenter, Ecole Polytechnique Fédérale de Lausanne, for chemical synthesis of the ^{13}C sugar-labeled RNAs. We thank the University of California at Los Angeles Microarray Core Facility for help with microarray experiments. This investigation was supported by a predoctoral fellowship from the Roche Research Foundation for Biology to F.C.O., the American Heart Association Western States Affiliate (A.L.), a Human Frontier Science Program postdoctoral fellowship to R.S., a US National Science Foundation IGERT DGE grant to M.J., US National Institutes of Health grant GM 61518 to G.C. and grants from the Swiss National Science Foundation, the Structural Biology National Center of Competence in Research and the Roche Research Fund for Biology at the Swiss Federal Institute of Technology in Zurich to F.H.-T.A. F.H.-T.A. is a European Molecular Biology Organization Young Investigator.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/nsmb/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Qiao, F. & Bowie, J.U. The many faces of SAM. *Sci. STKE [online]* **2005**, re7 (2005) (doi:10.1126/stke.2862005re7).
2. Hall, T.M. SAM breaks its stereotype. *Nat. Struct. Biol.* **10**, 677–679 (2003).
3. Sharrocks, A.D. The ETS-domain transcription factor family. *Nat. Rev. Mol. Cell Biol.* **2**, 827–837 (2001).
4. Jousset, C. *et al.* A domain of TEL conserved in a subset of ETS proteins defines a specific oligomerization interface essential to the mitogenic properties of the TEL-PDGFR beta oncoprotein. *EMBO J.* **16**, 69–82 (1997).
5. Yu, X., West, S.C. & Egelman, E.H. Structure and subunit composition of the RuvAB-Holliday junction complex. *J. Mol. Biol.* **266**, 217–222 (1997).
6. Aviv, T. *et al.* The RNA-binding SAM domain of Smaug defines a new family of post-transcriptional regulators. *Nat. Struct. Biol.* **10**, 614–621 (2003).
7. Green, J.B., Gardner, C.D., Wharton, R.P. & Aggarwal, A.K. RNA recognition via the SAM domain of Smaug. *Mol. Cell* **11**, 1537–1548 (2003).
8. Dilcher, M., Kohler, B. & von Mollard, G.F. Genetic interactions with the yeast Q-SNARE VTI1 reveal novel functions for the R-SNARE YKT6. *J. Biol. Chem.* **276**, 34537–34544 (2001).

9. Smibert, C.A., Wilson, J.E., Kerr, K. & Macdonald, P.M. smaug protein represses translation of unlocalized nanos mRNA in the *Drosophila* embryo. *Genes Dev.* **10**, 2600–2609 (1996).
10. Dahanukar, A., Walker, J.A. & Wharton, R.P. Smaug, a novel RNA-binding protein that operates a translational switch in *Drosophila*. *Mol. Cell* **4**, 209–218 (1999).
11. Nelson, M.R., Leidal, A.M. & Smibert, C.A. *Drosophila* Cup is an eIF4E-binding protein that functions in Smaug-mediated translational repression. *EMBO J.* **23**, 150–159 (2004).
12. Semotok, J.L. *et al.* Smaug recruits the CCR4/POP2/NOT deadenylase complex to trigger maternal transcript localization in the early *Drosophila* embryo. *Curr. Biol.* **15**, 284–294 (2005).
13. Tucker, M., Staples, R.R., Valencia-Sanchez, M.A., Muhlrad, D. & Parker, R. Ccr4p is the catalytic subunit of a Ccr4p/Pop2p/Notp mRNA deadenylase complex in *Saccharomyces cerevisiae*. *EMBO J.* **21**, 1427–1436 (2002).
14. Chen, J., Chiang, Y.C. & Denis, C.L. CCR4, a 3′-5′ poly(A) RNA and ssDNA exonuclease, is the catalytic component of the cytoplasmic deadenylase. *EMBO J.* **21**, 1414–1426 (2002).
15. Moore, P.B. Structural motifs in RNA. *Annu. Rev. Biochem.* **68**, 287–300 (1999).
16. Hofacker, I.L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31**, 3429–3431 (2003).
17. Qiao, F. *et al.* Derepression by depolymerization; structural insights into the regulation of Yan by Mae. *Cell* **118**, 163–173 (2004).
18. Thanos, C.D., Goodwill, K.E. & Bowie, J.U. Oligomeric structure of the human EphB2 receptor SAM domain. *Science* **283**, 833–836 (1999).
19. Stefl, R., Skrisovska, L. & Allain, F.H. RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. *EMBO Rep.* **6**, 33–38 (2005).
20. Wu, H., Henras, A., Chanfreau, G. & Feigon, J. Structural basis for recognition of the AGNN tetraloop RNA fold by the double-stranded RNA-binding domain of Rnt1p RNase III. *Proc. Natl. Acad. Sci. USA* **101**, 8307–8312 (2004).
21. Lu, D., Searles, M.A. & Klug, A. Crystal structure of a zinc-finger-RNA complex reveals two modes of molecular recognition. *Nature* **426**, 96–100 (2003).
22. Williamson, J.R. Induced fit in RNA-protein recognition. *Nat. Struct. Biol.* **7**, 834–837 (2000).
23. Maris, C., Dominguez, C. & Allain, F.H. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J.* **272**, 2118–2131 (2005).
24. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
25. Collart, M.A. & Timmers, H.T. The eukaryotic Ccr4-not complex: a regulatory platform integrating mRNA metabolism with cellular signaling pathways? *Prog. Nucleic Acid Res. Mol. Biol.* **77**, 289–322 (2004).
26. Denis, C.L. & Chen, J. The CCR4-NOT complex plays diverse roles in mRNA metabolism. *Prog. Nucleic Acid Res. Mol. Biol.* **73**, 221–250 (2003).
27. Sattler, M., Schleucher, J. & Griesinger, C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Magn. Reson. Spectrosc.* **34**, 93–158 (1999).
28. Bax, A., Kontaxis, G. & Tjandra, N. Dipolar couplings in macromolecular structure determination. *Methods Enzymol.* **339**, 127–174 (2001).
29. Peterson, R.D., Theimer, C.A., Wu, H. & Feigon, J. New applications of 2D filtered/edited NOESY for assignment and structure elucidation of RNA and RNA-protein complexes. *J. Biomol. NMR* **28**, 59–67 (2004).
30. Zwahlen, C. *et al.* Method for measurement of intermolecular NOEs by multinuclear NMR spectrocopy:application to a bacteriophage lambda N-peptide/boxB RNA complex. *J. Am. Chem. Soc.* **119**, 6711–6721 (1997).
31. Herrmann, T., Guntert, P. & Wuthrich, K. 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 (2002).
32. Guntert, P., Mumenthaler, C. & Wuthrich, K. Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273**, 283–298 (1997).
33. Case, D.A. *et al. AMBER Version 7* (University of California, San Francisco, USA, 2002).
34. Cornell, W.D. *et al.* A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).
35. Bashford, D. & Case, D. Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **51**, 129–152 (2000).
36. Padrta, P., Stefl, R., Kralik, L., Zidek, L. & Sklenar, V. Refinement of d(GCGAAGC) hairpin structure using one- and two-bond residual dipolar couplings. *J. Biomol. NMR* **24**, 1–14 (2002).
37. Tsui, V., Zhu, L., Huang, T.H., Wright, P.E. & Case, D.A. Assessment of zinc finger orientations by residual dipolar coupling constants. *J. Biomol. NMR* **16**, 9–21 (2000).
38. Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R. & Thornton, J.M. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8**, 477–486 (1996).
39. Koradi, R., Billeter, M. & Wuthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 29–32, 51–55 (1996).
40. DeLano, W.L. *The PyMOL Molecular Graphics System* (DeLano Scientific, San Carlos, California, USA, 2002).
41. Lee, A., Henras, A.K. & Chanfreau, G. Multiple RNA surveillance pathways limit aberrant expression of iron uptake mRNAs and prevent iron toxicity in *S. cerevisiae*. *Mol. Cell* **19**, 39–51 (2005).
42. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
43. Duchow, H.K., Brechbiel, J.L., Chatterjee, S. & Gavis, E.R. The nanos translational control element represses translation in somatic cells by a Bearded box-like motif. *Dev. Biol.* **282**, 207–217 (2005).
44. Mulder, F.A., Schipper, D., Bott, R. & Boelens, R. Altered flexibility in the substrate-binding site of related native and engineered high-alkaline *Bacillus subtilis*ins. *J. Mol. Biol.* **292**, 111–123 (1999).
45. Ariyoshi, M., Nishino, T., Iwasaki, H., Shinagawa, H. & Morikawa, K. Crystal structure of the holliday junction DNA in complex with a single RuvA tetramer. *Proc. Natl. Acad. Sci. USA* **97**, 8257–8262 (2000).

Structure *14*, 345–355, February 2006 DOI 10.1016/j.str.2005.11.013

Structure and Specific RNA Binding of ADAR2 Double-Stranded RNA Binding Motifs

Richard Stefl,[1,3] Ming Xu,[2,3] Lenka Skrisovska,[1] Ronald B. Emeson,[2] and Frédéric H.-T. Allain[1,*]
[1]Institute of Molecular Biology and Biophysics
ETH Zürich
8093 Zürich
Switzerland
[2]Department of Pharmacology
Vanderbilt University
Nashville, Tennessee 37232

Summary

Adenosine deaminases that act on RNA (ADARs) site-selectively modify adenosines to inosines within RNA transcripts, thereby recoding genomic information. How ADARs select specific adenosine moieties for deamination is poorly understood. Here, we report NMR structures of the two double-stranded RNA binding motifs (dsRBMs) of rat ADAR2 and an NMR chemical shift perturbation study of the interaction of the two dsRBMs with a 71 nucleotide RNA encoding the R/G site of the GluR-B. We have identified the protein and the RNA surfaces involved in complex formation, allowing us to present an NMR-based model of the complex. We have found that dsRBM1 recognizes a conserved pentaloop, whereas dsRBM2 recognizes two bulged bases adjacent to the editing site, demonstrating RNA structure-dependent recognition by the ADAR2 dsRBMs. In vitro mutagenesis studies with both the protein and the RNA further support our structural findings.

Introduction

Adenosine deaminases that act on RNA (ADARs) convert adenosine to inosine (A-to-I) by hydrolytic deamination in cellular and viral RNA transcripts containing either perfect or imperfect regions of double-stranded RNA (dsRNA) (Bass, 2002; Emeson and Singh, 2000; Gerber and Keller, 2001; Keegan et al., 2001). To date, two functional enzymes (ADAR1 and ADAR2), and one inactive enzyme (ADAR3), have been characterized in mammals. A-to-I modification is nonspecific within perfect dsRNA substrates, deaminating up to 50% of the adenosine residues (Bass, 2002; Emeson and Singh, 2000). The nonspecific reaction occurs as long as the double-stranded architecture of the RNA substrate is maintained, since ADARs unwind dsRNA by changing A-U base pairs to I-U mismatches (Bass and Weintraub, 1988). The majority of nonselective editing occurs in untranslated regions (UTRs) and introns, where large regular duplexes are formed (Levanon et al., 2004; Morse et al., 2002; Morse and Bass, 1999; Rueter et al., 1999). Such modifications can modulate gene silencing triggered by intramolecular structures in mRNA (Tonkin and Bass, 2003), nuclear retention of RNA transcripts (Zhang and Carmichael, 2001), or antiviral responses by extensive modification of viral transcripts (Wong et al., 1991).

A-to-I editing can also be highly specific within imperfect dsRNA regions containing bulges, loops, and mismatches, and it can modify a single or limited set of adenosine residues (Bass, 2002; Emeson and Singh, 2000). Selective editing within pre-mRNAs has been shown to affect the primary amino acid sequence of the resultant protein product to produce multiple protein isoforms from a single gene. For example, ADARs have been shown to produce functionally important isoforms of numerous proteins involved in synaptic neurotransmission, including ligand and voltage-gated ion channels and G protein-coupled receptors (Bhalla et al., 2004; Burns et al., 1997; Egebjerg and Heinemann, 1993; Hoopengardner et al., 2003; Kohler et al., 1993; Lomeli et al., 1994; Sommer et al., 1991). The pre-mRNA encoding the B subunit of the α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) subtype of glutamate receptor (GluR-B) has been studied extensively and is edited at multiple sites (Seeburg et al., 1998). One of these locations is the R/G site, where a genomically encoded AGA is modified to IGA, resulting in an arginine-to-glycine change (the ribosome interprets I as G due to its similar base-pairing properties). This change affects the biophysical properties of the ion channel allowing the edited isoform to recover faster from desensitization (Lomeli et al., 1994). The R/G site of the GluR-B pre-mRNA is often used as a model system for A-to-I editing studies, as it forms a small and well-conserved ~70 nucleotide (nt) stem-loop containing three mismatches (Aruscavage and Bass, 2000); this structure is referred to as the R/G stem-loop.

Like many RNA binding proteins, ADARs display a modular domain organization. ADARs contain from one to three tandem copies of double-stranded RNA binding motif (dsRBMs) in their N-terminal region and an adenosine deaminase domain, the structure of which has recently been determined, in its C-terminal portion (Macbeth et al., 2005). The dsRBMs of ADARs may play an important role in modulating the editing selectivity of ADARs (Carlson et al., 2003; Doyle and Jantsch, 2002; Stephens et al., 2004). The dsRBM is a 70–75 amino acid domain found in many eukaryotic proteins with diverse functions (Fierro-Monti and Mathews, 2000). The structures of several dsRBMs have been determined (Bycroft et al., 1995; Kharrat et al., 1995; Nanduri et al., 1998) and reveal a highly conserved $\alpha\beta\beta\beta\alpha$ protein topology in which the two α helices are packed along a face of a three-stranded antiparallel β sheet. Furthermore, structures of the dsRBMs from *Xenopus laevis* RNA binding protein A (Xlrbpa2) (Ryter and Schultz, 1998), *Drosophila* Staufen protein (Ramos et al., 2000), and *Aquifex aeolicus* RNase III (Blaszczyk et al., 2004), in complex with nonnatural synthetic dsRNA substrates, have been determined; the dsRBM of Rnt1p (an RNase III homolog from budding yeast) has been determined in complex with its natural RNA substrate (dsRNA capped by an AGAA tetraloop) (Wu et al., 2004). These

*Correspondence: allain@mol.biol.ethz.ch
[3]These authors contributed equally to this work.

structures revealed not only how dsRBMs can bind any dsRNA, regardless of base composition, but also how structure-specific recognition of RNA hairpins is achieved (reviewed in Stefl et al., 2005a).

While the enzymatic activity of ADARs and their biological role(s) have extensively been studied (Bass, 2002; Emeson and Singh, 2000), the determinants that control site-selective RNA modification are poorly understood. Here, we report the solution structure of the two dsRBMs of ADAR2 and their interactions with the conserved 71 nt RNA stem-loop containing the GluR-B R/G site (R/G stem-loop). We show that each dsRBM binds a different structural element of the R/G stem-loop, and that dsRBM1 and dsRBM2 recognize a stem capped by a pentaloop and a stem containing two A-C mismatches, respectively. Our structural study demonstrates that the dsRBMs of ADAR2 have the ability to distinguish between specific structural features of RNA, suggesting their importance for editing site selectivity.

Results

ADAR2 dsRBMs Are Independent Domains

We investigated the N-terminal region of rat ADAR2 (74–301) that includes the dsRBM1 (74–147), the interdomain linker (148–230), and dsRBM2 (231–301) by using NMR spectroscopy (Figure 1A). This protein construct was aminoterminally fused with noncleavable solubility-enhancement tag GB1 (Zhou et al., 2001) to improve its expression and solubility (Stefl et al., 2005b). ^{1}H, ^{13}C, and ^{15}N resonance assignments of GB1-dsRBM12 and of the two isolated dsRBMs were obtained as previously described (Stefl et al., 2005b). As the spectral quality of the 32 kDa GB1-dsRBM12 suffered from an increased transverse relaxation, we used hydrogen-to-deuteron substitutions at various levels to improve the relaxation properties of the protein (Stefl et al., 2005b). The comparison between the [^{1}H,^{15}N]-TROSY spectrum of a deuterated GB1-dsRBM12 and the [^{1}H,^{15}N]-HSQC spectra of both isolated dsRBM1 and dsRBM2 (Figure S1; see the Supplemental Data available with this article online) shows that the dsRBM resonances are identical in both contexts, except for a few N- and C-terminal residues. In addition, this comparison indicates that the interdomain linker is flexible, as the chemical shifts of the linker residues have random coil values (Stefl et al., 2005b). To determine whether the chemical shift differences of the terminal residues of the isolated dsRBMs and GB1-dsRBM12 are due to involvement in interdomain contacts or due to different flanking residues at the termini, we carefully analyzed and compared the [^{1}H,^{13}C]-HSQC, ^{15}N- and ^{13}C-separated NOESY data of a 50%-deuterated GB1-dsRBM12 with data from the corresponding experiments of the isolated dsRBM1 and dsRBM2. However, no interdomain NOEs could be observed. Furthermore, when the isolated dsRBM1 and dsRBM2 were mixed in *trans*, the [^{1}H,^{15}N]-HSQC spectrum showed no change of chemical shifts compared to the two [^{1}H,^{15}N]-HSQC spectra of the isolated domains (data not shown). These results indicate that the ADAR2 dsRBMs are independent domains separated by a flexible linker, similar to the two dsRBMs of PKR (Nanduri et al., 1998). Thus, we used separate dsRBM1 and dsRBM2 constructs to determine their structures by NMR.

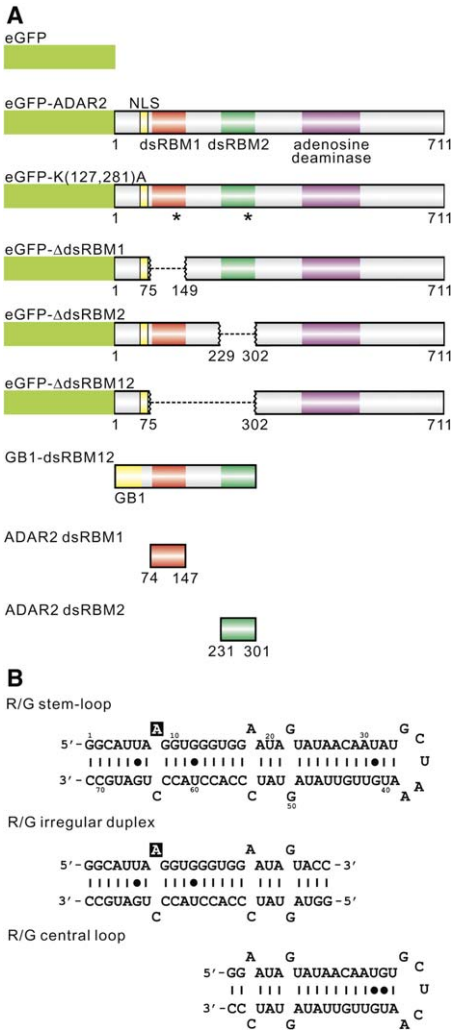


Figure 1. Protein and RNA Constructs Used for NMR and Functional Analysis

(A) A schematic diagram indicating the structure of eGFP, wild-type eGFP-ADAR2, eGFP-K(127,281)A double mutant (the position of point mutations are labeled by an asterisk), mutant fusion proteins (showing deletion of dsRBM1, dsRBM2, and dsRBM12), as well as GB1-dsRBM12, ADAR2 dsRBM1, and ADAR2 dsRBM2 is presented. The coordinates of each deletion are indicated, relative to the start codon. NLS, nuclear localization signal. GB1, immunoglobulin binding domain B1 of streptococcal protein G, is a noncleavable solubility-enhancement tag (Zhou et al., 2001).

(B) R/G stem-loop, a 71 nt stem-loop that includes the 67 nt of the human GluR-B mRNA stem-loop and is closed by two GC base pairs to improve the yield from in vitro transcription. R/G irregular duplex, a 52 nucleotide duplex embedding the R/G editing site of the rat R/G stem-loop, closed by two GC base pairs at both ends of the duplex. R/G central loop, a 41 nt stem-loop, represents the central part of the rat GluR-B R/G mRNA. Note that rat and human RNA sequences of the GluR-B mRNA R/G stem-loop are identical, except for two nucleotides in the central loop part, G32A and C37A, which do not affect the fold of this region (R.S. and F.H.-T.A., unpublished data). Furthermore, rat and human ADAR2 dsRBM1, a domain interacting with the central part of GluR-B R/G mRNA, are identical within the structured domain.

ADAR2 dsRBM Structures Are Not Identical

The structures of dsRBM1 and dsRBM2 were determined by using 1754 and 1459 conformationally restrictive NOE distance restraints, respectively, derived from

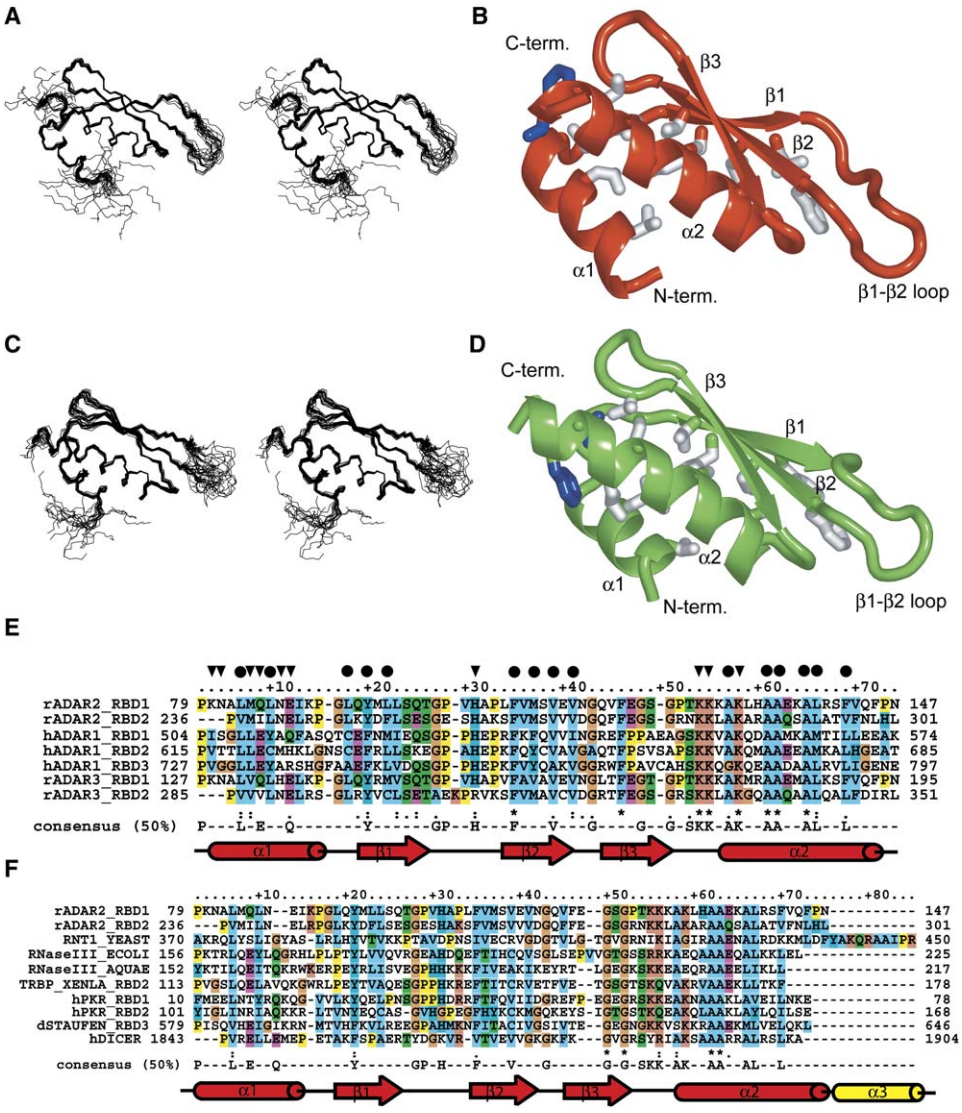


Figure 2. ADAR2 dsRBMs' Protein Structure

(A–D) NMR ensemble of (A) dsRBM1 and (C) dsRBM2; 20 lowest-energy structures. Ribbon representation of (B) dsRBM1 (in red) and (D) dsRBM2 (in green) with conserved hydrophobic core residues (sticks, in white).

(E) Sequence alignment of various dsRBMs of ADARs. Two dsRBMs of rat ADAR2, three dsRBMs of human ADAR1, and two dsRBMs of rat ADAR3. The consensus sequence (>50%) of dsRBMs (Fierro-Monti and Mathews, 2000) and a schematic of secondary structure elements of dsRBM are indicated below. The residues of the conserved hydrophobic core and the RNA binding surface are indicated by black dots and triangles, respectively.

(F) Sequence alignment of various dsRBMs. Two dsRBMs of rat ADAR2, dsRBM of Rnt1p (*S. cerevisiae*), dsRBM of RNase III (*E. coli*), dsRBM of RNase III (*A. aeolicus*), second dsRBM of Xlrbpa (*X. laevis*), dsRBM1 and dsRBM2 of PKR (*H. sapiens*), dsRBM3 of Staufen (*D. melanogaster*), and dsRBM of DICER (*H. sapiens*).

2D homonuclear and 3D ^{13}C- and ^{15}N-edited NOESYs. The resulting structures are represented by an ensemble of 20 conformers for both dsRBM1 and dsRBM2 (Figure 2). The ensembles have root mean square deviations of 0.45 ± 0.10 and 0.52 ± 0.12 Å over the backbone in the structured regions for dsRBM1 and dsRBM2, respectively. A full summary of structural statistics is given in Table 1. Both the ADAR2 dsRBM1 and dsRBM2 structures adopt the same fold as all other members of the dsRBM family, with an $\alpha\beta\beta\beta\alpha$ topology in which the two α helices are packed along a face of a three-stranded antiparallel β sheet (Figure 2). A central hydrophobic core composed of the residues shown in Figure 2 (Figures 2B and 2D for structural arrangement, and Figure 2E for sequence alignment) stabilizes the fold of the domain. Although the two dsRBMs of ADAR2 have 50% amino acid identity, the two structures differ slightly in the orientation of α helix 1 relative to the other secondary structure elements (Figure 2). This altered orientation is a result of a protein sequence difference in two amino acids at the C terminus of α helix 2, where Phe142 and Val143 in dsRBM1 are replaced by Val296 and Phe297 in dsRBM2. Phe297, compared to Val143, is bulkier, leading to a different interaction between the two α

Table 1. NMR Experimental Constraints and Structure Statistics

	dsRBM1	dsRBM2
Distance constraints		
NOE upper distance limits	1754	1459
Structure statistics[a]		
NOE violations		
Number (>0.3 Å)	0.85 ± 1.13	2.00 ± 1.34
Maximum violations (Å)	0.33 ± 0.13	0.39 ± 0.04
Rmsd from the mean coordinates (Å)[a]		
Only secondary structure elements (residues 6–25, 36–68 for dsRBM1, and 6–22, 33–65 for dsRBM2)		
Backbone	0.45 ± 0.10	0.52 ± 0.12
Heavy atoms	1.07 ± 0.16	1.29 ± 0.14
Entire domain		
Backbone	0.80 ± 0.25	1.34 ± 0.36
Heavy atoms	1.28 ± 0.23	2.13 ± 0.35
Rmsd from ideal geometry[a]		
Bonds lengths (Å)	0.0099 ± 0.0002	0.0106 ± 0.0002
Bond angles (°)	2.5127 ± 0.0309	2.5448 ± 0.0456
Ramachandran analysis (%)[b]		
Most favored region	86	82.4
Allowed region	13.1	16.2
Disallowed region	0.9	1.4

[a] The statistics (average ± SDs) calculated for the bundle of the 20 best-energy conformers.
[b] As determined by PROCHECK (Laskowski et al., 1996).

helices (Figure 2D, in blue). We found another difference between the two dsRBMs in the conformation of the β1-β2 loop. The β1-β2 loop of dsRBM1 is well defined, whereas the β1-β2 loop of dsRBM2 is conformationally heterogeneous (Figures 2A and 2C). In dsRBM2, several amide resonances of this loop are not observable in the spectra, probably due to a conformational exchange, whereas all of the amide proton resonances of the β1-β2 loop of dsRBM1 were observed and involved in many NOE correlations. These observations suggest that the β1-β2 loop of dsRBM1 is more rigid than the β1-β2 loop of dsRBM2, which is probably due to the presence of two prolines in dsRBM1 that are not found in dsRBM2 (Figure 2E). Flexible β1-β2 loops were also observed in other dsRBM structures (Leulliot et al., 2004; Ramos et al., 2000). Altogether, the longer α helix 1 and the conformationally preorganized β1-β2 loop of dsRBM1 might be important factors for ADAR2 RNA recognition.

Mapping of the RNA Binding Surface on the ADAR2 dsRBMs

To investigate how ADAR2 dsRBMs bind RNA, we performed an NMR chemical shift perturbation study with a 71 nt R/G stem-loop RNA (Figure 1B). This RNA is a 33 bp helix containing three mismatches (two A-C and one G-G) that is capped by a structured pentaloop (Stefl and Allain, 2005). A8 of this RNA can be specifically edited (up to 74%) by ADAR2 in vitro, but if the mismatches are replaced by Watson-Crick base pairs, the editing efficiency is reduced substantially (Kallman et al., 2003; Ohman et al., 2000).

First, we studied the interaction between GB1-dsRBM12 and the 71 nt R/G stem-loop (Figure 1). Upon RNA titration up to an equimolar ratio, the protein resonances showed significant chemical shift changes when followed by [^{1}H,^{15}N]-TROSY spectra; however, we could not assign the protein resonances of this 55 kDa GB1-dsRBM12-R/G stem-loop complex due to severe line broadening (Figure S2). As the dsRBMs are independent in the free form, we presumed that they could have different binding sites on the R/G stem-loop. Therefore, we used two truncations of the R/G stem-loop, a 52 nt R/G irregular duplex and a 41 nt R/G central loop (Figure 1). We prepared four complexes (the two truncated RNAs bound to each dsRBM) and measured a [^{1}H,^{15}N]-HSQC spectrum for each. The chemical shifts in the [^{1}H,^{15}N]-HSQC spectra of dsRBM1 bound to the R/G central loop and of dsRBM2 bound to the R/G irregular duplex complexes were virtually identical to the ones in the [^{1}H,^{15}N]-TROSY spectrum of the full-length complex (Figure S2). These observations indicate that the dsRBMs are bound in the same manner in these two subcomplexes and in the full-length complex (GB1-dsRBM12 bound to the R/G stem-loop). The two dsRBMs of ADAR2 bind two distinct locations on the R/G stem-loop; dsRBM1 binds close to the pentaloop, and dsRBM2 binds close to the editing site. The NMR data of the reciprocal complexes (dsRBM2 bound to the R/G central loop and dsRBM1 bound to the R/G irregular duplex complexes) indicated that such subcomplexes are formed; however, their chemical shift values do not resemble the ones observed in the full-length complex. Furthermore, the spectra of these two subcomplexes showed severe line broadening, probably resulting from exchange between multiple protein-RNA complexes of similar affinities. These observations indicated that the binding of both dsRBMs is specific.

To gain more detailed insights into the interactions between the ADAR2 dsRBMs and the R/G stem-loop, the backbone amide resonances of the dsRBM1 and dsRBM2 in the subcomplexes were assigned. In both dsRBMs, the largest chemical shift changes between the free and the bound forms were observed for the backbone amides of α helix 1 and the β1-β2 loop (Figure 3). In addition, large chemical shift changes were observed for the β3-α2 loop and the N terminus of α helix 2 of dsRBM1, whereas no large chemical shift changes were observed for the dsRBM2 in these two regions. These results are surprising since these regions of dsRBM1 and dsRBM2 are similar in sequence with the presence of three conserved lysines (Figure 2E). Taken together, the patterns of chemical shift perturbations indicate that the protein-RNA interactions are different between dsRBM1 and dsRBM2, reflecting the structural differences already observed in the free dsRBMs structures. The RNA binding surfaces identified by chemical shift perturbations in both dsRBMs agree well with the positive electrostatic potential calculated by using a nonlinear Poisson-Boltzmann equation (Figures 3C and 3F). In particular, the stretch of lysines located in the β3-α2 loop and the N terminus of α helix 2 found in both dsRBMs (Figure 2E) create the region with the highest potential. The RNA binding surfaces of ADAR2 dsRBM1 and dsRBM2, although not identical, are similar to the ones observed in other dsRBMs-RNA complexes (Blaszczyk et al., 2004; Ramos et al., 2000; Ryter and Schultz, 1998; Wu et al., 2004).

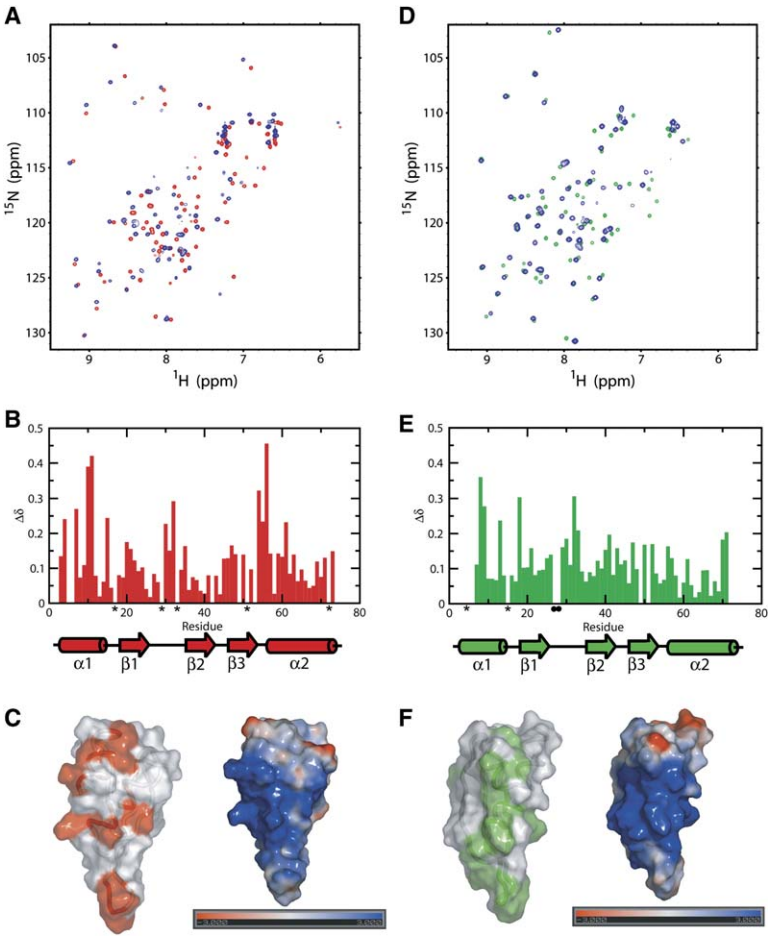


Figure 3. RNA Binding Surfaces of ADAR2 dsRBMs

(A–C) dsRBM1-R/G central loop. (A) Overlay of [^{1}H-^{15}N]-HSQC spectra of dsRBM1 (in red) and the dsRBM1-R/G central loop complex (in blue). (B) Chemical shift changes (Δδ) for dsRBM1 upon addition of the R/G central loop. The Δδ is calculated as $([\omega_{HN}\Delta\delta_{HN}]^2 + [\omega_N\Delta\delta_N]^2)^{1/2}$, where $\omega_{HN} = 1$ and $\omega_N = 0.154$ are weight factors of the nucleus (Mulder et al., 1999); asterisk, proline. (C) Δδ upon RNA binding mapped to the surface of the protein identifies the RNA binding surface of dsRBM1 (left; in red; $\Delta\delta \geq 0.15$ are displayed). The electrostatic surface of dsRBM1 (right).

(D–F) dsRBM2-R/G irregular duplex. (D) Overlay of [^{1}H-^{15}N]-HSQC spectra of dsRBM2 (in green) and the dsRBM2-R/G irregular duplex complex (in blue). (E) Δδ for dsRBM2 upon addition of R/G irregular duplex (asterisk, proline; closed, black circle, unassigned). (F) Δδ upon RNA binding mapped to the surface of the protein identifies the RNA binding surface of dsRBM2 (left; in green). The electrostatic surface of dsRBM2 (right).

Mapping of the Protein Binding Surface on the R/G Stem-Loop

To investigate the protein binding surface, the resonances of the 71 nt R/G stem-loop were assigned as described in the Supplemental Data. The NMR data showed the presence of a G22-G50 mismatch and two "open" A-C mismatches (A8-C64 and A18-C54). Based on these data, together with our NMR structure of the central pentaloop region of the human R/G stem-loop (Stefl and Allain, 2005), we built a structural model of the rat 71 nt R/G stem-loop (Figure 4D).

Upon protein binding to the 71 nt R/G stem-loop, the RNA imino proton resonances broaden continuously as a result of chemical exchange and increased molecular weight. However, no significant chemical shift changes and no new imino proton resonances were observed, indicating that no changes in the RNA secondary structure take place upon complex formation. In the two subcomplexes of dsRBM1 bound to the R/G central loop and dsRBM2 bound to the R/G irregular duplex, a precise analysis of the pyrimidine H5 and H6 chemical shift perturbations upon protein binding could be done with a series of 2D-[^{1}H,^{1}H]-TOCSY spectra (Figure 4). In the course of the protein titrations, the resonances moved from their initial positions, which correspond to the free form, in a stepwise directional manner until they reached their final positions, which correspond to the fully bound state (example shown in Figures 4A and 4B). These data indicate that, in both subcomplexes, the RNAs are in fast exchange between their free and bound forms relative to the NMR timescale. The binding of dsRBM1 to the R/G central loop induces a significant chemical shift perturbation of C37 and U40 (Figures 4A and 4B), and the binding of dsRBM2 to the R/G irregular duplex causes pyrimidine perturbations of C54, C55, C56, C63, C64, and U65; C54 and C64 experience the largest chemical shift changes (Figure 4C). These chemical shift changes strongly suggest that the above-mentioned RNA bases are interacting with the proteins or a significantly changed conformation (for C54 and C64) as they become stacked within the duplex upon protein binding. We find the latter explanation less likely, as the C54 and C64 H5 and H6 resonances are still in the chemical shift range of unpaired nucleotide and no new imino protons are observed upon protein binding. Figure 4D displays the pyrimidine residues with the largest chemical shift perturbations on the 3D NMR model of the R/G stem-loop.

The study of the reversed subcomplexes, dsRBM1 bound the R/G irregular duplex and dsRBM2 bound to the R/G central loop, showed that dsRBM1 contacts C54 and C64 on the R/G irregular duplex as well, whereas dsRBM2 only binds the stem (where one A-C mismatch, C64, is present), not the pentaloop of the central R/G stem-loop (data not shown).

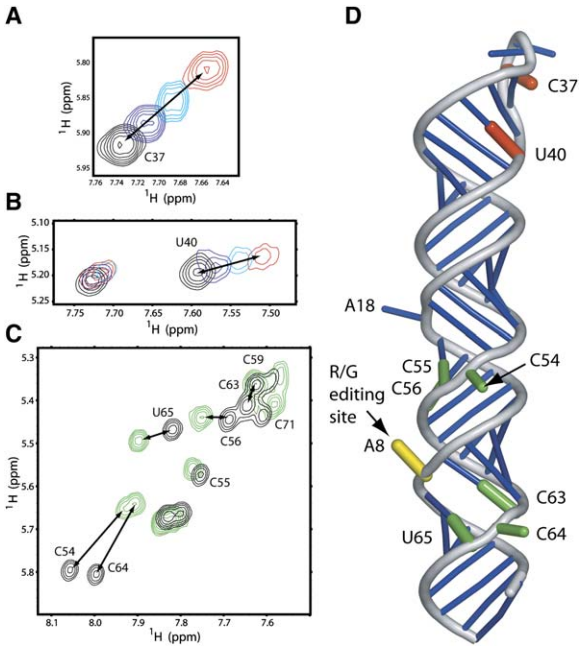


Figure 4. Protein Binding Surfaces of the R/G Stem-Loop

(A and B) dsRBM1-R/G central loop. Overlay of series 2D [^{1}H-^{1}H]-TOCSY spectra (focused on a portion with H5-H6 resonances of C37, U25, and U40) of the titration experiment in which the dsRBM1 was successively added in the following molar ratios: 0:1 (free RNA) (in black), 1/3:1 (in blue), 2/3:1 (in cyan), and 1:1 (fully bound RNA) (in red).

(C) dsRBM2-R/G irregular duplex. Overlay of 2D [^{1}H-^{1}H]-TOCSY spectra of free (in black) and fully bound RNA (in green), focused on a portion with strongly shifted H5-H6 resonances.

(D) Chemical shift changes upon protein binding mapped to an NMR model of the R/G stem-loop. The nucleotides in red and green are affected upon dsRBM1 and dsRBM2 binding, respectively.

ADAR2 dsRBMs Do Not Dimerize on the R/G Stem-Loop

It has been reported that ADAR activation involves RNA-dependent dimerization (Cho et al., 2003; Gallo et al., 2003; Jaikaran et al., 2002). To test whether more than two ADAR2 dsRBMs could bind the 71 nt R/G stem-loop, we performed light-scattering experiments. We incubated GB1-dsRBM12 with the R/G stem-loop at protein:RNA stochiometric ratios of 0.5:1, 1:1, 2:1, and 4:1, and we analyzed them by using gel filtration coupled with in-line laser light scattering. At the ratio of 1:1, a single peak with the expected size (MW ~55 kDa) of the bimolecular complex appeared. In all other ratios, peaks of either unbound RNA or protein, in addition to the peak of the bimolecular complex, appeared in the chromatogram. This indicates that only one molecule of GB1-dsRBM12 can be accommodated by the R/G stem-loop, and that the peptide sequence responsible for the dimerization is outside the RNA binding region of ADAR2 (74–301).

NMR Model of ADAR2 dsRBM12 in Complex with the R/G Stem-Loop

To understand the basic principles of this recognition, we constructed a model of ADAR2 dsRBMs in complex with the 71 nt R/G stem-loop based on our precise NMR identification of both the protein and RNA interaction surfaces and on the knowledge of the basic structural elements controlling dsRBMs-RNA recognition (Stefl et al., 2005a). We performed a docking search by using a methodology similar to the one implemented in HADDOCK (Dominguez et al., 2003). In docking calculations, we used the NMR ensembles of the dsRBM1 and dsRBM2 and multiple MD-generated conformations of the R/G stem-loop model as starting structures. We took advantage of the fact that there are no major changes in the backbone conformation of the dsRBMs upon RNA binding (Leulliot et al., 2004; Ramos et al., 2000; Wu et al., 2004), and that dsRBMs contact RNA via a well-conserved interaction scheme (Stefl et al., 2005a). Both facts provided the constraints that significantly reduced the degrees of freedom of the conformational docking search. Furthermore, the docking was guided by the information obtained from our NMR chemical shift perturbation studies that identified the protein and RNA interaction surfaces.

Figure 5A shows the resulting NMR model with the lowest energy of the ADAR2 dsRBM12-R/G stem-loop complex. In a similar manner to what was observed for the Rnt1p dsRBM-AGNN tetraloop-containing RNA complex (Wu et al., 2004), the dsRBM1 contacts the minor groove of the GCUCA pentaloop and the adjacent G-U mismatch of the central region of the R/G stem-loop. The dsRBM2 interacts with the bulged C54 and C64 opposite the editing site. Among the dsRBM-dsRNA complexes determined to date, the interaction of ADAR2 dsRBM2 is unique, since dsRBM2 appears to recognize two bulged cytosines. This is reminiscent of CCHH-type zinc fingers that are also able to recognize RNA bases that bulge out of the rigid structural architecture (Lu et al., 2003). dsRBMs are considered to be structure-specific rather than sequence-specific RNA binding proteins (Stefl et al., 2005a). Based on our NMR model of the complex, the dsRBMs of ADAR2 are not an exception, since dsRBM1 recognizes a stem-loop structure and dsRBM2 recognizes an RNA helix containing two A-C mismatches separated by ten base pairs.

Both ADAR2 dsRBMs Are Important for Efficient Editing of the R/G Site

To investigate whether both ADAR2 dsRBM-RNA interactions are important for ADAR2-mediated editing of the R/G site, either dsRBM1 or dsRBM2 was deleted from an eGFP-ADAR2 fusion protein (Figure 1A), which has been previously shown to have a comparable enzymatic activity to wild-type ADAR2 protein (Sansam et al., 2003). We took advantage of an in vitro editing system that used the R/G editing substrate and wild-type or mutant eGFP-ADAR2 proteins in HEK293 nuclear extracts. Preliminary time course analyses with wild-type eGFP-ADAR2 protein were used to define the linear range of the in vitro editing reaction (data not shown), and equivalent amounts of wild-type and mutant proteins, as determined by quantitative Western blotting, were incubated with an in vitro-transcribed R/G editing substrate (Dawson et al., 2004). Nuclear extracts from eGFP-transfected cells defined background editing levels for the in vitro system, while the wild-type eGFP-ADAR2 protein demonstrated robust editing of the R/G site (Figure 5E). Deletion of either dsRBM1 or dsRBM2 dramatically decreased the editing on R/G site by 3- to 10-fold, while

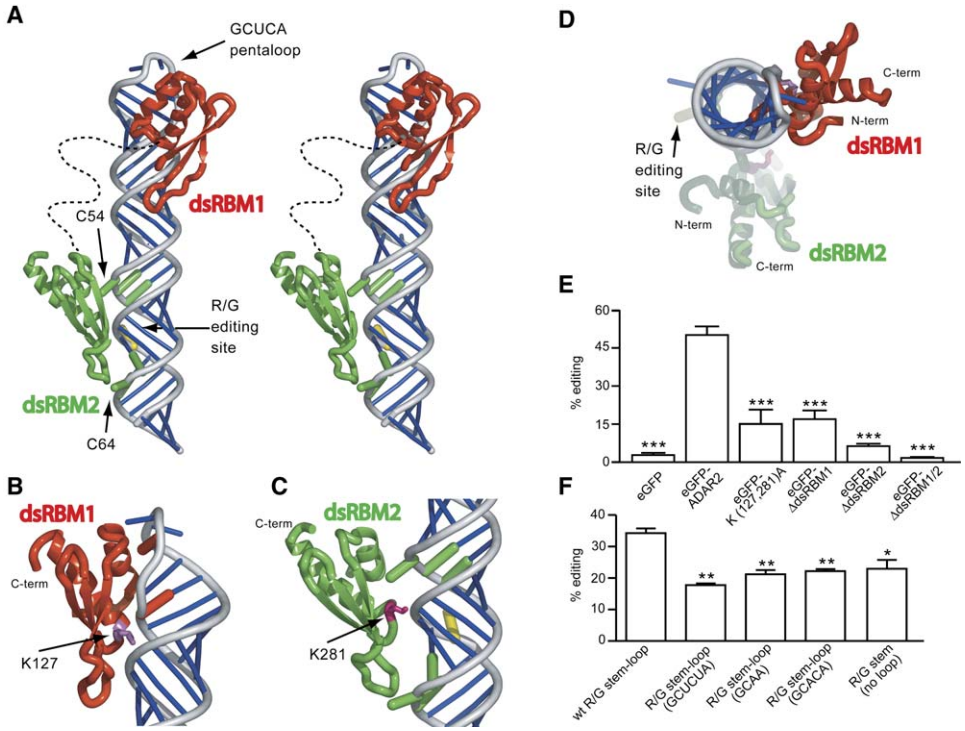


Figure 5. ADAR2-Mediated Editing of the R/G Site Requires Both dsRBMs

(A) Overall NMR model of ADAR2 dsRBM12 in complex with the R/G stem-loop (stereoview).

(B) dsRBM1 (in red) interacts with the central part of the R/G stem-loop; α helix 1 contacts the pentaloop and the adjacent G-U base pair.

(C) dsRBM2 interacts with bulged cytosines, opposite the editing site. K127 and K281, residues mutated to alanine in our functional studies are shown in magenta and pink, respectively.

(D) Overall NMR model (top view).

(E) Quantitative analysis of RNA editing for the R/G site from in vitro editing analyses with wild-type, eGFP-K(127,281)A, and mutant eGFP-ADAR2 fusion constructs lacking either dsRBM1 or dsRBM2 and both of them (mean ± SEM; n = 5); **p < 0.01; ***p < 0.001 compared to wild-type eGFP-ADAR2 (using the Student's t test).

(F) Various mutations introduced in the loop region of the R/G stem-loop and their role on editing efficiency at the R/G site. Mutations include: GCUCA pentaloop replaced by a GCUCUA hexaloop, by a GCAA tetraloop, and by a GCACA pentaloop. All of these mutations change the conformation of the loop. In addition, we used RNA in which the pentaloop of the R/G stem-loop is removed and in which the stem sequence is conserved ("R/G stem"). All mutants were assayed for editing activity at the R/G site in vitro by using wild-type eGFP-ADAR2 (mean ± SEM; n = 3); **p < 0.01; *p < 0.03 compared to wild-type eGFP-ADAR2 (using the Student's t test).

deletion of both dsRBMs (eGFP-ΔdsRBM12) completely eliminated A-to-I conversion at the R/G site (Figure 5E). In addition, simultaneous mutations of the two highly conserved K127 (dsRBM1) and K281 (dsRBM2) displayed significantly lower editing activity at the R/G site (Figure 5E), further confirming the importance of RNA binding of both domains for editing, as both side chains are predicted in our NMR model to interact with the sugar-phosphate backbone (Figures 5B and 5C).

Functional Importance of the R/G Stem-Loop Secondary Structure for Editing by ADAR2

Our NMR study shows that both dsRBM1 and dsRBM2 bind specific region of the R/G stem-loop; dsRBM1 binds near the pentaloop, and dsRBM2 binds the stem with two A-C mismatches in the neighborhood of the R/G editing site (Figure 5A). The functional importance of the A-C mismatches was previously shown, as their replacement by Watson-Crick base pairs decreases the editing from 74% to 41% (Ohman et al., 2000) and its selectivity for the R/G site from 80% to 30% (Kallman et al., 2003). To assess the functional importance of the pentaloop, we created several mutants in the loop region of the R/G stem-loop (Figure 5F), and we assayed them for editing activity at the R/G site in vitro. These mutants include a variation in the GCUCA pentaloop sequence (GCACA; a single mutation that changes the fold of the loop [Stefl and Allain, 2005]) and variations in the loop size (GCAA tetraloop and GCUCUA hexaloop). All of these mutants display lower editing efficiency at the R/G site compared to the wild-type (Figure 5F), indicating that the pentaloop sequence GCUCA and its specific structure are functional determinants of the editing at the R/G site. In addition, an R/G stem-loop mutant lacking the entire GCUCA pentaloop (R/G stem) also has lower editing activity at the R/G site (Figure 5F). Altogether, changes in the sequence or in the size of the pentaloop that lead to a different pentaloop topology result in lower editing efficiency at the R/G site, indicating the functional importance of the pentaloop structure (Stefl and Allain, 2005).

Discussion

Structure of ADAR2 dsRBMs: Comparison and Implications

In comparison to other dsRBMs, ADAR2 dsRBM1 and dsRBM2 differ from the canonical dsRBM fold like the

ones of Xlrbpa2 (Ryter and Schultz, 1998) and *Aquifex aeolicus* RNase III (Blaszczyk et al., 2004) (Figure S4A, in white). Interestingly, ADAR2 dsRBM1 resembles the dsRBM of Rnt1p (Figure S4B, in blue) (Leulliot et al., 2004); however, it lacks α helix 3, an additional element that imposes the conformation of the "recognition" α helix 1 in the dsRBM of Rnt1p. ADAR2 dsRBM2 appears to be unique among other members of the dsRBM family (Figure S4). This structural difference in the relative orientation of α helix 1 may be functionally important, as it is a key element that modulates the RNA binding specificity of dsRBMs (Ramos et al., 2000; Stefl et al., 2005a; Wu et al., 2004) (see below).

How Do ADAR2 dsRBMs Recognize the R/G Stem-Loop?

With dsRBM-containing proteins, questions regarding binding specificities have always been difficult to answer, as this abundant RNA binding domain is considered to bind any dsRNA in a non-sequence-specific manner. Structures of single dsRBMs in complex with dsRNA indeed revealed that dsRBMs are not sequence-specific RNA binders, but they raised the question of whether dsRBMs would rather recognize certain RNA structures, like stem-loops or irregular duplexes (Ramos et al., 2000; Stefl et al., 2005a; Wu et al., 2004). Our extensive binding study of ADAR2 dsRBMs with the GluR-B R/G stem-loop and our structural model further extends our understanding of how ADAR2 dsRBMs recognize their targets and, more generally, how dsRBMs recognize RNA.

dsRBMs are often present in multiple nonidentical copies in proteins. In studying the two dsRBMs of ADAR2, we provide one of the first structural studies on how two domains work together. Surprisingly, although both dsRBMs are essential for efficient RNA editing, they apparently bind the RNA independently, as the interdomain linker (147–231) that bridges the dsRBMs of ADAR2 is found to be unstructured in both the free and bound forms of the protein and does not appear to participate in the interaction with the R/G stem-loop. This contrasts with what was found for other RNA recognition motifs, in which the interdomain linkers play a critical role in RNA recognition (Allain et al., 2000; Deo et al., 1999; Handa et al., 1999). Another surprising result is that both dsRBMs are bound in a well-defined location on the R/G stem-loop (dsRBM1 and dsRBM2 are close to the pentaloop and the editing site, respectively), indicating that ADAR2 dsRBMs recognize this RNA substrate by themselves, without the deaminase domain. This finding was not obvious considering that the 71 nt RNA that we used contains 34 base pairs, providing potentially 20 different binding sites for a dsRBM, since each dsRBM binds across 15 base pairs (Ramos et al., 2000; Stefl et al., 2005a; Wu et al., 2004). This specific binding apparently originates from dsRBM2, which prefers an RNA duplex containing mismatches over a regular A-form duplex or a stem-loop, and from dsRBM1, which prefers a stem-loop over a regular duplex. The binding preference of ADAR2 dsRBM1 for a stem-loop containing a stable GCU(A/C)A pentaloop is reminiscent of Rnt1p dsRBM structure-specific recognition of the AGNN tetraloop (Wu et al., 2004) and to Staufen dsRBM3 bound to a stem-loop capped by a UUCG tetraloop (Ramos et al., 2000). Interestingly, all three dsRBMs have similar structures, especially regarding the position of α helix 1. This suggests that dsRBMs' binding preference for stem-loop over regular RNA duplexes might be more general than previously expected. In contrast, dsRBM2 favors RNA duplex substrates that contain mismatches and, more particularly, two cytosines involved in A-C mismatches. Although we cannot tell if this recognition is base specific or structure specific (the backbone deformation around the A-C mismatch), to our knowledge, this is the first structural indication that some dsRBMs specifically recognize RNA mismatches.

Scanning force microscopy also revealed that ADAR2 preferentially binds to the R/G stem-loop over regular duplex regions on an RNA mutant substrate in which the R/G stem-loop was inserted into a potato tuber viroid RNA (Klaue et al., 2003). Interestingly, when the duplex irregularities in the R/G stem-loop were mutated to form a more regular duplex, the ADAR2 was still localized in the vicinity of stem-loop structures, including the R/G stem-loop, but not exclusively. This further demonstrates a preference of ADAR2 for stem-loops that would originate from the presence of dsRBM1. Our findings are further supported by a recent biochemical study of ADAR2 in complex with the GluR-B Q/R site with hydroxyl radical cleavage, in which specific RNA binding of the dsRBMs was observed as well (Stephens et al., 2004).

Deletion of dsRBM1 from ADAR2 decreased the editing of the R/G site by 3-fold, and deletion of dsRBM2 decreased the editing of the R/G site by 10-fold. This highlights the importance of dsRBM2 and its exclusive binding to an RNA helix containing two A-C mismatches separated by ten base pairs adjacent to the R/G site. The weak editing activity of ADAR2 that lacks dsRBM2 could be explained by the dual ability of dsRBM1 for binding to both the stem-loop and the A-C mismatch regions of the R/G stem-loop. It suggests that, in certain circumstances, dsRBM1 can replace dsRBM2 in ADAR2 editing. The essential role of the dsRBM2 interaction with A-C mismatches is consistent with several biochemical experiments showing that ADAR2 forms multiple nonspecific complexes when bound to the R/G stem-loop lacking mismatches (Ohman et al., 2000), resulting in a dramatically reduced editing efficiency and selectivity at the R/G site (Kallman et al., 2003). The binding of dsRBM1 to the stem-loop region that contains the structured GCUCA pentaloop is also important, as the variations in the loop sequence and size have an effect on editing efficiency. This interaction is likely to contribute to the overall binding affinity (Macbeth et al., 2004). In conclusion, this structural study suggests that the dsRBMs of ADAR2 appear to preferentially recognize certain structural elements (the stem-loop and the mismatches) of the R/G stem-loop rather than its sequence, explaining why the secondary structure of the R/G stem-loop is very well conserved (Aruscavage and Bass, 2000).

Implication for ADAR Editing

Our structural study of the ADAR2 dsRBMs demonstrates that dsRBMs can specifically recognize certain secondary structure elements of the R/G stem-loop, a natural ADAR2 substrate encoding the B subunit of the AMPA-subtype of glutamate receptor. These

observations indicate that the R/G stem-loop recognition by the ADAR2 dsRBMs is an important determinant for directing the enzyme to the R/G editing site. How is this related to other editing sites? Recent bioinformatics analyses have predicted more than 12,000 new A-to-I editing sites, located predominantly in ALU repetitive elements in the human transcriptome (Athanasiadis et al., 2004; Blow et al., 2004; Levanon et al., 2004). These analyses showed that A-to-I editing is clearly more frequent at adenosines involved in A-C mismatches than at any other mismatches or base pairs. These findings correlate well with the binding preferences of ADAR2 dsRBM2 observed in our study and suggest that the dsRBM2 of ADAR2 may play a more general role in A-to-I editing site selection than previously expected. Of course, not all A-C mismatches are edited by ADAR2, indicating that dsRBM2 is not the only determinant for the specificity of A-to-I conversion. Our data showed that the dsRBM1 prefers to bind irregular RNA elements like stem-loops or non-Watson-Crick base pairs over regular RNA duplexes (in contrast to the dsRBM2 that binds mismatches but not loop regions). The dsRBM1 of ADAR2 may serve to anchor the protein on long, irregular RNA, consistent with the observation that most A-to-I editing sites are embedded within irregular RNA duplexes. ADAR2 dsRBM1 also prevents ADAR2 from editing small RNA duplexes, as suggested by a recent report describing an autoinhibitory role for dsRBM1 (Macbeth et al., 2004).

Experimental Procedures

Plasmids

Plasmids are described in the Supplemental Data.

Tissue Culture and In Vitro Editing Analysis

Human embryonic kidney (HEK293) cells were transiently cotransfected by calcium phosphate precipitation with cDNAs encoding either a control eGFP expression vector (pEGFP-C1; Clontech), eGFP-ADAR2, eGFP-ΔdsRBM1, or eGFP-ΔdsRBM2 in the presence of a 116 bp GluR-B minigene containing the R/G editing site. Crude nuclear extracts (Schreiber et al., 1989) were prepared from HEK293 cells expressing wild-type or mutant eGFP-ADAR2 and were diluted with dialysis buffer to maintain the enzymatic activity (30 mM HEPES [pH 7.6], 300 mM NaCl, 10% glycerol, 1 mM EDTA, 0.5 mM EGTA, 1 mM DTT, 1 mM PMSF, 2 μg/ml Leupeptin, 0.1% Aprotinin) prior to quantitative Western blotting analysis. Mutant proteins were diluted to achieve the same concentration as wild-type eGFP-ADAR2 protein and were incubated with 100 fmol R/G substrate for 30 min at 30ºC. The eGFP-K(127,281)A double mutant was prepared as previously described (Sansam et al., 2003).

In Vitro Editing of R/G Mutants

In vitro editing of R/G mutants is described in the Supplemental Data.

Recombinant Protein Expression and Purification

Several ADAR2 truncations were expressed and purified as previously described (Stefl et al., 2005b).

RNA Preparation

RNA preparation is described in the Supplemental Data.

NMR Spectroscopy

All of the NMR experiments were conducted at 293 K (or at 315 K) on Bruker Avance-900 and DRX-750, 600, and 500 MHz spectrometers. A detailed description of the resonance assignment procedure of the studied proteins is given elsewhere (Stefl et al., 2005b). RNAs were assigned mostly based on NOESY data, since through-bond techniques suffer from the loss of signal in large RNAs. Specifically, a 2D homonuclear NOESY and TOCSY measured on unlabeled R/G stem-loop RNA, 2D-filtered/edited NOESY experiments (Peterson et al., 2004), a 3D [^{1}H,^{13}C,^{1}H]-NOESY acquired on four base type-specific ^{13}C,^{15}N-(A)-,^{13}C,^{15}N-(U)-,^{13}C^{15}N-(G)-,^{13}C^{15}N-(C)-labeled R/G stem-loop RNAs, and the use of the resonance assignments of the subfragments (Figure 1) were essential in achieving sequential assignment. A full list of experiments used for free and bound proteins and RNAs is given in Table S1. All spectra were processed with XWINNMR (Bruker) and were analyzed with Sparky (Goddard and Kneller, 2004).

Structure Calculations

The preliminary structure determination was performed with the automated NOE assignment module CANDID (Herrmann et al., 2002) in the DYANA program (Guntert et al., 1997). CANDID/DYANA carries out automated assignment and distance calibration of NOE intensities, removal of meaningless restraints, structure calculation with torsion angle dynamics, and automatic upper distance limit violation analysis. The resultant NOE crosspeak assignments were subsequently confirmed by visual inspection of the spectra. In the next step, CANDID/DYANA-generated restraints were used for further refinement of the preliminary structures with AMBER 7.0 software (Case et al., 2002); this process employed a force field described by Cornell et al. (1995), a refinement protocol described in Padrta et al. (2002), and the generalized-Born solvation model (Bashford and Case, 2000). Molecular graphics were generated by using PyMOL (DeLano, 2002) and NUCCYL (Jovine, 2003).

Model for RNA Binding

The model for RNA binding is described in the Supplemental Data.

Use of Sequence/Structure Databases and Sequence Alignments

Sequence/structure databases and sequence alignments are described in the Supplemental Data.

Light-Scattering Experiments

Light-scattering experiments are described in the Supplemental Data.

Supplemental Data

Supplemental Data including Figures S1–S4 are available at http://www.structure.org/cgi/content/full/14/2/345/DC1/.

Acknowledgments

We thank Dr. E. Zobeley for assistance with the light-scattering instrument and Drs. J. D. Alfonso and M.-A. Rubio for insightful discussions. We are grateful to Prof. G. Wagner for the gift of the pET30-GB*Fusion*1 vector. The authors are supported by the Swiss National Science Foundation (Nr. 3100A0-107713), the Roche Research Fund for Biology at the Eidgenössisch Technische Hochschule Zurich (F.H.-T.A.), the National Institutes of Health (RBE; NS33323), and the European Molecular Biology Organization and the Human Frontier Science Program postdoctoral fellowships (R.S.). F.H.T.A. is a European Molecular Biology Organization Young Investigator.

Received: May 6, 2005
Revised: October 26, 2005
Accepted: November 1, 2005
Published: February 10, 2006

References

Allain, F.H., Bouvet, P., Dieckmann, T., and Feigon, J. (2000). Molecular basis of sequence-specific recognition of pre-ribosomal RNA by nucleolin. EMBO J. *19*, 6870–6881.

Aruscavage, P.J., and Bass, B.L. (2000). A phylogenetic analysis reveals an unusual sequence conservation within introns involved in RNA editing. RNA *6*, 257–269.

Athanasiadis, A., Rich, A., and Maas, S. (2004). Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. PLoS Biol. *2*, e391.

Bashford, D., and Case, D.A. (2000). Generalized born models of macromolecular solvation effects. Annu. Rev. Phys. Chem. *51*, 129–152.

Bass, B.L. (2002). RNA editing by adenosine deaminases that act on RNA. Annu. Rev. Biochem. *71*, 817–846.

Bass, B.L., and Weintraub, H. (1988). An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell *55*, 1089–1098.

Bhalla, T., Rosenthal, J.J., Holmgren, M., and Reenan, R. (2004). Control of human potassium channel inactivation by editing of a small mRNA hairpin. Nat. Struct. Mol. Biol. *11*, 950–956.

Blaszczyk, J., Gan, J., Tropea, J.E., Court, D.L., Waugh, D.S., and Ji, X. (2004). Noncatalytic assembly of ribonuclease III with double-stranded RNA. Structure (Camb) *12*, 457–466.

Blow, M., Futreal, P.A., Wooster, R., and Stratton, M.R. (2004). A survey of RNA editing in human brain. Genome Res. *14*, 2379–2387.

Burns, C.M., Chu, H., Rueter, S.M., Hutchinson, L.K., Canton, H., Sanders-Bush, E., and Emeson, R.B. (1997). Regulation of serotonin-2C receptor G-protein coupling by RNA editing. Nature *387*, 303–308.

Bycroft, M., Grunert, S., Murzin, A.G., Proctor, M., and Stjohnston, D. (1995). Nmr solution structure of a Dsrna binding domain from *Drosophila* Staufen protein reveals homology to the N-terminal domain of ribosomal-protein S5. EMBO J. *14*, 3563–3571.

Carlson, C.B., Stephens, O.M., and Beal, P.A. (2003). Recognition of double-stranded RNA by proteins and small molecules. Biopolymers *70*, 86–102.

Case, D.A., Pearlman, D.A., Caldwell, J.W., Cheatham, T.E., III, Wang, J., Ross, W.S., Simmerling, C.L., Darden, T.A., Merz, K.M., Stanton, R.V., et al. (2002). AMBER 7 (http://amber.scripps.edu/).

Cho, D.S., Yang, W., Lee, J.T., Shiekhattar, R., Murray, J.M., and Nishikura, K. (2003). Requirement of dimerization for RNA editing activity of adenosine deaminases acting on RNA. J. Biol. Chem. *278*, 17093–17102.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. (1995). A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. J. Am. Chem. Soc. *117*, 5179–5197.

Dawson, T.R., Sansam, C.L., and Emeson, R.B. (2004). Structure and sequence determinants required for the RNA editing of ADAR2 substrates. J. Biol. Chem. *279*, 4941–4951.

DeLano, W.L. (2002). The PyMOL Molecular Graphics System (http://www.pymol.org).

Deo, R.C., Bonanno, J.B., Sonenberg, N., and Burley, S.K. (1999). Recognition of polyadenylate RNA by the poly(A)-binding protein. Cell *98*, 835–845.

Dominguez, C., Boelens, R., and Bonvin, A.M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J. Am. Chem. Soc. *125*, 1731–1737.

Doyle, M., and Jantsch, M.F. (2002). New and old roles of the double-stranded RNA-binding domain. J. Struct. Biol. *140*, 147–153.

Egebjerg, J., and Heinemann, S.F. (1993). Ca^{2+} permeability of unedited and edited versions of the kainate selective glutamate receptor GluR6. Proc. Natl. Acad. Sci. USA *90*, 755–759.

Emeson, R.B., and Singh, M. (2000). Adenosine to inosine RNA editing: substrates and consequences. In RNA Editing: Frontiers in Molecular Biology, B.L. Bass, ed. (London: Oxford University Press), pp. 109–138.

Fierro-Monti, I., and Mathews, M.B. (2000). Proteins binding to duplexed RNA: one motif, multiple functions. Trends Biochem. Sci. *25*, 241–246.

Gallo, A., Keegan, L.P., Ring, G.M., and O'Connell, M.A. (2003). An ADAR that edits transcripts encoding ion channel subunits functions as a dimer. EMBO J. *22*, 3421–3430.

Gerber, A.P., and Keller, W. (2001). RNA editing by base deamination: more enzymes, more targets, new mysteries. Trends Biochem. Sci. *26*, 376–384.

Goddard, T.D., and Kneller, D.G. (2004). SPARKY 3 (http://www.cgl.ucsf.edu/home/sparky).

Guntert, P., Mumenthaler, C., and 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., and Yokoyama, S. (1999). Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein. Nature *398*, 579–585.

Herrmann, T., Guntert, P., and 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.

Hoopengardner, B., Bhalla, T., Staber, C., and Reenan, R. (2003). Nervous system targets of RNA editing identified by comparative genomics. Science *301*, 832–836.

Jaikaran, D.C., Collins, C.H., and MacMillan, A.M. (2002). Adenosine to inosine editing by ADAR2 requires formation of a ternary complex on the GluR-B R/G site. J. Biol. Chem. *277*, 37624–37629.

Jovine, L. (2003). NUCCYL (http://www.biosci.ki.se/groups/ljo/software/nuccyl.html).

Kallman, A.M., Sahlin, M., and Ohman, M. (2003). ADAR2 A→I editing: site selectivity and editing efficiency are separate events. Nucleic Acids Res. *31*, 4874–4881.

Keegan, L.P., Gallo, A., and O'Connell, M.A. (2001). The many roles of an RNA editor. Nat. Rev. Genet. *2*, 869–878.

Kharrat, A., Macias, M.J., Gibson, T.J., Nilges, M., and Pastore, A. (1995). Structure of the Dsrna binding domain of *Escherichia coli* Rnase-Iii. EMBO J. *14*, 3572–3584.

Klaue, Y., Kallman, A.M., Bonin, M., Nellen, W., and Ohman, M. (2003). Biochemical analysis and scanning force microscopy reveal productive and nonproductive ADAR2 binding to RNA substrates. RNA *9*, 839–846.

Kohler, M., Burnashev, N., Sakmann, B., and Seeburg, P.H. (1993). Determinants of Ca^{2+} permeability in both TM1 and TM2 of high affinity kainate receptor channels: diversity by RNA editing. Neuron *10*, 491–500.

Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R., and Thornton, J.M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J. Biomol. NMR *8*, 477–486.

Leulliot, N., Quevillon-Cheruel, S., Graille, M., Van Tilbeurgh, H., Leeper, T.C., Godin, K.S., Edwards, T.E., Sigurdsson, S.T., Rozenkrants, N., Nagel, R.J., et al. (2004). A new α-helical extension promotes RNA binding by the dsRBD of Rnt1p RNAse III. EMBO J. *23*, 2468–2477.

Levanon, E.Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z.Y., Shoshan, A., Pollock, S.R., Sztybel, D., et al. (2004). Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat. Biotechnol. *22*, 1001–1005.

Lomeli, H., Mosbacher, J., Melcher, T., Hoger, T., Geiger, J.R., Kuner, T., Monyer, H., Higuchi, M., Bach, A., and Seeburg, P.H. (1994). Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. Science *266*, 1709–1713.

Lu, D., Searles, M.A., and Klug, A. (2003). Crystal structure of a zinc-finger-RNA complex reveals two modes of molecular recognition. Nature *426*, 96–100.

Macbeth, M.R., Lingam, A.T., and Bass, B.L. (2004). Evidence for auto-inhibition by the N terminus of hADAR2 and activation by dsRNA binding. RNA *10*, 1563–1571.

Macbeth, M.R., Schubert, H.L., Vandemark, A.P., Lingam, A.T., Hill, C.P., and Bass, B.L. (2005). Inositol hexakisphosphate is bound in the ADAR2 core and required for RNA editing. Science *309*, 1534–1539.

Morse, D.P., and Bass, B.L. (1999). Long RNA hairpins that contain inosine are present in *Caenorhabditis elegans* poly(A)+ RNA. Proc. Natl. Acad. Sci. USA *96*, 6048–6053.

Morse, D.P., Aruscavage, P.J., and Bass, B.L. (2002). RNA hairpins in noncoding regions of human brain and Caenorhabditis elegans mRNA are edited by adenosine deaminases that act on RNA. Proc. Natl. Acad. Sci. USA *99*, 7906–7911.

Mulder, F.A., Schipper, D., Bott, R., and Boelens, R. (1999). Altered flexibility in the substrate-binding site of related native and engineered high-alkaline *Bacillus subtilisins*. J. Mol. Biol. *292*, 111–123.

Nanduri, S., Carpick, B.W., Yang, Y.W., Williams, B.R.G., and Qin, J. (1998). Structure of the double-stranded RNA-binding domain of the protein kinase PKR reveals the molecular basis of its dsRNA-mediated activation. EMBO J. *17*, 5458–5465.

Ohman, M., Kallman, A.M., and Bass, B.L. (2000). In vitro analysis of the binding of ADAR2 to the pre-mRNA encoding the GluR-B R/G site. RNA *6*, 687–697.

Padrta, P., Stefl, R., Kralik, L., Zidek, L., and Sklenar, V. (2002). Refinement of d(GCGAAGC) hairpin structure using one- and two-bond residual dipolar couplings. J. Biomol. NMR *24*, 1–14.

Peterson, R.D., Theimer, C.A., Wu, H., and 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.

Ramos, A., Grunert, S., Adams, J., Micklem, D.R., Proctor, M.R., Freund, S., Bycroft, M., St Johnston, D., and Varani, G. (2000). RNA recognition by a Staufen double-stranded RNA-binding domain. EMBO J. *19*, 997–1009.

Rueter, S.M., Dawson, T.R., and Emeson, R.B. (1999). Regulation of alternative splicing by RNA editing. Nature *399*, 75–80.

Ryter, J.M., and Schultz, S.C. (1998). Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. EMBO J. *17*, 7505–7513.

Sansam, C.L., Wells, K.S., and Emeson, R.B. (2003). Modulation of RNA editing by functional nucleolar sequestration of ADAR2. Proc. Natl. Acad. Sci. USA *100*, 14018–14023.

Schreiber, E., Matthias, P., Muller, M.M., and Schaffner, W. (1989). Rapid detection of octamer binding proteins with 'mini-extracts,' prepared from a small number of cells. Nucleic Acids Res. *17*, 6419.

Seeburg, P.H., Higuchi, M., and Sprengel, R. (1998). RNA editing of brain glutamate receptor channels: mechanism and physiology. Brain Res. Brain Res. Rev. *26*, 217–229.

Sommer, B., Kohler, M., Sprengel, R., and Seeburg, P.H. (1991). RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. Cell *67*, 11–19.

Stefl, R., and Allain, F.H. (2005). A novel RNA pentaloop fold involved in targeting ADAR2. RNA *11*, 592–597.

Stefl, R., Skrisovska, L., and Allain, F.H. (2005a). RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. EMBO Rep. *6*, 33–38.

Stefl, R., Skrisovska, L., Xu, M., Emeson, R.B., and Allain, F.H. (2005b). Resonance assignments of the double-stranded RNA-binding domains of adenosine deaminase acting on RNA 2 (ADAR2). J. Biomol. NMR *31*, 71–72.

Stephens, O.M., Haudenschild, B.L., and Beal, P.A. (2004). The binding selectivity of ADAR2's dsRBMs contributes to RNA-editing selectivity. Chem. Biol. *11*, 1239–1250.

Tonkin, L.A., and Bass, B.L. (2003). Mutations in RNAi rescue aberrant chemotaxis of ADAR mutants. Science *302*, 1725.

Wong, T.C., Ayata, M., Ueda, S., and Hirano, A. (1991). Role of biased hypermutation in evolution of subacute sclerosing panencephalitis virus from progenitor acute measles virus. J. Virol. *65*, 2191–2199.

Wu, H., Henras, A., Chanfreau, G., and Feigon, J. (2004). Structural basis for recognition of the AGNN tetraloop RNA fold by the double-stranded RNA-binding domain of Rnt1p RNase III. Proc. Natl. Acad. Sci. USA *101*, 8307–8312.

Zhang, Z., and Carmichael, G.G. (2001). The fate of dsRNA in the nucleus: a p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. Cell *106*, 465–475.

Zhou, P., Lugovskoy, A.A., and Wagner, G. (2001). A solubility-enhancement tag (SET) for NMR studies of poorly behaving proteins. J. Biomol. NMR *20*, 11–14.

Accession Numbers

Coordinates have been deposited into the Protein Data Bank with accession codes 2B7T and 2B7V for dsRBM1 and dsRBM2, respectively.

A novel RNA pentaloop fold involved in targeting ADAR2

RICHARD STEFL and FRÉDÉRIC H.-T. ALLAIN
Institute of Molecular Biology and Biophysics, Swiss Federal Institute of Technology Zurich, ETH-Hönggerberg, CH-8093 Zürich, Switzerland

ABSTRACT

Adenosine deaminases that act on RNA (ADARs) catalyze the site-specific conversion of adenosine to inosine in primary mRNA transcripts, thereby affecting coding potential of mature mRNAs. Structural determinants that define the adenosine moieties for specific ADARs-mediated deaminations are currently unknown. We report the solution structure of the central region of the human R/G stem–loop pre-mRNA, a natural ADAR2 substrate encoding the subunit B of the glutamate receptor (R/G site). The structure reveals that the GCU(A/C)A pentaloop that is conserved in mammals and birds adopts a novel fold. The fold is stabilized by a complex interplay of hydrogen bonds and stacking interactions. We propose that this new pentaloop structure is an important determinant of the R/G site recognition by ADAR2.

Keywords: adenosine deaminase; NMR; RNA editing; RNA structure

INTRODUCTION

Adenosine deaminases that act on RNA (ADARs) convert adenosines to inosines embedded in double-stranded RNA (dsRNA; Polson et al. 1991; Bass et al. 1997; Emeson and Singh 2000; Gerber and Keller 2001; Bass 2002; Schaub and Keller 2002). The enzymatic activity requires the C-terminal adenosine deaminase domain to catalyze the hydrolytic deamination of adenosine, whereas RNA substrate recognition is mediated by the N-terminal double-stranded RNA-binding motifs (dsRBMs; Stefl et al. 2005). Depending on the RNA substrate structure, ADARs can deaminate as many as 50% of the adenosine residues within a perfect dsRNA (in a nonselective fashion) or as little as a single, specific adenosine moiety (Emeson and Singh 2000; Bass 2002). The majority of nonselective editing occurs in untranslated regions (UTRs) and introns where large regular duplexes are formed (Morse and Bass 1999; Rueter et al. 1999; Morse et al. 2002; Athanasiadis et al. 2004; Levanon et al. 2004). Such modifications can modulate gene silencing triggered by intramolecular structures in mRNA (Tonkin and Bass 2003), lead to the nuclear retention of RNA transcripts (Kumar and Carmichael 1997; Zhang and Carmichael 2001), or participate in an antiviral response (Wong et al. 1991) by extensive modification of viral RNAs. Conversely, selective editing has been shown to take place largely within codons, so that multiple RNA and protein isoforms can be created from a single genomic locus. For example, ADARs have been shown to produce functionally important isoforms for several proteins involved in synaptic neurotransmission, including ligand and voltage-gated ion channels and G-protein coupled receptors (Sommer et al. 1991; Egebjerg and Heinemann 1993; Kohler et al. 1993; Lomeli et al. 1994; Burns et al. 1997; Hoopengardner et al. 2003).

ADAR2 (a member of the mammalian ADAR family) specifically modifies the glutamate receptor subunit B (GluR-B) pre-mRNA at the R/G site (referred as R/G stem–loop), where the genomic **A**GA encoding an arginine is modified after editing into **I**GA encoding a glycine (the ribosome interprets I as G). This amino acid change affects the biophysical properties of the ion channel, the R/G edited protein recovering faster from desensitization (Lomeli et al. 1994). The R/G site is located to a 71-nt intramolecular stem–loop in its pre-mRNA consisting of a stem with a few mismatches and a pentaloop (Fig. 1A). The secondary structure of the R/G stem–loop is strongly conserved in mammals and birds with the GCUMA pentaloop consensus (where M is A/C) (Aruscavage and Bass 2000). To date, there is still no structural insight into the process of A-to-I editing.

RESULTS AND DISCUSSION

We report the solution structure of the central region of the human R/G stem–loop, embedding the pentaloop GCUAA (referred as R/G central loop; Fig. 1B). Prior to structure

Reprint requests to: Frédéric H.-T. Allain, Institute of Molecular Biology and Biophysics, Swiss Federal Institute of Technology Zurich, ETH-Hönggerberg, CH-8093 Zürich, Switzerland; e-mail: allain@mol.biol.ethz.ch; fax: +41 (0) 1 63 31294.

Article and publication are at http://www.rnajournal.org/cgi/doi/10.1261/rna.7276805.

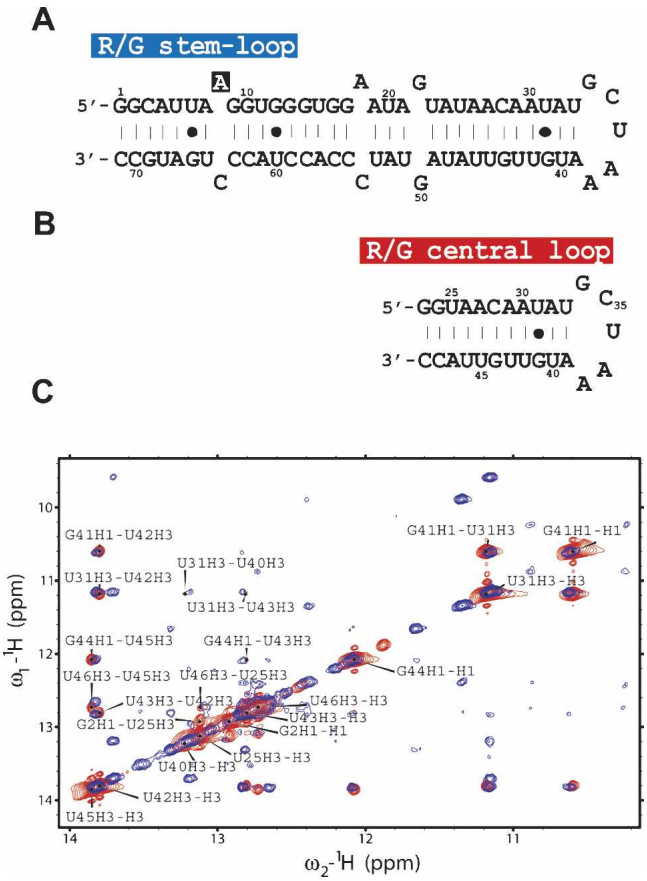


FIGURE 1. Sequence and folding of the R/G stem–loop and central loop. (*A*) R/G stem–loop, a 71-nt intramolecular stem–loop that includes the 67 nt of the human GluR-B R/G pre-mRNA sequence and is closed by two G·C base pairs to improve yield of in vitro transcription. The R/G stem–loop pre-mRNA involves base pairing of intron and exon; the intron/exon junction occurs at the +2 position (G10) relative to the R/G editing site (A8). (*B*) R/G central loop, a 27-nt intramolecular stem–loop that includes the 23 nt of the central part of the human GluR-B R/G pre-mRNA closed by two G·C base pairs. (*C*) 2D [^{1}H-^{1}H] NOESY spectra (in 90% H_2O/10% D_2O) of a 71-nt R/G stem–loop RNA recorded at 900 MHz spectrometer (in blue) and 27-nt R/G central loop recorded at 600 MHz spectrometer (in red). The assignments are labeled by the one-letter code of nucleotides accompanied by a sequence number and a proton name.

determination, we assessed the folding of the R/G central loop and the full-length R/G stem–loop (Fig. 1) by comparing their [^{1}H,^{1}H]-TOCSY and [^{1}H,^{1}H]-NOESY spectra (Fig. 1C). These data confirmed that the central fragment (nt 25–47; Fig. 1) maintains the same conformation adopted within the context of the full-length R/G stem–loop. The structure of the R/G central loop was determined using 602 conformationally restrictive NOE-derived distance restraints, 128 torsion angle restraints, and 85 restraints derived from residual dipolar couplings. The R/G central loop adopts a stable stem–loop structure with a well-defined stem, adopting a standard A-form double-helical architecture, and with a structured GCUAA pentaloop, stabilized by a complex interplay of hydrogen bonding and stacking interactions (Fig. 2). The resulting ensemble of structures is well defined, as demonstrated by the superimposition of the final 20 lowest energy conformers (Fig. 2A). Inclusion of residual dipolar coupling restraints that contain long-range angular information improved the global

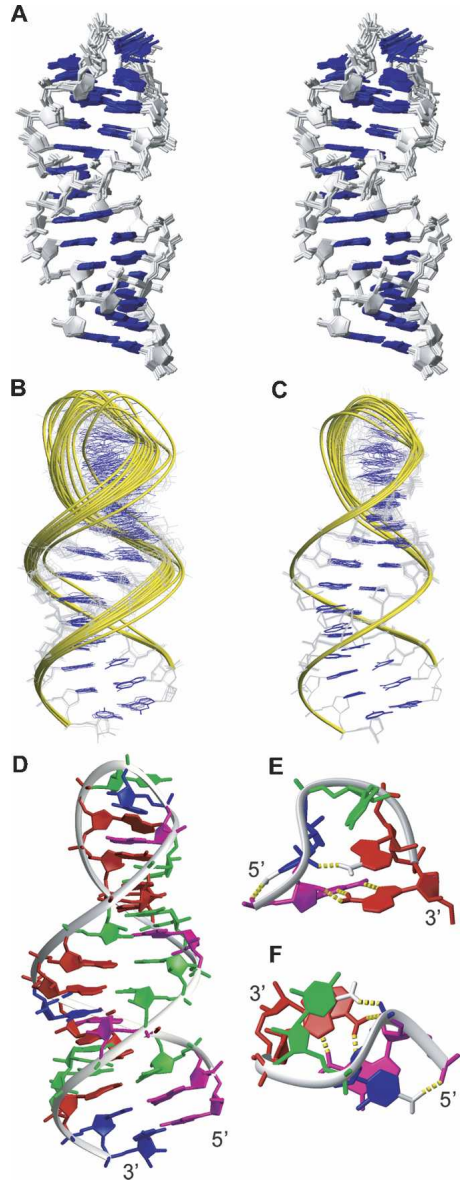


FIGURE 2. Structure of the R/G central loop. (*A*) Stereo view of the NMR ensemble of the R/G central loop, the 13 final lowest energy structures are superimposed over all residues, sugar-phosphate backbones (in gray sticks), and RNA bases (in blue sticks) are shown. (*B*) NMR ensemble calculated without RDCs (13 lowest energy structures) superimposed over the three terminal base pairs. Overall RMSD is 0.84 ± 0.23 Å. (*C*) NMR ensemble calculated with RDCs (13 lowest energy structures) superimposed over the three terminal base pairs. Overall RMSD is 0.72 ± 0.20 Å. (*D*) The lowest energy structure of the R/G central loop represented as a ribbon-and-stick model. Nucleotides are colored red (A), green (U), magenta (G), and blue (C). Only heavy atoms are shown. (*E*,*F*) Close-up views of the GCUAA pentaloop from the minor groove and from the top, respectively. Hydrogen bonds are shown by yellow dashed sticks and the hydrogens involved are displayed.

TABLE 1. NMR experimental restraints and structure statistics

NMR restraints[a]		
Total restraints	840 (31 per residue)	
NOEs	602	
Torsion angles[b]	128	
H-bonds	25	
RDCs	85	
	No. RDCs	RDCs
Structure statistics[c]		
NOE violations		
Number (>0.2 Å)	2.61 ± 1.19	3.69 ± 1.32
Maximum violations (Å)	0.25	0.28
Torsion violations		
Number (>5°)	0	1.0 ± 1.29
Maximum violations (°)	4.8	9.8
RCDs violations		
Number (>2 Hz)		7.38 ± 4.13
Maximum violations (Hz)		4.3
R.M.S.D. from the mean coordinates (Å)[c]		
Heavy atoms	0.84 ± 0.23	0.72 ± 0.20
R.M.S.D. from ideal geometry[c]		
Bonds lengths (Å)	0.011 ± 0.0002	0.012 ± 0.0002
Bonds angles (°)	2.789 ± 0.0309	2.941 ± 0.0320

[a]Final force constants used for square-well penalty functions of individual types of restraints were 32 kcal·mol^{-1}·Å^{-2} for NOEs and Watson–Crick hydrogen bonds, 200 kcal·mol^{-1}·rad^{-2} for torsion angles and 0.3 kcal·mol^{-1}·Hz^{-2} for residual dipolar couplings (RDCs). The force constants were ramped up during the simulated annealing protocol as described (Padrta et al. 2002). Error bounds of ±2 Hz were used for all residual dipolar coupling restraints.
[b]The A-form duplex backbone torsion angles derived from high-resolution crystal structures were used for the double-helical part as described in Materials and Methods: α (270°–330°), β (150°–210°), γ (30°–90°), δ (50°–110°), ε (180°–240°), and ζ (260°–320°).
[c]The statistics (average ± standard deviations) calculated for the bundle of the 13 lowest energy conformers.

precision of the resulting NMR ensemble (Fig. 2B,C). A full summary of structural statistics is given in Table 1. The GCUAA pentaloop adopts a novel fold in which G34 and A38 form a sheared G·A mismatch, C35 is flipped out toward the major groove, and the central U36 stacks over the following A37 in the minor groove (Fig. 2E,F). The flipped out C35 is stabilized by a hydrogen bond between an amino proton and a phosphate oxygen of the preceding residue G34. Similarly, A37 is stabilized by a hydrogen bond between an amino proton and a phosphate oxygen of C35. This latter hydrogen bond explains the A/C conservation at this position in mammals and birds, as the alternative C base (present in rat and mouse species) also has an amino group at a similar position that is likely to be hydrogen bonded to a phosphate oxygen of C35 as well (Fig. 3).

Surprisingly, the structure of the GCUAA pentaloop significantly differs from the structures of GCAUA (Huppler et al. 2002; Fig. 4B), GAAGA (Legault et al. 1998), and GAAAA (Scharpf et al. 2000), which all belong to the GNR(N)A pentaloop family (where N is any nucleotide, R is a purine, and (N) denotes any bulged nucleotide). The only similar feature between the GCUAA pentaloop structure and the GNR(N)A family is the G·A sheared mismatch formed between the 5′-G and 3′-A. In contrast to the GNR(N)A fold that is reminiscent of the most frequently occurring GNRA-type fold (Jucker et al. 1996), except for the additional bulged (N) nucleotide, the GCUAA pentaloop displays rather structural features reminiscent of the UUCG (Allain and Varani 1995; Fig. 4D) and AGNN (Lebars et al. 2001; Wu et al. 2001; Fig. 4C) tetraloop folds. Specifically, the C of the UU**C**G and the G of A**G**NN tetraloops are, similar to the C of the G**C**UAA pentaloop, flipped out toward the major groove, stacked over the first base of the loop, and hydrogen bonded to the closest phosphate-oxygen through its amino group (Fig. 4). The GCUAA pentaloop and UUCG and AGNN tetraloops also have a very similar "S-shape" conformation of the sugar-phosphate backbone (Fig. 4, in black).

The structure of the GCUAA pentaloop explains well the phylogenetic conservation of GCUMA (where M is A/C; Aruscavage and Bass 2000). The first and last positions are invariantly G and A, respectively, to allow formation of the G·A sheared mismatch. Furthermore, the second position is invariantly a C and the fourth position an A or a C, nucleotides conserved to stabilize the conformation of the pentaloop via their amino-phosphate hydrogen bonds. Finally the central nucleotide is a U, probably to prevent the pentaloop from adopting the GNR(N)A-fold.

The structural similarity between the AGNN and UUCG tetraloops and the GCUAA pentaloop is particularly interesting considering that both the AGNN and the UUCG tetraloops are specifically recognized by the dsRBMs of Rnt1p (Wu et al. 2004) and Staufen (Ramos et al. 2000), respectively. In both dsRBM–RNA complexes, the α-helix 1 of the dsRBMs recognizes the specific shapes of the minor groove created by the conserved AGNN and UUCG tetraloop folds (Ramos et al. 2000; Wu et al. 2004). Analogously, we recently observed that the GCUCA pentaloop, rat analog of the human GCUAA pentaloop of the R/G stem–loop pre-mRNA, is specifically recognized by the first dsRBM of rat ADAR2 (R. Stefl and F.H.-T. Allain, unpubl.). In addition, pre-mRNA encoding the R/G site of subunit C of the glutamate receptor that is also specifically edited by ADAR2 has a UCCR tetraloop (Aruscavage and Bass 2000) that belongs to UNCG-fold. Altogether, similar architectures of the reported GCUCA pentaloop and AGNN and UUCG tetraloops indicates that these loops might play a

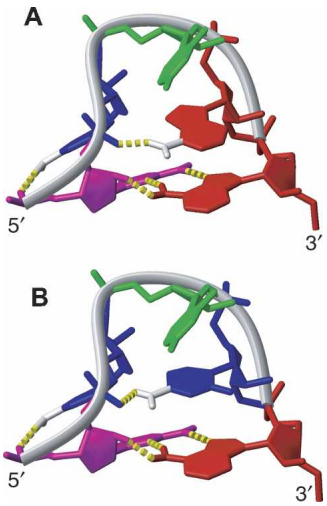


FIGURE 3. Comparison of GCUAA and GCUCA pentaloops, demonstrating the conservation of the GCUMA fold. (*A*) NMR structure of GCUAA pentaloop. Nucleotides are colored red (A), green (U), magenta (G), and blue (C). (*B*) Model of GCUCA pentaloop structure created based on the GCUAA pentaloop structure. The A of GCU**A**A was replaced by the C of GCU**C**A while keeping the conformation of the χ torsion angle.

more important role in the recognition processes mediated by dsRBM-containing proteins than was previously expected.

MATERIALS AND METHODS

RNA sample preparations

Unlabeled, uniformly ^{13}C,^{15}N-labeled RNA, and base-type-specifically ^{13}C,^{15}N-AC/GU-labeled RNA oligonucleotides were prepared by in vitro transcription using T7 polymerase and DNA synthetic primers (Price et al. 1998) and purified either on a denaturing polyacrylamide gel or using HPLC under denaturing conditions. RNA oligonucleotides were annealed at low salt, pH 6.0, and dilute conditions by heating to 95°C and snap cooling on ice to ensure homogenous formation of the monomeric stem–loop conformation. The homogenous formation of the monomeric stem–loop conformation was confirmed by gel filtration coupled with in-line laser light scattering measurement (Wyatt Technology) and native PAGE analysis.

NMR spectroscopy

All the NMR experiments were conducted at 303 K on Bruker Avance-900 and DRX-750, 600, and 500 MHz spectrometers. Exchangeable proton resonances were assigned using two-dimensional (2D) NOESY spectrum (mixing time of 200 msec) of the RNA in 90% H_2O/10% D_2O at 278 K. Nonexchangeable proton resonances were assigned using 2D NOESY, 2D TOCSY, [^1H-^{13}C] HSQC, 3D [^1H-^{13}C-^1H] HCCH TOCSY, and 3D [^1H-^{13}C-^1H] NOESY-HSQC spectra in 99.99% D_2O. NOE restraints of nonexchangeable protons were obtained from 2D NOESY experiments with mixing times of 30, 90, 150, and 250 msec, and 3D NOESY-HSQC experiment with 100 msec mixing time. The NOEs were semi-quantitatively classified based on their intensities in the 2D and 3D NOESY spectra as follows: very strong (1.8–2.4 Å), strong (1.8–3.0 Å) medium (1.8–4.0 Å), weak (1.8–5.0 Å), and very weak (1.8–6.0 Å). NOEs involving exchangeable protons were observed in 2D NOESY in 90% H_2O/10% D_2O and all classified as very weak (1.8–6.0 Å). Hydrogen-bonding distance restraints were used only for the Watson–Crick base pairs, when the imino-proton was observed experimentally. The imino protons of U33 and G34 were not observed in the 2D NOESY, probably due to the exchange with solvent. For G34, it is expected since G34 H1 is not hydrogen bonded. However, the absence of U33 H3 is surprising since our structure shows that U33 is involved in a Watson–Crick base pair with A39. It is possible that the presence of the G34–A38 non-Watson–Crick base pair above U33 made the U33–A39 base pair less stable and therefore U33 H3 more prone to exchange with the solvent. Residual dipolar couplings (RDCs) were measured from $^1J_{CH}$ differences in t_2-coupled HSQCs in the presence or absence of a C12E9/hexanol mixture (Ruckert and Otting 2000). All spectra were processed using XWINNMR (Bruker) and analyzed using Sparky (T.D. Goddard and D.G. Kneller, SPARKY 3, University of California, San Francisco).

Structure calculations

Structures were calculated using AMBER 7.0 software (University of California, San Francisco) with the Cornell et al. (1995) force field, with the generalized-Born solvation model (Bashford and Case 2000), and with a refinement protocol as described (Padrta et al. 2002). Square-well penalty functions with force constants as indicated in Table 1 were used for all experimental restraints. Refinements were performed in two steps. In the first step, refinement with NOEs and torsion angles was done starting from com-

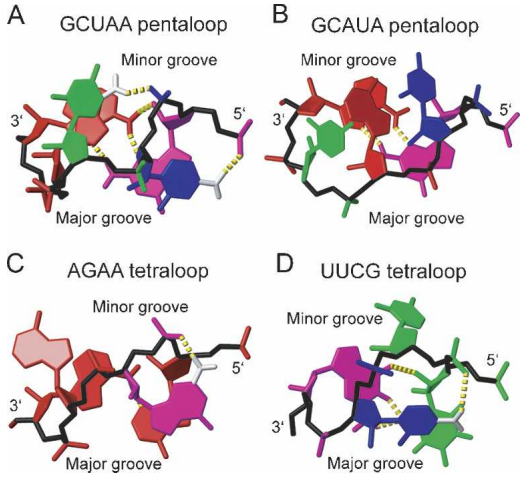


FIGURE 4. Comparison between the structures of the GCUAA (*A*), GCAUA (Huppler et al. 2002) (*B*), AGAA (Lebars et al. 2001; Wu et al. 2001) (*C*), and UUCG (Allain and Varani 1995) loops (*D*). Nucleotides are colored red (A), green (U), magenta (G), and blue (C). Sugar-phosphate backbones are shown in black. Hydrogen bonds are shown as yellow dashed sticks. Amino groups involved in H bonding are depicted; those involved in contacting phosphate oxygen are in white. Otherwise only heavy atoms are shown.

pletely randomized structures. In the second step, the RDCs were employed in addition to NOE and torsion angle restraints. Initial estimates of the alignment tensors were obtained from the preliminary structures resulting from the first refinement step, as described (Tsui et al. 2000). To impose better convergence of the ensembles some artificial torsion angle restraints were used for α (270°–330°), β (150°–210°), γ (30°–90°), δ (50°–110°), ε (180°–240°), and ζ (260°–320°); these torsion angle restraints cover the A-form RNA double-helical range derived from high-resolution crystal structures of RNA (NDB ID: AR0010 [R = 1.16 Å], AR0013 [R = 1.2 Å], AR0020 [R = 1.20 Å], AR0021 [R = 1.20 Å], ARB002 [R = 0.8 Å], ARB003 [R = 1.1 Å], ARB004 [R = 0.9 Å], ARB005 [R = 0.8 Å], ARF0108 [R = 1.8 Å], and ARH064 [R = 1.8 Å]). These torsion angle restraints do not exert any force on the final structures. To prevent high violations of local geometries while accommodating the RDC restraints, additional angle restraints to maintain proper local geometries were employed (Tsui et al. 2000). The family of best structures (13 out of 30 calculated) was selected on the basis of lowest energy criteria. Molecular graphics was generated using MOLMOL (Koradi et al. 1996). The hydrogen bonds found in the pentaloop were not observed experimentally but were deduced based on distances and angles between the proton, the donor, and the acceptor in the ensemble of structures. Atomic coordinates and NMR restraints have been deposited into the Protein Data Bank under accession code (1YSV). The chemical shifts have been deposited into the BioMagResBank Database (accession code BMRB-6485).

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation and the Roche Research Fund for Biology at the ETH Zurich (F.H.T.A.) and the European Molecular Biology Organization and the Human Frontier Science Program postdoctoral fellowships (R.S.). F.H.T.A. is an EMBO Young Investigator.

Received December 20, 2004; accepted February 2, 2005.

REFERENCES

Allain, F.H. and Varani, G. 1995. Structure of the P1 helix from group I self-splicing introns. *J. Mol. Biol.* **250:** 333–353.

Aruscavage, P.J. and Bass, B.L. 2000. A phylogenetic analysis reveals an unusual sequence conservation within introns involved in RNA editing. *RNA* **6:** 257–269.

Athanasiadis, A., Rich, A., and Maas, S. 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS* **2:** e391.

Bashford, D. and Case, D.A. 2000. Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **51:** 129–152.

Bass, B.L. 2002. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* **71:** 817–846.

Bass, B.L., Nishikura, K., Keller, W., Seeburg, P.H., Emeson, R.B., O'Connell, M.A., Samuel, C.E., and Herbert, A. 1997. A standardized nomenclature for adenosine deaminases that act on RNA. *RNA* **3:** 947–949.

Burns, C.M., Chu, H., Rueter, S.M., Hutchinson, L.K., Canton, H., Sanders-Bush, E., and Emeson, R.B. 1997. Regulation of serotonin-2C receptor G-protein coupling by RNA editing. *Nature* **387:** 303–308.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W., and Kollman, P.A. 1995. A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. *J. Am. Chem. Soc.* **117:** 5179–5197.

Egebjerg, J. and Heinemann, S.F. 1993. Ca^{2+} permeability of unedited and edited versions of the kainate selective glutamate receptor GluR6. *Proc. Natl. Acad. Sci.* **90:** 755–759.

Emeson, R.B. and Singh, M. 2000. Adenosine to inosine RNA editing: Substrates and consequences. In *RNA editing: Frontiers in molecular biology* (ed. B.L. Bass), pp. 109–138. Oxford University Press, London.

Gerber, A.P. and Keller, W. 2001. RNA editing by base deamination: Enzymes, more targets, new mysteries. *Trends Biochem. Sci.* **26:** 376–384.

Hoopengardner, B., Bhalla, T., Staber, C., and Reenan, R. 2003. Nervous system targets of RNA editing identified by comparative genomics. *Science* **301:** 832–836.

Huppler, A., Nikstad, L.J., Allmann, A.M., Brow, D.A., and Butcher, S.E. 2002. Metal binding and base ionization in the U6 RNA intramolecular stem-loop structure. *Nat. Struct. Biol.* **9:** 431–435.

Jucker, F.M., Heus, H.A., Yip, P.F., Moors, E.H., and Pardi, A. 1996. A network of heterogeneous hydrogen bonds in GNRA tetraloops. *J. Mol. Biol.* **264:** 968–980.

Kohler, M., Burnashev, N., Sakmann, B., and Seeburg, P.H. 1993. Determinants of Ca^{2+} permeability in both TM1 and TM2 of high affinity kainate receptor channels: Diversity by RNA editing. *Neuron* **10:** 491–500.

Koradi, R., Billeter, M., and Wuthrich, K. 1996. MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14:** 29–32, 51–55.

Kumar, M. and Carmichael, G.G. 1997. Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts. *Proc. Natl. Acad. Sci.* **94:** 3542–3547.

Lebars, I., Lamontagne, B., Yoshizawa, S., Aboul-Elela, S., and Fourmy, D. 2001. Solution structure of conserved AGNN tetraloops: Insights into Rnt1p RNA processing. *EMBO J.* **20:** 7250–7258.

Legault, P., Li, J., Mogridge, J., Kay, L.E., and Greenblatt, J. 1998. NMR structure of the bacteriophage λ N peptide/boxB RNA complex: Recognition of a GNRA fold by an arginine-rich motif. *Cell* **93:** 289–299.

Levanon, E.Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z.Y., Shoshan, A., Pollock, S.R., Sztybel, D., et al. 2004. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol.* **22:** 1001–1005.

Lomeli, H., Mosbacher, J., Melcher, T., Hoger, T., Geiger, J.R., Kuner, T., Monyer, H., Higuchi, M., Bach, A., and Seeburg, P.H. 1994. Control of kinetic properties of AMPA receptor channels by nuclear RNA editing. *Science* **266:** 1709–1713.

Morse, D.P. and Bass, B.L. 1999. Long RNA hairpins that contain inosine are present in *Caenorhabditis elegans* poly(A)+ RNA. *Proc. Natl. Acad. Sci.* **96:** 6048–6053.

Morse, D.P., Aruscavage, P.J., and Bass, B.L. 2002. RNA hairpins in noncoding regions of human brain and *Caenorhabditis elegans* mRNA are edited by adenosine deaminases that act on RNA. *Proc. Natl. Acad. Sci.* **99:** 7906–7911.

Padrta, P., Stefl, R., Kralik, L., Zidek, L., and Sklenar, V. 2002. Refinement of d(GCGAAGC) hairpin structure using one- and two-bond residual dipolar couplings. *J. Biomol. NMR* **24:** 1–14.

Polson, A.G., Crain, P.F., Pomerantz, S.C., McCloskey, J.A., and Bass, B.L. 1991. The mechanism of adenosine to inosine conversion by the double-stranded RNA unwinding/modifying activity: A high-performance liquid chromatography–mass spectrometry analysis. *Biochemistry* **30:** 11507–11514.

Price, S., Oubridge, C., Varani, G., and Nagai, K. 1998. Preparation of RNA–protein complexes for X-ray crystallography and NMR. In *RNA–protein interaction: Practical approach* (ed. C. Smith), pp. 37–74. Oxford University Press, Oxford, UK.

Ramos, A., Grunert, S., Adams, J., Micklem, D.R., Proctor, M.R., Freund, S., Bycroft, M., St Johnston, D., and Varani, G. 2000. RNA

recognition by a Staufen double-stranded RNA-binding domain. *EMBO J.* **19:** 997–1009.

Ruckert, M. and Otting, G. 2000. Alignment of biological macromolecules in novel nonionic liquid crystalline media for NMR experiments. *J. Am. Chem. Soc.* **122:** 7793–7797.

Rueter, S.M., Dawson, T.R., and Emeson, R.B. 1999. Regulation of alternative splicing by RNA editing. *Nature* **399:** 75–80.

Scharpf, M., Sticht, H., Schweimer, K., Boehm, M., Hoffmann, S., and Rosch, P. 2000. Antitermination in bacteriophage λ. The structure of the N36 peptide–boxB RNA complex. *Eur. J. Biochem.* **267:** 2397–2408.

Schaub, M. and Keller, W. 2002. RNA editing by adenosine deaminases generates RNA and protein diversity. *Biochimie* **84:** 791–803.

Sommer, B., Kohler, M., Sprengel, R., and Seeburg, P.H. 1991. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell* **67:** 11–19.

Stefl, R., Skrisovska, L., and Allain, F.H. 2005. RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. *EMBO Rep.* **6:** 33–38.

Tonkin, L.A. and Bass, B.L. 2003. Mutations in RNAi rescue aberrant chemotaxis of ADAR mutants. *Science* **302:** 1725.

Tsui, V., Zhu, L., Huang, T.H., Wright, P.E., and Case, D.A. 2000. Assessment of zinc finger orientations by residual dipolar coupling constants. *J. Biomol. NMR* **16:** 9–21.

Wong, T.C., Ayata, M., Ueda, S., and Hirano, A. 1991. Role of biased hypermutation in evolution of subacute sclerosing panencephalitis virus from progenitor acute measles virus. *J. Virol.* **65:** 2191–2199.

Wu, H., Yang, P.K., Butcher, S.E., Kang, S., Chanfreau, G., and Feigon, J. 2001. A novel family of RNA tetraloop structure forms the recognition site for *Saccharomyces cerevisiae* RNase III. *EMBO J.* **20:** 7240–7249.

Wu, H., Henras, A., Chanfreau, G., and Feigon, J. 2004. Structural basis for recognition of the AGNN tetraloop RNA fold by the double-stranded RNA-binding domain of Rnt1p RNase III. *Proc. Natl. Acad. Sci.* **101:** 8307–8312.

Zhang, Z. and Carmichael, G.G. 2001. The fate of dsRNA in the nucleus: A p54(nrb)-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. *Cell* **106:** 465–475.

EMBO reports

review

RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle

Richard Stefl, Lenka Skrisovska & Frédéric H.-T. Allain[+]
Swiss Federal Institute of Technology Zürich, Zürich, Switzerland

At all stages of its life (from transcription to translation), an RNA transcript interacts with many different RNA-binding proteins. The composition of this supramolecular assembly, known as a ribonucleoprotein particle, is diverse and highly dynamic. RNA-binding proteins control the generation, maturation and lifespan of the RNA transcript and thus regulate and influence the cellular function of the encoded gene. Here, we review our current understanding of protein–RNA recognition mediated by the two most abundant RNA-binding domains (the RNA-recognition motif and the double-stranded RNA-binding motif) plus the zinc-finger motif, the most abundant nucleic-acid-binding domain. In addition, we discuss how not only the sequence but also the shape of the RNA are recognized by these three classes of RNA-binding protein.

Keywords: double-stranded RNA-binding motif; RNA-binding proteins; RNA recognition; RNA-recognition motif; zinc-finger motif
EMBO reports (2005) **6,** 33–38. doi:10.1038/sj.embor.7400325

Introduction

The association of RNA-binding proteins (RBPs) with RNA transcripts begins during transcription. Some of these early-binding RBPs remain bound to the RNA until it is degraded, whereas others recognize and transiently bind to RNA at later stages for specific processes such as splicing, processing, transport and localization (Dreyfuss *et al*, 2002). The RBPs cover the RNA transcripts and control their fate. Some RBPs function as RNA chaperones (Lorsch, 2002) by helping the RNA, which is initially single-stranded, to form various secondary or tertiary structures. When folded, these structured RNAs, together with specific RNA sequences, act as a signal for other RBPs that mediate gene regulation. Here, we review our current structural understanding of protein–RNA recognition mediated by the two most abundant RNA-binding domains, the RNA-recognition motif (RRM) and the double-stranded RNA-binding motif (dsRBM), and by the most abundant nucleic-acid-binding motif, the CCHH-type zinc-finger domain. We discuss how these three small domains recognize RNA: some bind single-stranded RNA by direct readout of the primary sequence, whereas others recognize primarily the shape of the RNA or both the sequence and the shape. Other types of RNA-binding domains, such as the K-homology (KH) domain or the oligonucleotide/oligosaccharide-binding (OB) fold, have recently been reviewed and are not discussed here (Messias & Sattler, 2004).

Institute for Molecular Biology and Biophysics, Swiss Federal Institute of Technology Zürich, ETH-Hönggerberg, CH-8093 Zürich, Switzerland
[+]Corresponding author. Tel: +41 (0)1 63 33940; Fax: +41 (0)1 63 31294;
E-mail: allain@mol.biol.ethz.ch

Submitted 22 September 2004; accepted 26 November 2004

RNA shape-dependent recognition by double-stranded RBM

The dsRBM is a 70–75 amino-acid domain with a conserved αββββα protein topology in which the two α-helices are packed along one face of a three-stranded anti-parallel β-sheet (Fig 1; Fierro-Monti & Mathews, 2000; St Johnston *et al*, 1992). These domains occur mostly in multiple copies (up to five) and have so far been found in 388 eukaryotic proteins, 72 of which are human (data taken from the SMART database; Letunic *et al*, 2004). These proteins have an essential role in RNA interference, RNA processing, RNA localization, RNA editing and translational repression (Doyle & Jantsch, 2002; Saunders & Barber, 2003).

So far, only three structures of dsRBMs in complex with dsRNA have been determined (Table 1): a 1.9 Å crystal structure of the second dsRBM of *Xenopus laevis* RNA-binding protein A (Xlrbpa2) bound to two coaxially stacked dsRNA molecules, each 10 bp long (Ryter & Schultz, 1998); a nuclear magnetic resonance (NMR) structure of the third dsRBM from the *Drosophila* Staufen protein in complex with a symmetrical GC-rich 12-bp duplex capped by a UUCG tetraloop (Ramos *et al*, 2000); and an NMR structure of the dsRBM of Rnt1p (an RNase III homologue from budding yeast) bound to a 14-bp RNA duplex capped by an AGAA tetraloop (Wu *et al*, 2004). All three structures have several common features that reveal how a dsRBM is able to bind to any dsRNA but not to dsDNA, regardless of its base composition. The dsRBMs interact along one face of the RNA duplex through both α-helices and their β1–β2 loop (Fig 1). The contacts with the RNA cover 15 bp that span two consecutive minor grooves separated by a major groove. In all three structures, the contacts to the sugar-phosphate backbone of the major groove and of one minor groove (Fig 1) are mediated by the β1–β2 loop and the amino-terminal part of α-helix 2. These interactions are non-sequence-specific as they involve 2′-hydroxyls and phosphate oxygens and are perfectly adapted to the shape of an RNA double helix. By contrast, the interactions mediated by α-helix 1 are different in all three complexes. In the dsRBM of Xlrbpa2, α-helix 1 interacts non-specifically with the other minor groove of the RNA (Fig 1A), with a few contacts to the bases. In the dsRBM of Staufen, α-helix 1

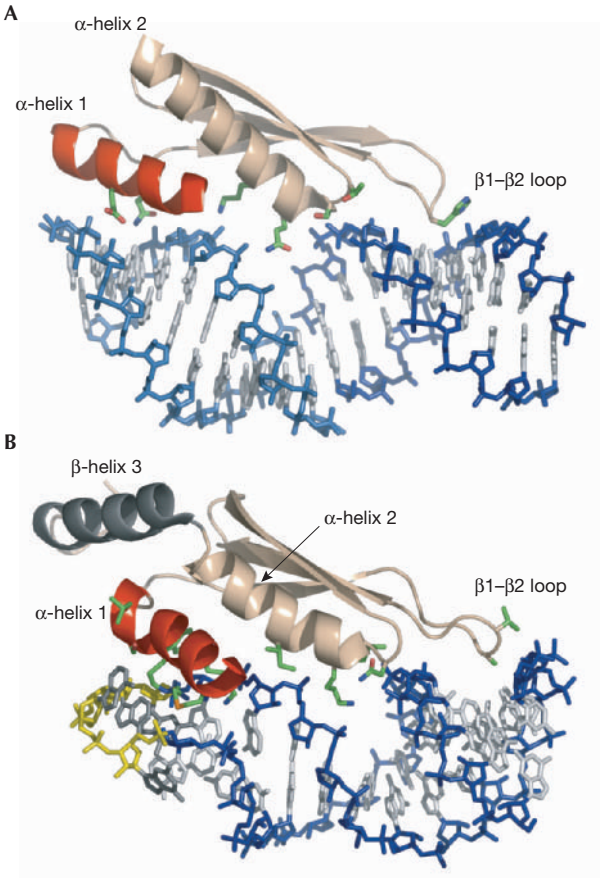


Fig 1 | Double-stranded RNA recognition by double-stranded RNA-binding motifs. (**A**) The double-stranded RNA-binding motif (dsRBM) of Xlrbpa2 bound to dsRNA (Ryter & Schultz, 1998). The α-helix 1 (in red), amino-terminal part of α-helix 2, and β1–β2 loop recognize non-sequence-specifically the shape of dsRNA. Backbones are coloured blue and light blue for two co-axially stacked duplexes. (**B**) The dsRBM of Rnt1p bound to the stem–loop closed by the AGNN tetraloop (backbone in yellow and bases in black; Wu *et al*, 2004). The α-helix 1, a key element for shape-specific recognition by dsRBMs, is highlighted in red. The dsRBM of Rnt1p has an additional carboxy-terminal α-helix 3 (in black), that modulates the conformation of α-helix 1 (Leulliot *et al*, 2004). Side chains involved in intermolecular interactions are shown.

interacts with a UUCG tetraloop that caps the RNA double helix. Although the UUCG tetraloop is not a natural substrate of Staufen, this finding led to the proposal that α-helix 1 modulates the specificity of individual dsRBMs (Ramos *et al*, 2000). Indeed, this was recently confirmed by the structure of the dsRBM of Rnt1p bound to its natural RNA substrate (Fig 1B), in which α-helix 1 recognizes the specific shape of the minor groove created by the conserved AGNN tetraloop (Wu *et al*, 2004). The α-helix 1, the conformation of which is stabilized by an additional carboxy-terminal α-helix 3 (Fig 1B; Leulliot *et al*, 2004), is tightly inserted into the RNA minor groove and contacts the sugar-phosphate backbone and the two non-conserved tetraloop bases, whereas the conserved A and G bases are not involved in the interactions (Wu *et al*, 2004). This structure illustrates how this dsRBM recognizes the specific shape of its RNA target but not its sequence. dsRBMs are highly conserved and have the same structural framework, but are chemically distinct through variations in key residues. The structure of the dsRBM of Rnt1p in complex with RNA highlights the essential role of the α-helix 1 in the recognition of structured elements that deviate from regular dsRNA. The α-helix 1 is the least-conserved secondary structure element among various dsRBMs and seems to have a different spatial arrangement relative to the rest of the domain in different dsRBMs. This variability may be an important factor as many biochemical experiments have shown that dsRBM-containing proteins have binding specificity for a variety of RNA structures, such as stem–loops, internal loops, bulges or helices with mismatches (Doyle & Jantsch, 2002; Fierro-Monti & Mathews, 2000; Ohman *et al*, 2000; Stephens *et al*, 2004). Clearly, further structures are needed to decipher the extent of RNA shape-dependent recognition by dsRBMs.

RNA sequence- and shape-dependent recognition by an RRM

The RRM is the most common RNA-binding motif. It is a small protein domain of 75–85 amino acids with a typical βαββαβ topology that forms a four-stranded β-sheet packed against two α-helices (Mattaj, 1993). RRMs are found in about 0.5%–1% of human genes (Venter *et al*, 2001) and are often present in multiple copies (up to six per protein). RRM-domain-containing proteins are involved in many cellular functions, particularly messenger RNA and ribosomal RNA processing, splicing and translation regulation, RNA export and RNA stability (Dreyfuss *et al*, 2002).

So far, ten structures of an RRM in complex with RNA have been determined using either NMR spectroscopy or X-ray crystallography (Table 1). These structures reveal the complexity of protein–RNA recognition mediated by the RRM, which often involves not only protein–RNA interactions but also RNA–RNA and protein–protein interactions. All ten structures reveal some common features. The main protein surface of the RRM involved in the interaction with the RNA is the four-stranded β-sheet, which usually contacts two or three nucleotides (exemplified here by the RRM1 of sex-lethal; Fig 2A; Handa *et al*, 1999). The nucleotides are located on the surface of the β-sheet, with the bases oriented parallel to the β-sheet plane and often packed against conserved hydrophobic side-chains (usually aromatics). These two or three nucleotides are recognized sequence-specifically by interactions with the protein side-chains of the β-sheet and with the main-chain and side-chains of the residues carboxy-terminal to the β-sheet. Interestingly, it seems that almost all possible sequences (doublets or triplets) can be accommodated on such a surface as the RNA sequences are different in each structure (Table 1).

Often, RRM-containing proteins bind more than three nucleotides and recognize longer single-stranded RNA (for example, poly(A)-binding protein (PAPB; Deo *et al*, 1999), sex-lethal (Handa *et al*, 1999), Hu protein D (HuD; Wang & Hall, 2001), heterogeneous nuclear RNP A1 (hnRNP A1; Ding *et al*, 1999), nucleolin (Allain *et al*, 2000; Johansson *et al*, 2004), RNA stem–loops U1A (Oubridge *et al*, 1994), U2B″ (Price *et al*, 1998), nucleolin (Allain *et al*, 2000)) or even internal loops (U1A; Allain *et al*, 1997; Varani *et al*, 2000), all with high affinity ($K_d \approx 10^{-9}M^{-1}$). In U1A, U2B″, nucleolin and sex-lethal, two loops between the secondary-structure elements of the RRM (the β2–β3 loop and the β1–α1 loop) are essential for additional contacts with the RNA (Fig 2B). These loops vary significantly in size and amino-acid sequence between the different RRMs. In the RRM of CBP20, the C- and N-terminal extensions (which are stabilized by the cognate

Table 1 | Various structures of RNA-binding proteins bound to RNA

Complex	RNA secondary structure	No. of RBDs in structures vs in full-length protein	RNA sequence recognized specifically by RRM β-sheet	Function	Reference; Protein Data Bank (PDB) ID
dsRBM type					
Second dsRBM of Xlrbpa2	Duplex[a]	1/3	–	hnRNP association, translation repression	Ryter & Schultz, 1998; 1DI2
Third dsRBM of Staufen	Stem–loop[b]	1/5	–	mRNA localization, translation control	Ramos *et al*, 2000; 1EKZ
dsRBM of Rnt1p	Stem–loop	1/1	–	RNA processing	Wu *et al*, 2004; 1T4L
RRM type					
N-terminal RRM of U2B" (in U2B"–U2A'–RNA complex)	Stem–loop	1/2	CAG	Pre-mRNA splicing	Price *et al*, 1998; 1A9N
N-terminal RRM of U1A	Stem–loop	1/2	CAC	Pre-mRNA splicing	Oubridge *et al*, 1994; 1URN
N-terminal RRM of U1A	Internal loop	1/2	CAC	Pre-mRNA splicing	Allain *et al*, 1997; 1AUD, Varani *et al*, 2000; 1DZ5
Two N-terminal RRMs of nucleolin	Stem–loop	2/4	RRM1-CG RRM2-UC	Ribosome biogenesis	Allain *et al*, 2000; 1FJE, Johansson *et al*, 2004; 1RKJ
Two N-terminal RRM of PABP	Single strand	2/4	RRM1-AAA RRM2-AAA	Translation initiation	Deo *et al*, 1999; 1CVJ
Two RRMs of sex-lethal	Single strand	2/2	RRM1-UUU RRM2-UGU	Alternative splicing	Handa *et al*, 1999; 1H2T
Two N-terminal RRMs of HuD	Single strand	2/3	RRM1-UUU RRM2-UU	mRNA stability, translation regulation	Wang & Hall, 2001; 1FXL, 1G2E
RRM of CBP20	Single strand	1/1	m7GpppG	Maturation of pre-mRNA and U-rich snRNA	Mazza *et al*, 2002; 1H2V
Zinc finger					
Fourth, fifth and sixth zinc fingers (CCHH-type) of TFIIIA	Truncated 5S RNA[c]	3/9	–	Transcription regulation	Lu *et al*, 2003; 1UN6
First and second zinc fingers (CCCH-type) of TIS11d	Single strand	2/2	–	RNA processing and degradation	Hudson *et al*, 2004; 1RGO

[a]Two coaxially stacked dsRNA (each 10-bp long). [b]12-bp duplex capped by a non-physiologically relevant UUCG tetraloop. [c]Consists of loop A, loop E and helices I, IV and V.
CBP20, cap-binding protein 20; dsRBM, double-stranded RNA-binding motif; hnRNP, heterogeneous nuclear ribonucleoprotein; HuD, Hu protein D; PABP, poly(A)-binding protein; RBD, RNA-binding domain; Rnt1p, RNase III homologue; RRM, RNA-recognition motif; TFIIIA, transcription factor IIIA; Xlrbpa2, *Xenopus laevis* RNA-binding protein A.

protein CBP80) provide a tight binding pocket for the 5′ capped RNAs (7-methyl-G(5′)ppp(5′)N, where N is any nucleotide; Fig 2C; Mazza *et al*, 2002). In proteins that contain several RRMs, high-affinity binding can only be achieved by the cooperative binding of at least two RRMs to the RNA (for example, in nucleolin (Fig 2D), PABP and sex-lethal). In addition to the β-sheet–RNA contacts, interactions between the inter-domain linker and the RNA and between the RRMs themselves contribute to the marked increase in affinity compared with the binding of the individual domain alone. These structures show that the RRM is a platform with a large capacity for variation in order to achieve high RNA-binding affinity and specificity. For example, it is remarkable that a single domain like nucleolin RRM2 contacts only two nucleotides, whereas U1A RRM1 contacts 12 nucleotides and the RRM of Y14 (Fribourg *et al*, 2003) does not contact RNA but rather another protein. This fascinating plasticity of the RRM explains why it is so abundant and why it is involved in so many different biological functions; however, this plasticity makes it difficult to predict how the RRM achieves RNA recognition.

RNA recognition by zinc fingers

CCHH-type zinc-finger domains are the most common DNA-binding domain found in eukaryotic genomes. Typically, several fingers are used in a modular fashion to achieve high sequence-specific recognition of DNA (Miller *et al*, 1985). Each finger displays a ββα protein fold in which a β-hairpin and an α-helix are pinned together by a Zn^{2+} ion. DNA-sequence-specific recognition is achieved by the interactions between protein side-chains of the α-helix (at position –1, 2, 3 and 6, for the canonical arrangement) and the DNA bases in the major groove (Fig 3A; Wolfe *et al*, 2000). However, there is increasing evidence that zinc fingers are also used to recognize RNA (Finerty & Bass, 1997; Mendez-Vidal *et al*, 2002; Picard & Wegnez, 1979; Theunissen *et al*, 1992). The crystal structure of three zinc fingers (fingers 4–6) of transcription factor IIIA (TFIIIA) in complex with a 61-nucleotide fragment of the 5S RNA (Lu *et al*, 2003) provided the first insight into RNA recognition by CCHH-type zinc fingers. In this structure, finger 4 binds to loop E, finger 5 to helix V, and finger 6 to loop A (Fig 3B). Finger 4 recognizes loop E by specifically interacting with a bulged guanosine (Fig 3C) and, similarly, finger 6 recognizes loop A by specifically

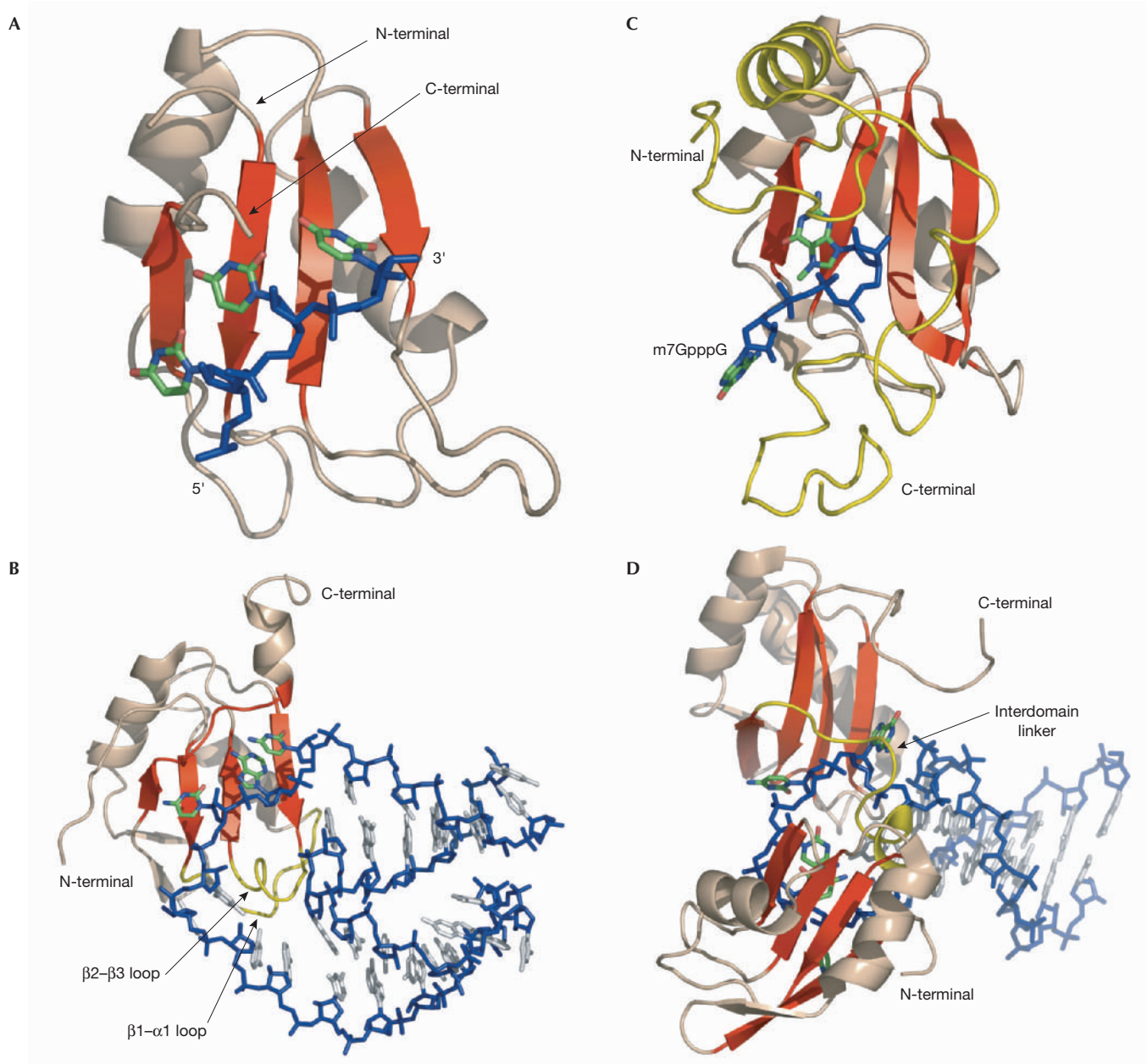


Fig 2 | RNA recognition by RNA-recognition motifs. The similarities and differences are highlighted in red and yellow, respectively. (**A**) RNA-recognition motif 1 (RRM1) of sex-lethal (shown as a ribbon model) interacts with the triplet UUU (shown as a stick model; Handa *et al*, 1999). The four-stranded β-sheet (in red) recognizes three nucleotides—this is the canonical mode of RRM–RNA interaction. (**B**) RRM1 of U1A bound to an RNA internal loop (Allain *et al*, 1997). The four-stranded β-sheet (in red) recognizes three nucleotides. The β2–β3 loop and the β1–α1 loop (in yellow) contact additional RNA residues. (**C**) RRM of CBP20 complexed with m^7GpppG (7-methyl-G(5′)ppp(5′)G; Mazza *et al*, 2002). C- and N-terminal extensions (in yellow) provide additional protein–RNA contacts that creates a specific binding pocket for 5′ capped RNAs. (**D**) RRMs 1 and 2 of nucleolin bound to an RNA stem–loop (Allain *et al*, 2000). The four-stranded β-sheet of RRM2 (in red) recognizes only two nucleotides. The interdomain linker (in yellow) participates in the recognition of the hairpin architecture.

interacting with two bases (an adenine and a cytosine) that also bulge out from the rest of the RNA (Fig 3B). The specific recognition of the RNA by both fingers 4 and 6 is achieved by side-chain contacts from the N-terminal parts of the α-helix (at position –1, 1 and 2; Fig 3C). The interaction of finger 5 with helix V differs from the ones made by fingers 4 and 6. In this case, finger 5 recognizes a short RNA double helix by multiple contacts between basic amino acids of the α-helix and the RNA sugar-phosphate backbone (Fig 3D).

In contrast to the above-mentioned CCHH zinc fingers, another class of zinc fingers (CCCH-type) was recently found to adopt a different fold and to recognize sequence-specifically single-stranded RNA (Hudson *et al*, 2004). In this NMR structure, sequence-specific

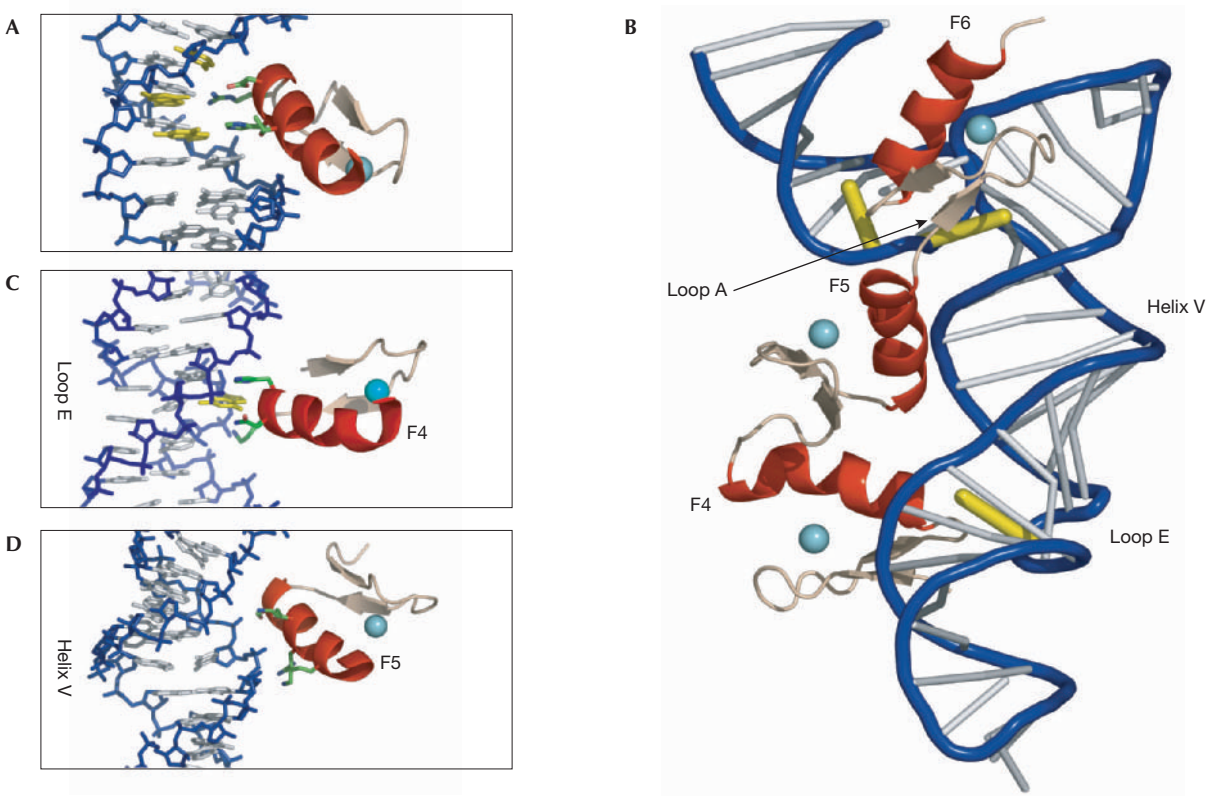


Fig 3 | DNA vs RNA recognition by CCHH-type zinc fingers. (**A**) Zinc finger 2 of Zif268 bound to double-stranded DNA (Pavletich & Pabo, 1991). The α-helix of the zinc finger (in red) inserts into the DNA major groove; base contacts are made from positions –1, 2, 3 and 6 of the α-helix (the protein side-chains are shown as an element-type coloured stick model). The DNA bases that are recognized by the finger are coloured yellow. (**B**) Overall view of the complex of transcription factor IIIA (TFIIIA) fingers 4–6 (F4–F6) and 61-nucleotide 5S RNA (Lu *et al*, 2003). The protein and RNA are represented as ribbon models. The bulged bases involved in the recognitions are highlighted in yellow. Cyan balls represent zinc ions. (**C**) TFIIIA finger 4 (F4) bound to loop E (Lu *et al*, 2003). The α-helix (in red) of the finger 4 specifically interacts with a guanosine base that bulges out (in yellow); the base contacts are made from the side-chain at position –1, 1 and 2 of the α-helix. (**D**) TFIIIA finger 5 (F5) bound to helix V (Lu *et al*, 2003). The α-helix (in red) of finger 5 recognizes the dsRNA shape by non-sequence-specific contacts to the RNA sugar-phosphate backbone.

RNA recognition is achieved by a network of intermolecular hydrogen bonds between the protein main-chain functional groups and the Watson–Crick edges of the bases (Hudson *et al*, 2004). These structures reveal that zinc fingers bind to RNA differently to the way they do to DNA. The CCHH-type zinc fingers have two modes of RNA binding. First, the zinc fingers interact non-specifically with the backbone of a double helix, and second, the zinc fingers specifically recognize individual bases that bulge out of a structurally rigid element. The CCCH-type zinc fingers show a third mode of RNA binding, in which the single-stranded RNA is recognized in a sequence-specific manner. Taken together, zinc fingers represent a unique class of nucleic-acid-binding proteins that are capable of a direct readout of the DNA sequence within a DNA double helix, a direct readout of the RNA sequence within single-stranded RNA, and an indirect readout of the RNA as they recognize the shape of the RNA rather than its sequence. Of course, more structures of CCHH-type and CCCH-type zinc fingers in complex with RNA will need to be determined to generalize their mode of RNA recognition.

Conclusions

Proteins that contain RNA-binding domains and their interactions with RNA have important roles in all aspects of gene expression and regulation. The enormous diversity of interactions observed in protein–RNA complexes indicates that a simple recognition code is unlikely to exist in the world of protein–RNA interactions. However, two unifying themes may be inferred from the known complexes: the recognition of the primary RNA sequence and/or the recognition of the RNA shape by individual RBPs. In a simplistic view, the RRMs, dsRBMs and CCHH-type zinc fingers seem to be shaped to recognize single-stranded RNA, double-stranded RNA and RNA bulges, respectively. However, we have shown here by reviewing several recent protein–RNA complex structures that, the RRM and, to a lesser extent, the dsRBMs and the CCHH-type zinc fingers have evolved to recognize specifically a rich repertoire of RNAs in terms of length, sequence and structure. This is achieved in three ways: first, by the subtle amino-acid change in variable regions of the domains, namely the β2–β3 and the β1–α1 loops in the RRM, α-helix 1 in the dsRBM and the α-helix in the zinc fingers; second, by multiplication of the domains to achieve higher affinity through cooperative binding; and third, by extension of the protein domain. Although more structures still need to be determined, it might soon be possible to predict which RBP binds to which RNA, and how it recognizes its target. As a consequence, post-transcriptional gene expression and its regulation could be understood and controlled at the atomic level.

ACKNOWLEDGEMENTS

We apologize to authors whose work could not be cited due to space constraints. The authors are supported by the Swiss National Science Foundation (No. 31-67098.01), the Roche Research Fund for Biology at the ETH Zurich (F.H.-T.A.), and the European Molecular Biology Organization and the Human Frontier Science Program postdoctoral fellowships (R.S.).

REFERENCES

Allain FH, Howe PW, Neuhaus D, Varani G (1997) Structural basis of the RNA-binding specificity of human U1A protein. *EMBO J* **16:** 5764–5772

Allain FH, Bouvet P, Dieckmann T, Feigon J (2000) Molecular basis of sequence-specific recognition of pre-ribosomal RNA by nucleolin. *EMBO J* **19:** 6870–6881

Deo RC, Bonanno JB, Sonenberg N, Burley SK (1999) Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell* **98:** 835–845

Ding J, Hayashi MK, Zhang Y, Manche L, Krainer AR, Xu RM (1999) Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA. *Genes Dev* **13:** 1102–1115

Doyle M, Jantsch MF (2002) New and old roles of the double-stranded RNA-binding domain. *J Struct Biol* **140:** 147–153

Dreyfuss G, Kim VN, Kataoka N (2002) Messenger-RNA-binding proteins and the messages they carry. *Nat Rev Mol Cell Biol* **3:** 195–205

Fierro-Monti I, Mathews MB (2000) Proteins binding to duplexed RNA: one motif, multiple functions. *Trends Biochem Sci* **25:** 241–246

Finerty PJ, Bass BL (1997) A *Xenopus* zinc finger protein that specifically binds dsRNA and RNA–DNA hybrids. *J Mol Biol* **271:** 195–208

Fribourg S, Gatfield D, Izaurralde E, Conti E (2003) A novel mode of RBD-protein recognition in the Y14–Mago complex. *Nat Struct Biol* **10:** 433–439

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

Hudson BP, Martinez-Yamout MA, Dyson HJ, Wright PE (2004) Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nat Struct Mol Biol* **11:** 257–264

Johansson C, Finger LD, Trantirek L, Mueller TD, Kim S, Laird-Offringa IA, Feigon J (2004) Solution structure of the complex formed by the two N-terminal RNA-binding domains of nucleolin and a pre-rRNA target. *J Mol Biol* **337:** 799–816

Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res* **32:** D142–D144

Leulliot N, Quevillon-Cheruel S, Graille M, Van Tilbeurgh H, Leeper TC, Godin KS, Edwards TE, Sigurdsson ST, Rozenkrats N, Nagel RJ, Ares M, Varani G (2004) A new α-helical extension promotes RNA binding by the dsRBD of Rnt1p RNAse III. *EMBO J* **23:** 2468–2477

Lorsch JR (2002) RNA chaperones exist and DEAD box proteins get a life. *Cell* **109:** 797–800

Lu D, Searles MA, Klug A (2003) Crystal structure of a zinc-finger—RNA complex reveals two modes of molecular recognition. *Nature* **426:** 96–100

Mattaj IW (1993) RNA recognition: a family matter? *Cell* **73:** 837–840

Mazza C, Segref A, Mattaj IW, Cusack S (2002) Large-scale induced fit recognition of an m(7)GpppG cap analogue by the human nuclear cap-binding complex. *EMBO J* **21:** 5548–5557

Mendez-Vidal C, Wilhelm MT, Hellborg F, Qian W, Wiman KG (2002) The p53-induced mouse zinc finger protein wig-1 binds double-stranded RNA with high affinity. *Nucleic Acids Res* **30:** 1991–1996

Messias AC, Sattler M (2004) Structural basis of single-stranded RNA recognition. *Acc Chem Res* **37:** 279–287

Miller J, Mclachlan AD, Klug A (1985) Repetitive zinc-binding domains in the protein transcription factor Iiia from *Xenopus* oocytes. *EMBO J* **4:** 1609–1614

Ohman M, Kallman AM, Bass BL (2000) *In vitro* analysis of the binding of ADAR2 to the pre-mRNA encoding the GluR-B R/G site. *RNA* **6:** 687–697

Oubridge C, Ito N, Evans PR, Teo CH, Nagai K (1994) Crystal structure at 1.92 Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature* **372:** 432–438

Pavletich NP, Pabo CO (1991) Zinc finger–DNA recognition: crystal structure of a Zif68-DNA complex at 2.1 Å. *Science* **252:** 809–817

Picard B, Wegnez M (1979) Isolation of a 7s particle from *Xenopus laevis* oocytes—5s RNA–protein complex. *Proc Natl Acad Sci USA* **76:** 241–245

Price SR, Evans PR, Nagai K (1998) Crystal structure of the spliceosomal U2B″–U2A′ protein complex bound to a fragment of U2 small nuclear RNA. *Nature* **394:** 645–650

Ramos A, Grunert S, Adams J, Micklem DR, Proctor MR, Freund S, Bycroft M, St Johnston D, Varani G (2000) RNA recognition by a Staufen double-stranded RNA-binding domain. *EMBO J* **19:** 997–1009

Ryter JM, Schultz SC (1998) Molecular basis of double-stranded RNA–protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. *EMBO J* **17:** 7505–7513

Saunders LR, Barber GN (2003) The dsRNA binding protein family: critical roles, diverse cellular functions. *FASEB J* **17:** 961–983

St Johnston D, Brown NH, Gall JG, Jantsch M (1992) A conserved double-stranded RNA-binding domain. *Proc Natl Acad Sci USA* **89:** 10979–10983

Stephens OM, Haudenschild BL, Beal PA (2004) The binding selectivity of ADAR2's dsRBMs contributes to RNA-editing selectivity. *Chem Biol* **11:** 1239–1250

Theunissen O, Rudt F, Guddat U, Mentzel H, Pieler T (1992) RNA and DNA-binding zinc fingers in *Xenopus* Tfiiia. *Cell* **71:** 679–690

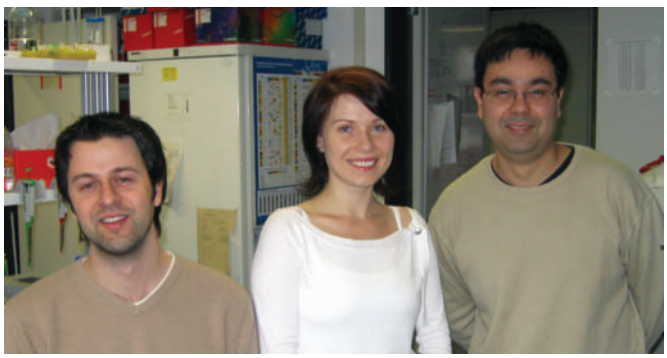
Varani L, Gunderson SI, Mattaj IW, Kay LE, Neuhaus D, Varani G (2000) The NMR structure of the 38 kDa U1A protein–PIE RNA complex reveals the basis of cooperativity in regulation of polyadenylation by human U1A protein. *Nat Struct Biol* **7:** 329–335

Venter JC *et al* (2001) The sequence of the human genome. *Science* **291:** 1304–1351

Wang XQ, Hall TMT (2001) Structural basis for recognition of AU-rich element RNA by the HuD protein. *Nat Struct Biol* **8:** 141–145

Wolfe SA, Nekludova L, Pabo CO (2000) DNA recognition by Cys(2)His(2) zinc finger proteins. *Annu Rev Biophys Biomol Struct* **29:** 183–212

Wu H, Henras A, Chanfreau G, Feigon J (2004) Structural basis for recognition of the AGNN tetraloop RNA fold by the double-stranded RNA-binding domain of Rnt1p RNase III. *Proc Natl Acad Sci USA* **101:** 8307–8312



Richard Stefl, Lenka Skrisovska & Frédéric H.-T. Allain, who is an EMBO Young Investigator

Journal of Biomolecular NMR (2005) 31: 71–72

Letter to the Editor: Resonance assignments of the double-stranded RNA-binding domains of adenosine deaminase acting on RNA 2 (ADAR2)

Richard Stefl[a,*], Lenka Skrisovska[a], Ming Xu[b], Ronald B. Emeson[b] & Frederic H.-T. Allain[a,*]

[a]*Institute of Molecular Biology and Biophysics, Swiss Federal Institute of Technology Zürich, ETH-Hönggerberg, CH-8093 Zürich, Switzerland;* [b]*Department of Pharmacology, Vanderbilt University, Nashville, TN 37232, U.S.A.*

Received 20 September 2004; Accepted 2 November 2004

Key words: NMR resonance assignment, RNA editing, RNA-binding domains

Biological context

Adenosine deaminases that act on RNA (ADARs) convert adenosine to inosine (A-to-I) by hydrolytic deamination in cellular and viral RNA transcripts containing either perfect or imperfect RNA duplexes (Bass, 2002). A-to-I editing can be either specific or non-specific, deaminating up to 50% of the adenosine residues within a perfect RNA duplex, while modifying a single or limited set of adenosine residues within imperfect double-stranded RNA (dsRNA) regions containing bulges, loops, and mismatches (Bass, 2002). The majority of non-selective editing occurs in untranslated regions (UTRs) and introns where large regular duplexes are formed. Such modifications can modulate gene silencing triggered by intramolecular structures in mRNA (Tonkin and Bass, 2003), the nuclear retention of hyperedited RNA transcripts, or participate in the antiviral response by extensive modification of viral RNAs. Selective editing has been shown to take place largely within codons, so that multiple RNA and protein isoforms can be created from a single genomic locus. For example, ADARs have been shown to produce functionally important isoforms for numerous proteins involved in synaptic neurotransmission, including ligand and voltage-gated ion channels and G-protein coupled receptors (Bass, 2002).

Like many RNA-binding proteins, ADARs display a modular domain organization, having two or three tandem copies of double-stranded RNA-binding domain (dsRBDs) at its N-terminal, and a C-terminal adenosine deaminase domain. The dsRBDs of ADARs play an important role in modulating the editing selectivity of ADARs (Carlson et al., 2003). To gain insight into this intriguing protein-RNA recognition process, we have initiated an NMR study of the two dsRBDs of rat ADAR2 (74–301) (designated dsRBD12). Here, we report ^{1}H, ^{13}C, and ^{15}N resonance assignments of the dsRBD12 of ADAR2.

*To whom correspondence should be addressed. E-mails: {allain, stefl}@mol.biol.ethz.ch

Methods and results

Initial attempts to study the 228-residue N-terminal dsRBD12 of rat ADAR2 (74–301) failed due to the low solubility of the protein in common NMR buffers. To enhance the solubility limit of dsRBD12, a non-cleavable solubility-enhancement tag, GB1 (Zhou et al., 2001), was fused to the protein. dsRBD12 was cloned into pET30-GBF*usion*1 vector yielding N-terminal GB1-tagged and C-terminal His-tagged construct of dsRBD12 (294 residues; designated GB1-dsRBD12). The protein was overexpressed in BL21(DE3) *E. coli* in 100% and 60% D_2O minimal medium (0.5 l) containing [$^{13}C_6$] glucose and $^{15}NH_4Cl$ in order to obtain a ~85% and ~50% deuterated (^{2}H), [^{13}C, ^{15}N]-labeled proteins, respectively. Cells were grown at 37 °C to OD_{600} ~1 and induced with 1 mM IPTG. The protein was purified under denaturating conditions and refolded into native buffer on Ni-NTA affinity column. The refolded GB1-dsRBD12 provides the same [^{1}H-^{15}N]-TROSY spectrum as the one purified under native conditions (data not shown). The protein was concentrated to ~0.8 mM in 250 µl of 50 mM sodium phosphate buffer (pH 8.0) containing 200 mM NaCl.

With the same protocol, N-terminal His-tagged [^{13}C, ^{15}N]-labeled dsRBD1 (74–147) and dsRBD2 (231–301)

fragments were also expressed, purified, and refolded, separately. These individual domains were important to confirm the resonance assignment of the long protein construct, because dsRBD1 and dsRBD2 share 52% sequence identity and contain repeated residues in the sequence such as leucines and lysines, which cause chemical shift degeneracy.

All the NMR experiments were conducted at 20 °C using Bruker Avance-900 and DRX-750, 600, and 500 MHz spectrometers. The following spectra were recorded on a [^{2}H(85%),^{13}C,^{15}N]-labeled GB1-dsRBD12 sample for assignment of the backbone and C_β resonances: [^{1}H,^{15}N]-TROSY (Figure 1), [^{1}H,^{15}N]-TROSY-HNCA (Salzmann et al., 1998), [^{1}H,^{15}N]-TROSY-HNCACB (Salzmann et al., 1999), 3D NOESY-[^{1}H,^{15}N,^{1}H]-TROSY (Pervushin et al., 1999). Side chain resonance assignments were achieved using a 3D NOESY-[^{1}H,^{15}N,^{1}H]-TROSY and a 3D [^{1}H,^{13}C,^{1}H]-NOESY experiments ran on a [^{2}H(50%),^{13}C,^{15}N]-labeled GB1-dsRBD12 sample. These latter assignments were made by comparison with the spectra of the individual domains where the side-chain resonances were assigned using a 3D [^{1}H,^{15}N,^{1}H]-NOESY, a 3D [^{1}H,^{13}C,^{1}H]-NOESY, and a 3D HCCH-TOCSY (Bax and Grzesiek, 1993) experiments on [^{13}C,^{15}N]-labeled dsRBD1 and dsRBD2 samples. Aromatic side chains were assigned using 2D [^{1}H,^{1}H]-TOCSY, 2D [^{1}H,^{1}H]-NOESY.

Extent of assignment

The [^{1}H,^{15}N]-TROSY spectrum of the [^{2}H(85%),^{13}C,^{15}N]-labeled GB1-dsRBD12 is shown in Figure 1 with indicated assignments. With the combined information from all heteronuclear experiments on the three different protein constructs, 99% of the backbone amide protons and ^{15}N nuclei (prolines not counted), and 99% of the C_α were assigned for the three structured domains of GB1-dsRBD12. Assignments could also be made for 35% of the 82-residue linker (133-215), mostly for amino acids flanking prolines. The chemical shift of the latter residues (near random coil values) and the absence of NOEs indicate that this region is highly flexible and is likely to be unstructured. In total, ~75% of the backbone were assigned of the entire protein sequence (1–294). This represents almost 100% of the resonances which appeared in the TROSY-type experiments. The assignment of side chain aliphatic ^{1}H and ^{13}C resonances is approximately 95% complete for the three structured domains, the remaining portion could not have been assigned due to resonance overlap. Aromatic proton assignment is 90% complete for the structured part of GB1-dsRBD12. ^{1}H chemical shifts

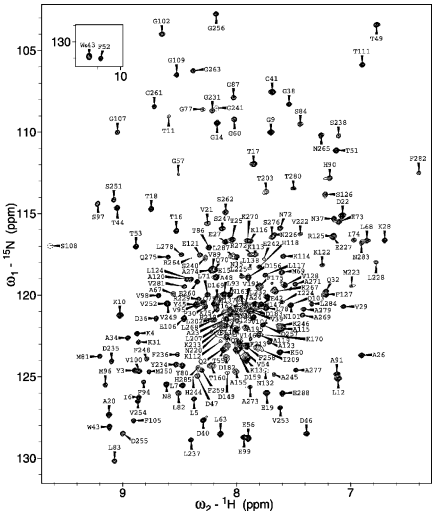


Figure 1. [^{1}H-^{15}N]-TROSY spectrum of ~0.8 mM [^{2}H(85%),^{13}C,^{15}N]-labeled GB1-dsRBD12 (32 kDa) in 50 mM sodium phosphate buffer (pH 8.0), 200 mM NaCl, acquired at 293 K on a Bruker DRX-600 MHz spectrometer. The assignments are labeled by the one-letter code of amino acids accompanied by a sequence number. S108, N132, and G241 are only visible at a lower contour level and their positions are indicated by the dashed circles.

are referenced to internal 2,2-dimethyl- silapentane-5-sulfonic acid (DSS), whereas ^{13}C and ^{15}N chemical shifts are referenced indirectly to DSS, using absolute frequency ratios. Chemical shifts are deposited in the BioMagResBank Database (accession code BMRB-6280).

Acknowledgements

The work was supported by Swiss National Science Foundation and the Roche Research Fund for Biology at the ETH Zurich (to FHTA), by the National Institutes of Health (to RBE) and by European Molecular Biology Organization and Human Frontier Science Program postdoctoral fellowships to RS.

References

Bass, B.L. (2002) *Annu. Rev. Biochem.*, **71**, 817–846.
Bax, A. and Grzesiek, S. (1993) *Acc. Chem. Res.*, **26**, 131–138.
Carlson, C.B., Stephens, O.M. and Beal, P.A. (2003) *Biopolymers*, **70**, 86–102.
Pervushin, K.V., Wider, G., Riek, R. and Wuthrich, K. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 9607–9612.
Salzmann, M., Pervushin, K., Wider, G., et al. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 13585–13590.
Salzmann, M., Wider, G., Pervushin, K., et al. (1999) *J. Am. Chem. Soc.*, **121**, 844–848.
Tonkin, L.A. and Bass, B.L. (2003) *Science*, **302**, 1725.
Zhou, P., Lugovskoy, A.A. and Wagner, G. (2001) *J. Biomol. NMR*, **20**, 11–14.