Published in final edited form as:

Adv Exp Med Biol. 2010 ; 693: 37–53.

Insights into the structural basis of RNA recognition by STAR domain proteins

Sean P. Ryder and Francesca Massi

Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 364 Plantation Street, LRB-906, Worcester, MA, 01605, Phone: 508-856-1372, Fax: 508-856-6464

Abstract

STAR proteins regulate diverse cellular processes and control numerous developmental events. They function at the post-transcriptional level by regulating the stability, sub-cellular distribution, alternative splicing, or translational efficiency of specific mRNA targets. Significant effort has been expended to define the determinants of RNA recognition by STAR proteins, in hopes of identifying new mRNA targets that contribute their role in cellular metabolism and development. This work has led to the extensive biochemical characterization of the nucleotide sequence specificity of a handful of STAR proteins. In contrast, little structural information is available to analyze the molecular basis of sequence specific RNA recognition by this protein family. This chapter reviews the relevant literature on STAR domain protein structure, and provides insights into how these proteins discriminate between different RNA sequences.

Keywords

Post-transcriptional Regulation; RNA-binding protein; NMR spectroscopy; protein-RNA interactions; thermodynamics; consensus sequence; dynamics; conservation; gel mobility shift; SELEX

Introduction

RNA-binding proteins play fundamental roles in cellular physiology. They guide decoding of the genome, they comprise the machinery that synthesizes proteins, and they regulate the intensity, duration, and sub-cellular distribution of gene expression.^{1–6} Most RNA-binding proteins must distinguish between specific RNA sequences within the cellular milieu in order to function. The thermodynamic, kinetic, and structural basis for sequence discrimination by RNA-binding proteins defines the mechanism of RNA recognition, which yields insights into specificity and biological function.

Sequence specific RNA-binding proteins employ one of three methods to discriminate between their RNA targets. Some recognize linear RNA sequences through base specific interactions, including hydrogen bonds and shape-specific van der Waals interactions.^{7–10}

Specificity arises through decoding the chemical differences between the four bases. Others recognize a specific three-dimensional structure, including stem-loops, pseudoknots, or helical junctions.^{11–14} Here, specificity derives from the ability of a set of RNA sequences to fold into a recognizable shape. Lastly, some RNA-binding proteins bind to a protein-RNA complex, and only achieve specificity through a combination of protein-protein and protein-RNA interactions.^{15–17}

In this review, we focus on RNA binding by the STAR (signal transduction and activation of RNA) domain family of RNA-binding proteins (Fig. 1).¹⁸ Genes that encode STAR proteins are found in the genomes of all metazoan species and have recently been identified in plants.¹⁹ Highly studied examples include *Caenorhabditis elegans* *GLD-1*, *Drosophila melanogaster* *HOW*, and their vertebrate homologs *Quaking (QKI)* and *Sam68*.^{20–30} STAR proteins play a key role in developmental processes and have been implicated in human disease.^{31,32} They couple cellular signaling events to post-transcriptional regulation of gene expression.

The STAR domain

The STAR domain is a region of extended conservation flanking a canonical maxi-KH RNA binding domain (Fig. 1).¹⁸ The region N-terminal to the maxi-KH domain is termed the QUA1 motif, and the region C-terminal to the maxi-KH domain is termed the QUA2 motif. These regions are named after the mouse STAR domain protein Quaking (QKI), and define the difference between the STAR domain and other KH domain RNA-binding proteins.

Several lines of evidence indicate that the QUA1 domain is a homodimerization motif critical for STAR protein function.^{33–35} First, in situ crosslinking studies in cell culture demonstrate that the STAR protein QKI6 self-associates.^{33,34} Mutagenesis experiments reveal that self-association requires the QUA1 region. Moreover, immunoprecipitation experiments from mixed lysates of HeLa cells transfected with either myc-tagged or HA-tagged QKI show the two variants interact with each other, and epitope-tagged variants are retained on a GST column when incubated with GST-QKI expressed in bacteria. A single point mutation within the QUA1 region of QKI eliminates dimerization both in vitro and in cell culture, and causes an embryonic lethal phenotype in mice.³³ Subsequent experiments performed using recombinant variants of the *C. elegans* STAR protein (GLD-1) confirm this finding, and demonstrate that the QUA1 domain is sufficient for stable dimerization in vitro.³⁶ Together, the data show that the QUA1 domain is both necessary and sufficient for dimerization and that dimerization is crucial for function in cells.

In contrast, the KH and QUA2 domains form an extended RNA-binding interface. Numerous qualitative and quantitative assays including UV crosslinking, column retention, gel mobility shift, and fluorescence polarization (FP) experiments demonstrate that the KH and QUA2 domains bind to short penta- or hexanucleotide consensus sequences with moderate to high affinity.^{27,36–42} The KH and QUA2 domains are sufficient for RNA-binding activity.³⁶ Dimerization improves affinity, likely mediated by direct interactions with RNA from both subunits of the dimer, although exactly how dimerization influences binding specificity remains unresolved. ENU-induced point mutations within the KH and

QUA2 domains of QKI yield an embryonic lethal phenotype, demonstrating that both regions are required for function.⁴³ Similarly, mutations within the KH and QUA2 domain of GLD-1 have significant pleiotropic effects on germline development.²² The data demonstrate that the activity of the RNA-binding subunits of STAR proteins is required for their biological function. In the next section, we review the quantitative data that defines the nucleotide sequence specificity of RNA recognition by GLD-1 and QKI, highlighting similarities and contrasting differences.

RNA recognition by STAR proteins

The nucleotide sequence specificity of four STAR domain proteins (GLD-1, QKI, HOW, and Sam68) has been investigated in detail using quantitative in vitro methods. Two approaches have proven useful. In the first, the binding specificity is determined through identification of a regulatory target and characterization of the minimal binding sequence, followed by comprehensive mutagenesis.^{36,44} In the second, the binding specificity is determined from a randomized sequence library using systematic evolution of ligands through exponential enrichment (SELEX) followed by computational comparison of the “winner” sequences to identify similarities.^{27,38,45} Both methods yield comparable results, outlined below, and indicate that GLD-1, QKI, and HOW bind to RNA with similar though not identical specificity, while Sam68 binds to a different sequence.

Recognition of RNA by GLD-1

GLD-1 regulates the switch from spermatogenesis to oogenesis in *C. elegans* hermaphrodite development by controlling the expression of *tra-2*, a bifunctional membrane protein and transcription factor required for promoting oocyte cell fate.^{42,46–48} Gel mobility shift and yeast 3-hybrid experiments reveal that GLD-1 binds to a repeat element in the 3'-UTR of the *tra-2* mRNA termed the TGE, for *tra-2* and *gli-1* element, with very high affinity (K_d , app ~ 10 nM) and a 2:1 apparent protein to RNA stoichiometry.^{35,42} The TGE is 28 nucleotides in length and is primarily comprised of uridines and adenosines. Gel shift experiments with a battery of mutant variants across the TGE identify a bipartite recognition element that includes a UA dinucleotide near the 5'-end of the sequence and a contiguous UACUCA hexanucleotide element near the 3'-end of the sequence.³⁵ Because GLD-1 is a homodimer and because the recognition element is bipartite, it was proposed that one subunit recognizes the hexanucleotide element while the second subunit recognizes the UA dinucleotide as a partial or incomplete version of the hexanucleotide binding site.

Consistent with this hypothesis, a 12-nucleotide element comprised of the UACUCA hexanucleotide element flanked by three additional nucleotides on either side efficiently competes with the full-length TGE in competition gel shift experiments.³⁵ Moreover, isothermal titration calorimetry experiments demonstrate that the 12-nucleotide RNA binds to GLD-1 with an apparent 1:1 protein to RNA stoichiometry. Thus, each subunit of the dimer is capable of binding to an identical copy of the hexanucleotide sequence. Because there are three copies of the UACUCA hexanucleotide within the region of the *tra-2* UTR that contain the TGE repeats, it is not clear if the upstream UA dinucleotide is relevant to

binding in worms, or if its apparent contribution to binding is an artifact of the minimal in vitro system.

To delineate the consensus GLD-1 binding sequence, a comprehensive library of single nucleotide mutations of the UACUCA sequence was analyzed within the context of the 12-nucleotide RNA.³⁵ Competition gel shifts were performed to determine the IC₅₀ of the mutant sequence relative to the wild-type 12-mer RNA. The consensus recognition sequence, termed the STAR binding element (SBE), is 5'-UACU(C/A)A-3' (Table 1). Only the C to A mutation at the fifth position is tolerated without a reduction in competition efficiency. Allowing for mutations that reduce binding by up to 10-fold, a more relaxed consensus of 5'-(U>G>A/C)A(C>A)U(C/A>U)A-3' was also proposed. To date, the relative affinity and number of binding sites required for regulation have not been assessed in any functional assay in worms. Thus, it is not clear which consensus is more relevant to GLD-1 regulatory activity in worms, or if additional requirements beyond the determinants of binding in vitro are needed to select targets for regulation.

Recognition of RNA by QKI

QKI was identified as the gene responsible for the phenotype observed in the *Quaking* mouse (*Qk^v*), a spontaneous mutant that arose in a mouse colony over forty years ago.²⁰ These mice fail to form compact myelin in their central nervous system, leading to a characteristic tremor upon movement. The *Qk^v* allele is a large 1 MB deletion of chromosome 17 that modifies the expression pattern of the adjacent *Quaking* locus.^{26,49,50} The dysmyelination phenotype is due to the reduction of QKI expression in the oligodendrocyte lineage. There are three predominant isoforms of QKI, termed QKI5, QKI6, and QKI7 after their respective transcript lengths.⁵⁰ QKI5 is nuclear, while QKI6 and QKI7 are cytoplasmic.⁵¹ All variants share the STAR domain and differ only in their C-terminus and 3'-UTR sequence. Several functions are proposed for QKI, including the regulation of mRNA stability, translation efficiency, and alternative splicing.^{41,52–55} Abundant evidence demonstrates that a major role of QKI in the oligodendrocyte lineage is in regulating the translation and stability of myelin basic protein (MBP) mRNA, which encodes a major myelin structural protein.^{41,52} Reporter experiments and qualitative binding experiments indicate that QKI regulates MBP expression through specific association with its 3'-UTR.⁴¹

Due to the high sequence similarity of QKI and GLD-1 within the STAR domain, it was predicted that these factors would bind to RNA with similar specificity. Indeed, QKI is capable of binding to TGE RNA in vitro, and QKI-6 can functionally substitute for GLD-1 in a reporter assay in worms.^{39,56} To directly measure the specificity of QKI, competition binding experiments were performed using an FP assay and the same 12-mer library used to map the GLD-1 consensus sequence.³⁹ The data reveal that QKI binds to RNA with similar, though not identical, specificity as GLD-1. The consensus sequence, termed the QKI STAR binding element (QSBE), is 5'-NA(A/C)UAA-3' (Table 1). This sequence is similar to the SBE recognized by GLD-1, but not identical. One difference is that the identity of the first nucleotide of the hexanucleotide consensus is not specified. A second difference is that an adenosine is permitted at the third position, where GLD-1 recognizes only a cytidine. The final, and most dramatic difference is that an adenosine substitution at the fifth position of

the consensus leads to a 40-fold enhancement of binding, whereas the same substitution in GLD-1 leads to a modest 2-fold increase. Several QKI consensus sites are present in the 3'-UTR of myelin basic protein transcripts.³⁹ QKI binds to all of these in vitro, but it binds with highest affinity (~10 nM) to a sequence present within a previously characterized region required for silencing MBP translation during localization.⁵⁷ The data suggest, but do not prove, that QKI regulates MBP mRNA through specific association with this element in its 3'-UTR. Functional studies are required to demonstrate that these binding sites are true *cis*-regulatory elements that confer QKI-dependent regulation of MBP expression in cells.

In an independent study, the binding specificity of QKI was determined using an in vitro SELEX protocol.³⁸ Several rounds of selection, enrichment, and amplification lead to the identification of a number of aptamer sequences, many of which contained separate “core” 5'-NACUAAY-3' and “half site” 5'-YAAAY-3' motifs with variable spacing (Table 1). Limited mutagenesis studies indicate that both elements are required for QKI to associate with the selected aptamers by gel mobility shift. There are two interesting differences between these results and the previous mutagenesis studies. First, the data suggest that an additional pyrimidine nucleotide is specified after the sixth position in the consensus. This requirement was not explicitly tested in previous binding experiments. Second, the selected consensus suggests a stricter requirement for a C at the third position, while the mutagenesis experiments indicate that a C or A is tolerated. Finally, the data suggest that QKI, like GLD-1, recognizes a bipartite element. It remains to be seen if either element is required for binding and function in vivo, or if both subunits of the QKI dimer can recognize either two copies of the “core” or two copies of the “half-site” element. Notably, the high affinity QKI binding site in the MBP 3'-UTR lacks a separate 5'-YAAAY-3' half-site.³⁹ As with GLD-1, functional experiments are needed to identify the minimal requirements for QKI-dependent regulation in cells.

RNA recognition by other STAR domain proteins

A similar SELEX strategy was applied to Sam68, a STAR domain protein that is phosphorylated by Src and implicated in several aspects of cellular signaling and post-transcriptional regulation of gene expression.^{24,25,27,58} The results suggest that Sam68 binds with high affinity to RNA sequences that contain the four nucleotide element 5'-UAAA-3' (Table 1).²⁷ Mutation of the element to 5'-UACA-3' eliminates binding in vitro, as does mutation of the STAR domain of the protein. Because re-selection or comprehensive mutagenesis was not performed, the results do not comprehensively define the Sam68 consensus binding sequence. However, they clearly demonstrate that this protein binds to an element that is different than GLD-1 and QKI binding sites. The binding determinant is four nucleotides in length instead of five or six. Second, the site is more purine-rich than the sites recognized by the other proteins, including a run of three adenosine nucleotides. These results were recently confirmed in an independent selection, and extended to another STAR protein Sam68-like mammalian protein 2 (SLM2, also known as Khdrbs3), which binds the same sequence.⁴⁵ SLM2 shows a greater degree of sequence similarity to Sam68 than to GLD-1, QKI, or other STAR proteins, and as such is expected to bind RNA with specificity similar to Sam68.

The binding specificity of the *Drosophila* STAR domain protein HOW has also been investigated.³⁷ HOW is a post-transcriptional regulator of *stripe* and other transcripts required for wing tendon development.^{30,40} The binding specificity of HOW was determined by mapping its binding sites in the 3'-UTR of *stripe* transcripts using biotinylated RNA fragments and streptavidin resin in HOW pull down assays.³⁷ HOW binds to RNA with almost identical specificity to QKI, recognizing the sequence 5'-NA(C>A)UAA-3' (Table 1). Intriguingly, HOW can bind to this sequence when it comprises the loop of a stem-loop structure, if the length of the loop is at least 12 nucleotides. Though not quantitative, an approximation of the relative affinity based upon relative signal intensity suggests that HOW binds more tightly to stem-loop RNA sequences than to unstructured RNAs. This result is in contrast to findings with QKI, where stem-loop sequences inhibit binding.³⁸ If RNA structure context is needed for HOW binding specificity in the functional context, then mRNA target specificity could arise from a combination of sequence and structure-based recognition, unique among the STAR domain proteins.

STAR domain structure

What is the molecular basis for RNA-recognition by STAR domain proteins, and what accounts for the differences in specificity between members of this family? To date, there are no structural data for an intact STAR domain from any protein. However, a pair of partial structures begins to reveal the three-dimensional architecture of this domain and the structural basis for sequence-specific RNA recognition.^{8,59} The next few sections will review the structures, including an NMR structure of the KH-QUA2 domain of Splicing Factor 1 (SF1, Fig. 1)—the mammalian branch point binding protein, and the NMR structure of the KH-QUA2 domain of QKI. A comparison of structure-based sequence alignments, homology modeling, and protein mutagenesis experiments provide a starting point to understand how RNA binding specificity is achieved.

The NMR structure of SF1 bound to RNA

SF1 is a component of the eukaryotic splicing apparatus that is critical for recognition of the branch site adenosine in introns.^{60–62} SF1 is the mammalian homolog of the yeast branch point binding protein (BBP), which recognizes the branch site adenosine within the context of the branch point sequence (BPS) 5'-UACUAAC-3'.⁶² SF1 contains recognizable maxi-KH and QUA2 domain, but lacks the QUA1 domain typical of STAR proteins (Fig. 1).⁸ Instead, it binds to RNA cooperatively with U2AF35/65, which recognizes the polypyrimidine tract and 3'-splice site to form the initial intron recognition complex.¹⁷ Intriguingly, the BPS sequence lies within the binding consensus sequence recognized by GLD-1, QKI, and HOW.^{36–39} The structure of mammalian SF1 bound to the yeast BPS has been determined by nuclear magnetic resonance spectroscopy (NMR), providing the first glimpse into the structural basis of RNA recognition by the KH and QUA2 domains.⁸ In the following section, we provide an overview of the NMR structure of SF1, outlining the amino acids that comprise the RNA-binding interface.

Overview of the structure

The three dimensional structure of SF1 KH-QUA2 region bound to an 11-nucleotide RNA containing the yeast BPS (5'-UAUACUAACAA-3') has been determined using NMR spectroscopy.⁸ The high quality of the experimental data, including the large number of intermolecular restraints, enables the high-resolution structural characterization of the SF1/RNA complex (Fig. 2a). The most interesting feature of the structure is the relative organization of the KH and QUA2 domains. QUA2 forms an alpha helix that packs against the maxi-KH domain, forming an expanded structure with topology $\beta 1-\alpha 1-\alpha 2-\beta 2-\beta 3-\alpha 3-\alpha 4$. This organization forms an extended hydrophobic surface between the two domains that comprises part of the RNA binding groove. The RNA molecule is bound in an extended conformation onto a large RNA binding surface that includes helix $\alpha 4$ of QUA2 and the following elements of KH: helices $\alpha 1$ and $\alpha 2$, strand $\beta 2$, the conserved GXXG loop (situated between $\alpha 1$ and $\alpha 2$), and the variable loop (also known as the “thumb” region, between $\beta 2$ and $\beta 3$) that encloses the RNA. The QUA2 domain is involved in the recognition of the nucleotides at the 5' end of the RNA (A4-C5), while the KH domain is critical for the recognition of the 3' end of the BPS RNA (U6-A7-A8-C9). The SF-1/RNA complex is stabilized by a combination of extensive hydrophobic interactions with aliphatic side chains, base-specific and backbone hydrogen bonds, as well as electrostatic interactions. A peculiar feature of SF1, relative to other single-stranded RNA binding proteins, is the absence of aromatic-base stacking interactions between the protein and the RNA.^{7,9}

Recognition of RNA by SF1

Analysis of the NMR secondary chemical shifts collected in the absence of RNA indicates that the overall structure of SF1 is largely maintained in the free state.⁸ Thus, it is likely that SF1 associates with BPS RNA through a pre-formed RNA interface, rather than through large structural rearrangements induced by association of the RNA. This does not preclude local structural re-arrangements necessary to form the stable interactions observed in the structure calculations. The detailed interactions responsible for sequence-specific recognition of BPS RNA by SF1 are described below.

Residues in the QUA2 domain of SF1 recognize A4 and C5. In particular, A4 is recognized through hydrophobic interactions between the base and the side chains of Arg 255 and Ala 248. The shape of the pocket formed by these amino acids mediates specificity for an adenosine residue. The base and sugar of C5 form hydrophobic interactions with the side chains of Leu 244 and of Leu 247 of QUA2, respectively. Specificity for a cytidine is provided in the form of a hydrogen bond between the side chain amide of Asn 151 and the N5 of C5.

U6 binds to the interface between the QUA2 and KH domains. This position is stabilized through van der Waals interactions between the base and the side chains of Leu 244 and of Leu 247, located in the QUA2 domain, as well as by hydrogen bonds with the backbone of Leu 155 and Gly158 of the KH domain (Fig. 3a). The sugar of U6 is stabilized by a hydrogen bond with the backbone amide of Arg 160 and is stacked against the side chain of Pro 159, in the conserved GPRG loop. This nucleotide is packed against the backbone atoms of Gly 154, explaining the conservation of a glycine residue at this position. The steric

constraints and the specific hydrogen bond interactions formed by the U6 recognition pocket between the KH and QUA2 domains of SF1 explain the strong selectivity of this position.

A7 interacts exclusively with residues located in the KH domain. The conformation of A7 in the complex is stabilized by a π -cation interaction with the sidechain of a highly conserved lysine (Lys 184) located in the variable loop (Fig 3b). In addition, a hydrogen bond between N6 of A7 and the sidechain of Glu 159 and van der Waals interaction with Val 153 contribute to the recognition of this base in the protein-RNA complex.

The branch point adenosine A8 is bound through specific hydrogen bonds with the protein backbone amide and carbonyl groups of Ile 177 (Fig. 3c). An additional hydrogen bond between N7 of A8 and the 2' hydroxyl of A7 contributes to the unique recognition of an adenine base at this position. A series of conserved aliphatic residues of the KH domain, Ile 157, Leu 164, Ile 175, Ile 177 and Val 183, form a hydrophobic envelope that surrounds A8. The close proximity of the phosphate groups of A7 and A8 to the backbone of the conserved GPRG loop is consistent with the abolished or severely diminished RNA binding affinity associated with mutations of these residues to negatively charged glutamate or aspartate.

C9 recognition is achieved through hydrophobic interactions of the pyrimidine base with the side chain of Val 183, and of the ribose ring with the side chain of Met 176. Compared to other positions, the identity of the C residue at this position is not well specified by the nature of the interactions. A8, C9, and A10 are stacked in an A-helix-like conformation, thus base stacking interactions stabilize the nucleotides in this conformation.

The NMR structure of the KH and QUA2 regions of QKI

The structure of the KH-QUA2 region of the *Xenopus laevis* Quaking protein (XQUA) has been solved using NMR spectroscopic methods.⁵⁹ In contrast to the SF1 structure described above, the QKI structure was obtained in the absence of RNA. The ensemble of calculated structures indicate that the KH domain of the protein is well defined, with the exception of the variable loop which lacks long-range restraints. The QUA2 domain clearly presents an α -helical secondary structure, but its orientation relative to the KH domain is undefined as no long-range distance restraints between the two domains were identified. A representative structure of the ensemble is shown in Figure 2b. Apart from the difference in defined orientation between the QUA2 domains of SF1 and *Xenopus* QKI, a major variation between the two structures is between the variable loops, colored in orange. In QKI, this region is α -helical, and more precisely organized into two α -helices. As a consequence, the KH domain is characterized by the following topology: $\beta 1$ - $\alpha 1$ - $\alpha 2$ - $\beta 2$ - $\alpha 3$ - $\alpha 4$ - $\beta 3$ - $\alpha 5$.

Characterization of the backbone dynamics of QKI through measurement of the ¹⁵N relaxation rates demonstrates that the KH domain is well structured in solution in its entirety, including the variable loop region. Thus, the heterogeneity of the structures observed in this region is entirely due to the paucity of long-range distance restraints. The structure of this region is well defined internally, but its orientation relative to the rest of the KH domain is less clear. In agreement with the NMR studies of SF1 that indicate that the QUA2 region is α -helical in the absence of RNA, this study of QKI in the free state supports the presence of an α -helix (residues 189–201) in the QUA2 region.^{8,59} ¹⁵N relaxation studies indicate that

the QUA2 domain of the protein is more dynamic than the KH domain. The heterogeneity of the ensemble of structures in the QUA2 domain is due to the presence of a highly flexible linker connecting the α -helix region of QUA2 and the KH domain. Hence, the QUA2 α -helix is not docked against the KH domain, as observed in the bound structure of SF1, but instead is mobile in solution.

Figure 2c shows a comparison of the QKI structure to SF-1/BPS RNA complex structure, where the RNA has been removed for clarity.^{8,59} The major differences lie within the orientation of the GXXG loop and the presence of two α -helical elements in the variable loop. In models where the RNA from the SF1 structure is docked into the QKI structure without structural refinement, the topology of the QKI GXXG loop would cause steric clash with U6 and A7. Moreover, α 4 within the QKI variable loop would clash with A11, although the differences in the amino acids interacting with this base and biochemical evidence that the nucleotide identity is not specified at this position in QKI suggest differences in RNA recognition at this position. It is likely that such dissimilarities between the structures of the two proteins arise from small differences between the conformations of the free and bound states. The highly dynamical character of QUA2 observed for QKI in the free state support a mechanism where the KH domain of the STAR domain proteins is the first to interact with the RNA, followed by the docking of the QUA2 domain helix against the KH domain. Kinetic studies are needed to tease apart the contribution of each region towards the rate of RNA binding.

Conservation of RNA contact residues between SF1 and STAR proteins

Structure based homology modeling reveals a high degree of conservation in the RNA contact residues between SF1, GLD-1, QKI, and HOW (Fig. 4).^{63,64} The GLD-1, QKI, and HOW residues predicted to contact RNA in the structure-based alignments are 100% identical with each other.³⁶ Of these, eighteen out of twenty-five (72%) are identical to those in the SF1 structure. All amino acids in the QUA2 region are conserved, suggesting that GLD-1, QKI, and HOW recognize nucleotides corresponding to A4 and C5 in an identical fashion. Four out of five amino acids that comprise the U6 binding pocket are conserved, with the lone exception being a substitution of arginine for Leu 155. While at first glance the leucine to arginine substitution would seem a non-conservative substitution, the aliphatic side-chain of arginine can form the same hydrophobic stacking interactions as leucine with the π -orbitals of U6, consistent with experimental evidence revealing that a uridine residue is strongly specified at this position in GLD-1, QKI, and HOW. Likewise, a single conservative substitution is present in the A9 binding pocket, where an aspartic acid replaces the glutamate (Glu 149) in SF1. The identity of the amino acids that recognize A8 and C9 are somewhat less well conserved, where Val 183 is replaced with a methionine, Leu 164 is replaced with an alanine, and Ile 157 is replaced with a leucine. It should be noted, however, that all of these substitutions are conservative in nature, replacing a hydrophobic residue with another, and Ile 177, which forms a specific hydrogen bond with A10, is conserved in GLD-1, QKI, and HOW. Biochemical data reveals that GLD-1, QKI, and HOW specify an adenosine residue at a similar position within their consensus binding sites (Table 1).^{36–39}

Because of the high degree of sequence identity in the RNA contact residues, and because the limited substitutions between species do not affect binding specificity experimentally, it is not surprising that all four proteins recognize similar sequence determinants.^{36–39,62} The minor variances in binding specificity between GLD-1, QKI, and HOW are not explained by this modeling exercise, as the RNA contact amino acids are 100% identical between all three proteins. Future structural and biochemical analyses will be needed to dissect the structural basis for the differences in RNA recognition by these proteins.

In contrast, there is significant variance between the SF1 RNA contact residues and their counterparts in Sam68.^{36,65} Only 40% are identical (10/25), with non-conservative substitutions in amino acids that make base-specific contacts. Only two of the five amino acids in QUA2 that contact A4 and C5 are conserved. Notably, Arg 255, which defines the shape of the A4 binding pocket, is a glutamate in Sam68 (Glu 279). Glutamate is significantly smaller than arginine, and as such cannot form the same contacts. Moreover, the glutamate for arginine substitution causes a charge reversal that could have a significant impact on the architecture of the pocket and electrostatic environment needed to recognize polyanionic RNA.

Three of the five amino acids that recognize U8 are conserved, with two substitutions: Leu 155 is replaced with a lysine, and Arg 160 is replaced with a glutamine. Lysine can form similar hydrophobic π -orbital stacking interactions as leucine in SF1, and arginine in GLD-1, QKI, and HOW. Thus this substitution is not expected to significantly compromise U6 recognition. Arg 160 forms a protein backbone to RNA backbone interaction, thus the amino acid side chain does not directly contribute to nucleotide sequence discrimination. The contact amino acids in Sam68 remain consistent with specification of a uridine nucleotide.

A7 is recognized by a base specific hydrogen bond between the N6 exocyclic amine and a glutamate side chain (Glu 149). In Sam68, the equivalent residue is a lysine (Lys 169). This amino acid cannot form the same contact, thus, adenosine specification cannot occur through a similar mechanism. Only one of the four amino acids that form base specific interactions with A8 and C9 are conserved in Sam68. The substitutions are relatively conservative changes from one small hydrophobic amino acid to another. It is not clear how these changes affect specificity.

Because the QUA2 domain is poorly conserved in Sam68 relative to SF1, GLD-1, QKI, and HOW, it is possible that this region does not contact RNA at all. SELEX data suggests that Sam68 binds to a four nucleotide specificity determinant (5'-UAAA-3') as opposed to the five or six nucleotide elements recognized by SF1, GLD-1, QKI, and HOW. Also, two basic amino acids within the QUA2 domain (R240 and K241) that have been experimentally demonstrated to contribute to binding affinity and specificity without directly contacting RNA are not conserved in Sam68, but are in GLD-1, QKI, and HOW.⁶⁶ Together, the evidence suggests that the Sam68 QUA2 domain diverges significantly from the others, and as such its role in RNA recognition is not clear.

In contrast, the U6 binding pocket in Sam68 is similar to the other STAR proteins. We suggest that the first U of the Sam68 recognition sequence is equivalent to U6 in the BPS RNA, while the second and third adenosines are equivalent to A7 and A8. Thus, the KH domain alone could drive specific RNA recognition by Sam68. However, because of the amino acid differences described above, specification of A7 must occur through a different mechanism, and it is not at all clear how C9 can be replaced by an adenosine due to and fit within the binding pocket. The structural basis of specific RNA recognition by Sam68 and its variants remains an important question that merits continued investigation.

Concluding remarks

STAR domain proteins bind to short linear RNA sequences with defined sequence specificity. While much biochemical work has been done to address STAR protein binding specificity, little is known about the mechanism by which sequence specificity is achieved. Some insight can be obtained through homology modeling of the KH and QUA2 regions of the STAR domain using the NMR structure of SF1 as a guide. The high level of conservation of the RNA contact residues between SF1, GLD-1, QKI, and HOW, and the relative similarity of their consensus binding sites, greatly facilitate this effort. Such comparisons fall short of explaining the basis for the remaining differences in affinity and specificity, and do not provide a compelling basis for the mechanism of RNA recognition by Sam68 and related proteins, which falls into a different specificity class. There is much left to be understood about RNA recognition by STAR proteins, and as such additional structures of STAR-RNA complexes are needed.

A pressing issue is the role of dimerization in defining binding specificity. STAR proteins form stable dimers, mediated by the QUA1 domain, and as such contain two subunits capable of associating with RNA independently. Do both subunits bind to the same sequence? If so, regulatory targets might be expected to have two copies of the consensus binding site. Or does each subunit bind to RNA with different specificity? If so, a bipartite recognition element comprised of different sequences, as observed by SELEX for QKI, may in fact represent the required element *in vivo*. The shape of the dimer and the path of the RNA through the dimer will likely influence the energetics of RNA recognition. Resolution of these issues awaits functional studies and the determination of the high resolution structure of a ternary complex between a STAR domain dimer and its cognate RNA target sequence.

Acknowledgments

The authors thank Ruth Zearfoss and John Pagano for critical comments on the manuscript. S.P.R. is supported by NIH R01 GM081422, NIH R21 NS059380, and a Basil O'Connor Starter Scholar Award from the March of Dimes.

References

1. Farley BM, Ryder SP. Regulation of maternal mRNAs in early development. *Crit Rev Biochem Mol Biol.* 2008; 43:135–162. [PubMed: 18365862]
2. Moore MJ. From birth to death: the complex lives of eukaryotic mRNAs. *Science.* 2005; 309:1514–1518. [PubMed: 16141059]

3. Keene JD, Lager PJ. Post-transcriptional operons and regulons co-ordinating gene expression. *Chromosome Res.* 2005; 13:327–337. [PubMed: 15868425]
4. Verdel A, Jia S, Gerber S, et al. RNAi-mediated targeting of heterochromatin by the RITS complex. *Science.* 2004; 303:672–676. [PubMed: 14704433]
5. Green R, Noller HF. Ribosomes and translation. *Annu Rev Biochem.* 1997; 66:679–716. [PubMed: 9242921]
6. Spirin AS. “Masked” forms of mRNA. *Curr Top Dev Biol.* 1966; 1:1–38. [PubMed: 4943434]
7. Hudson BP, Martinez-Yamout MA, Dyson HJ, et al. Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nat Struct Mol Biol.* 2004; 11:257–264. [PubMed: 14981510]
8. Liu Z, Luyten I, Bottomley MJ, et al. Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science.* 2001; 294:1098–1102. [PubMed: 11691992]
9. Wang X, Zamore PD, Hall TM. Crystal structure of a Pumilio homology domain. *Mol Cell.* 2001; 7:855–865. [PubMed: 11336708]
10. Edwards TA, Pyle SE, Wharton RP, et al. Structure of Pumilio reveals similarity between RNA and peptide binding motifs. *Cell.* 2001; 105:281–289. [PubMed: 11336677]
11. Chao JA, Patskovsky Y, Almo SC, et al. Structural basis for the coevolution of a viral RNA-protein complex. *Nat Struct Mol Biol.* 2008; 15:103–105. [PubMed: 18066080]
12. Vargason JM, Szittyi G, Burgyan J, et al. Size selective recognition of siRNA by an RNA silencing suppressor. *Cell.* 2003; 115:799–811. [PubMed: 14697199]
13. Valegard K, Murray JB, Stockley PG, et al. Crystal structure of an RNA bacteriophage coat protein-operator complex. *Nature.* 1994; 371:623–626. [PubMed: 7523953]
14. Oubridge C, Ito N, Evans PR, et al. Crystal structure at 1.92 Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature.* 1994; 372:432–438. [PubMed: 7984237]
15. Agalarov SC, Sridhar Prasad G, Funke PM, et al. Structure of the S15,S6,S18-rRNA complex: assembly of the 30S ribosome central domain. *Science.* 2000; 288:107–113. [PubMed: 10753109]
16. Sonoda J, Wharton RP. Recruitment of Nanos to hunchback mRNA by Pumilio. *Genes Dev.* 1999; 13:2704–2712. [PubMed: 10541556]
17. Berglund JA, Abovich N, Rosbash M. A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition. *Genes Dev.* 1998; 12:858–867. [PubMed: 9512519]
18. Vernet C, Artzt K. STAR, a gene family involved in signal transduction and activation of RNA. *Trends Genet.* 1997; 13:479–484. [PubMed: 9433137]
19. Vega-Sanchez ME, Zeng L, Chen S, et al. SPIN1, a K homology domain protein negatively regulated and ubiquitinated by the E3 ubiquitin ligase SPL11, is involved in flowering time control in rice. *Plant Cell.* 2008; 20:1456–1469. [PubMed: 18586868]
20. Sidman RL, Dickie MM, Appel SH. Mutant Mice (Quaking and Jimpy) with Deficient Myelination in the Central Nervous System. *Science.* 1964; 144:309–311. [PubMed: 14169723]
21. Francis R, Barton MK, Kimble J, et al. gld-1, a tumor suppressor gene required for oocyte development in *Caenorhabditis elegans*. *Genetics.* 1995; 139:579–606. [PubMed: 7713419]
22. Francis R, Maine E, Schedl T. Analysis of the multiple roles of gld-1 in germline development: interactions with the sex determination cascade and the glp-1 signaling pathway. *Genetics.* 1995; 139:607–630. [PubMed: 7713420]
23. Jones AR, Francis R, Schedl T. GLD-1, a cytoplasmic protein essential for oocyte differentiation, shows stage- and sex-specific expression during *Caenorhabditis elegans* germline development. *Dev Biol.* 1996; 180:165–183. [PubMed: 8948583]
24. Taylor SJ, Shalloway D. An RNA-binding protein associated with Src through its SH2 and SH3 domains in mitosis. *Nature.* 1994; 368:867–871. [PubMed: 7512694]
25. Lock P, Fumagalli S, Polakis P, et al. The human p62 cDNA encodes Sam68 and not the RasGAP-associated p62 protein. *Cell.* 1996; 84:23–24. [PubMed: 8548822]
26. Ebersole TA, Chen Q, Justice MJ, et al. The quaking gene product necessary in embryogenesis and myelination combines features of RNA binding and signal transduction proteins. *Nat Genet.* 1996; 12:260–265. [PubMed: 8589716]

27. Lin Q, Taylor SJ, Shalloway D. Specificity and determinants of Sam68 RNA binding. Implications for the biological function of K homology domains. *J Biol Chem.* 1997; 272:27274–27280. [PubMed: 9341174]
28. Baehrecke EH. who encodes a KH RNA binding protein that functions in muscle development. *Development.* 1997; 124:1323–1332. [PubMed: 9118803]
29. Zaffran S, Astier M, Gratecos D, et al. The held out wings (how) *Drosophila* gene encodes a putative RNA-binding protein involved in the control of muscular and cardiac activity. *Development.* 1997; 124:2087–2098. [PubMed: 9169854]
30. Nabel-Rosen H, Dorevitch N, Reuveny A, et al. The balance between two isoforms of the *Drosophila* RNA-binding protein how controls tendon cell differentiation. *Mol Cell.* 1999; 4:573–584. [PubMed: 10549289]
31. Aberg K, Saetre P, Jareborg N, et al. Human QKI, a potential regulator of mRNA expression of human oligodendrocyte-related genes involved in schizophrenia. *Proc Natl Acad Sci U S A.* 2006; 103:7482–7487. [PubMed: 16641098]
32. Haroutunian V, Katsel P, Dracheva S, et al. The human homolog of the QKI gene affected in the severe dysmyelination “quaking” mouse phenotype: downregulated in multiple brain regions in schizophrenia. *Am J Psychiatry.* 2006; 163:1834–1837. [PubMed: 17012699]
33. Chen T, Richard S. Structure-function analysis of Qk1: a lethal point mutation in mouse quaking prevents homodimerization. *Mol Cell Biol.* 1998; 18:4863–4871. [PubMed: 9671495]
34. Chen T, Damaj BB, Herrera C, et al. Self-association of the single-KH-domain family members Sam68, GRP33, GLD-1, and Qk1: role of the KH domain. *Mol Cell Biol.* 1997; 17:5707–5718. [PubMed: 9315629]
35. Ryder SP, Frater L, Abramovitz DL, et al. RNA target specificity of the STAR/GSG domain post-transcriptional regulatory protein GLD-1. *Nat Struct Mol Biol.* 2004; 11:20–28. [PubMed: 14718919]
36. Ryder SP, Frater LA, Abramovitz DL, et al. RNA target specificity of the STAR/GSG domain post-transcriptional regulatory protein GLD-1. *Nat Struct Mol Biol.* 2004; 11:20–28. [PubMed: 14718919]
37. Israeli D, Nir R, Volk T. Dissection of the target specificity of the RNA-binding protein HOW reveals dpp mRNA as a novel HOW target. *Development.* 2007; 134:2107–2114. [PubMed: 17507411]
38. Galarneau A, Richard S. Target RNA motif and target mRNAs of the Quaking STAR protein. *Nat Struct Mol Biol.* 2005; 12:691–698. [PubMed: 16041388]
39. Ryder SP, Williamson JR. Specificity of the STAR/GSG domain protein Qk1: implications for the regulation of myelination. *Rna.* 2004; 10:1449–1458. [PubMed: 15273320]
40. Nabel-Rosen H, Volohonsky G, Reuveny A, et al. Two isoforms of the *Drosophila* RNA binding protein, how, act in opposing directions to regulate tendon cell differentiation. *Dev Cell.* 2002; 2:183–193. [PubMed: 11832244]
41. Larocque D, Pilotte J, Chen T, et al. Nuclear retention of MBP mRNAs in the quaking viable mice. *Neuron.* 2002; 36:815–829. [PubMed: 12467586]
42. Jan E, Motzny CK, Graves LE, et al. The STAR protein, GLD-1, is a translational regulator of sexual identity in *Caenorhabditis elegans*. *Embo J.* 1999; 18:258–269. [PubMed: 9878068]
43. Justice MJ, Bode VC. Three ENU-induced alleles of the murine quaking locus are recessive embryonic lethal mutations. *Genet Res.* 1988; 51:95–102. [PubMed: 3410318]
44. Ryder SP, Williamson JR. Specificity of the STAR/GSG domain protein Qk1: Implications for the regulation of myelination. *RNA.* 2004; 10:1449–1458. [PubMed: 15273320]
45. Galarneau A, Richard S. The STAR RNA binding proteins GLD-1, QKI, SAM68 and SLM-2 bind bipartite RNA motifs. *BMC Mol Biol.* 2009; 10:47. [PubMed: 19457263]
46. Doniach T. Activity of the sex-determining gene *tra-2* is modulated to allow spermatogenesis in the *C. elegans* hermaphrodite. *Genetics.* 1986; 114:53–76. [PubMed: 3770471]
47. Goodwin EB, Okkema PG, Evans TC, et al. Translational regulation of *tra-2* by its 3' untranslated region controls sexual identity in *C. elegans*. *Cell.* 1993; 75:329–339. [PubMed: 8402916]

48. Kuwabara PE, Okkema PG, Kimble J. tra-2 encodes a membrane protein and may mediate cell communication in the *Caenorhabditis elegans* sex determination pathway. *Mol Biol Cell*. 1992; 3:461–473. [PubMed: 1498366]
49. Ebersole T, Rho O, Artzt K. The proximal end of mouse chromosome 17: new molecular markers identify a deletion associated with quakingviable. *Genetics*. 1992; 131:183–190. [PubMed: 1592235]
50. Kondo T, Furuta T, Mitsunaga K, et al. Genomic organization and expression analysis of the mouse qkI locus. *Mamm Genome*. 1999; 10:662–669. [PubMed: 10384037]
51. Wu J, Zhou L, Tonissen K, et al. The quaking I-5 protein (QKI-5) has a novel nuclear localization signal and shuttles between the nucleus and the cytoplasm. *J Biol Chem*. 1999; 274:29202–29210. [PubMed: 10506177]
52. Li Z, Zhang Y, Li D, et al. Destabilization and mislocalization of myelin basic protein mRNAs in quaking dysmyelination lacking the QKI RNA-binding proteins. *J Neurosci*. 2000; 20:4944–4953. [PubMed: 10864952]
53. Wu JI, Reed RB, Grabowski PJ, et al. Function of quaking in myelination: regulation of alternative splicing. *Proc Natl Acad Sci U S A*. 2002; 99:4233–4238. [PubMed: 11917126]
54. Larocque D, Galarneau A, Liu HN, et al. Protection of p27(Kip1) mRNA by quaking RNA binding proteins promotes oligodendrocyte differentiation. *Nat Neurosci*. 2005; 8:27–33. [PubMed: 15568022]
55. Zhao L, Ku L, Chen Y, et al. QKI binds MAP1B mRNA and enhances MAP1B expression during oligodendrocyte development. *Mol Biol Cell*. 2006; 17:4179–4186. [PubMed: 16855020]
56. Saccomanno L, Loushin C, Jan E, et al. The STAR protein QKI-6 is a translational repressor. *Proc Natl Acad Sci U S A*. 1999; 96:12605–12610. [PubMed: 10535969]
57. Ainger K, Avossa D, Diana AS, et al. Transport and localization elements in myelin basic protein mRNA. *J Cell Biol*. 1997; 138:1077–1087. [PubMed: 9281585]
58. Lukong KE, Richard S. Sam68, the KH domain-containing superSTAR. *Biochim Biophys Acta*. 2003; 1653:73–86. [PubMed: 14643926]
59. Maguire ML, Guler-Gane G, Nietlispach D, et al. Solution structure and backbone dynamics of the KH-QUA2 region of the *Xenopus* STAR/GSG quaking protein. *J Mol Biol*. 2005; 348:265–279. [PubMed: 15811367]
60. Kramer A, Utans U. Three protein factors (SF1, SF3 and U2AF) function in pre-splicing complex formation in addition to snRNPs. *EMBO J*. 1991; 10:1503–1509. [PubMed: 1827409]
61. Arning S, Gruter P, Bilbe G, et al. Mammalian splicing factor SF1 is encoded by variant cDNAs and binds to RNA. *RNA*. 1996; 2:794–810. [PubMed: 8752089]
62. Berglund JA, Chua K, Abovich N, et al. The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell*. 1997; 89:781–787. [PubMed: 9182766]
63. Gasteiger E, Gattiker A, Hoogland C, et al. ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*. 2003; 31:3784–3788. [PubMed: 12824418]
64. Schwede T, Kopp J, Guex N, et al. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res*. 2003; 31:3381–3385. [PubMed: 12824332]
65. Lehmann-Blount KA, Williamson JR. Shape-specific nucleotide binding of single-stranded RNA by the GLD-1 STAR domain. *J Mol Biol*. 2005; 346:91–104. [PubMed: 15663930]
66. Garrey SM, Cass DM, Wandler AM, et al. Transposition of two amino acids changes a promiscuous RNA binding protein into a sequence-specific RNA binding protein. *RNA*. 2008; 14:78–88. [PubMed: 18000034]

Fig. 1. Domain structure of STAR domain protein examples GLD-1, HOW, QKI, and Sam68. The domain structure of SF1 is shown for comparison. The QUA1 region, responsible for dimerization, is in orange. The KH and QUA2 regions, which form the RNA interface, are in blue. The approximate limits of each region are denoted. Other notable domains are presented in gray.

Fig. 2.

Structures of the KH and QUA2 regions. Each image represents a single model from within a family of structures deposited within the respective PDB files. **A.** NMR structure of *H. sapiens* SF1 KH and QUA2 region bound to BPS RNA. The 5' and 3' ends of the RNA are labeled, as is the N and C termini of the protein. The KH domain is in blue, the QUA2 region in red, the variable loop is colored orange, and the GXXG loop is black. The RNA sequence is gray. **B.** NMR structure of the KH and QUA2 region of *X. laevis* QKI in the absence of RNA. The coloring scheme is the same as panel A. **C.** Overlay of the SF1 structure with the QKI structure, with RNA removed for clarity. The KH, QUA2, variable loop, and GXXG loop of SF1 are identical to panel A. The KH domain of QKI is represented in green, the QUA2 region in magenta, the variable loop in yellow, and the GXXG loop in white.

Fig. 3.

Close up view of sequence specific RNA interactions in the SF1-RNA complex NMR structure. **A.** The U6 binding pocket is shown. Amino acids side chains and the uridine nucleotide are labeled. A hydrogen bond between the backbone amide and the N5 of cytidine is represented with a dashed line. **B.** The A7 binding pocket. The amino acid side chains involved in the interaction are labeled. A base-specific hydrogen bond between the N6 exocyclic amine of A7 and a Glu 149 is represented with a dashed line. **C.** The A8 binding pocket. Amino acids and base specific hydrogen bonds are labeled as in panels A and B.

Fig. 4.

Diagram of the protein-RNA contacts observed in the SF1-BPS RNA complex structure. The sequence of the RNA is shown next to a backbone diagram. The UACUAA element is bold. The SF1 amino acids that contact RNA are presented next to the nucleotide they interact with in the structure. The corresponding amino acid in GLD-1, HOW, QKI, and Sam68 is given in a box above. Conserved amino acids are in gray. Blue boxes represent positions where all four proteins differ from SF1, but the identity of the amino acid difference is the same in GLD-1, QKI, and HOW. Red boxes represent positions where all four proteins differ from SF1 but the difference is identical in all four STAR proteins. Red font indicates a position where only Sam68 differs from the SF1 sequence.

Table 1
Nucleotide sequence specificity of STAR domain proteins

RNA recognition by STAR domain proteins. The STAR protein identity is listed in the first column. The second column contains the RNA-binding consensus sequence. Degenerate nucleotides are contained within parentheses. The third column annotates the experimental method used to measure the specificity.

Protein	Specificity	Method
GLD-1	5'-UACU(C/A)A-3' (conservative) 5'(U>C>G/A)A(C>A)U(C/A)A-3' (relaxed)	Gel Shift / Mutagenesis ³⁶
QKI ^a	5'-A(C/A)UAA-3'	FP / Mutagenesis ⁴⁴
QKI ^b	5'-ACUAA-3' (core) 5'-(U/C)AA(U/C)-3' (half-site)	SELEX ³⁸
HOW	5'-A(C>A)UAA-3'	Pull Down ³⁷
Sam68	5-UAAA-3'	SELEX ²⁷

^aBinding specificity determined by Ryder and Williamson⁴³

^bBinding specificity determined by Galarneau et al.³⁷