The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain

Kirk B. Jensen*, Kiran Musunuru*†, Hal A. Lewis†‡, Stephen K. Burley†§, and Robert B. Darnell*¶

Laboratories of *Molecular Neuro-Oncology and †Molecular Biophysics, and §Howard Hughes Medical Institute, The Rockefeller University, 1230 York Avenue, New York, NY 10021

Edited by Larry Gold, NeXstar Pharmaceuticals, Inc., Boulder, CO, and approved February 16, 2000 (received for review December 17, 1999)

The Nova family of proteins are target antigens in the autoimmune disorder paraneoplastic opsoclonus-myoclonus ataxia and contain K-homology (KH)-type RNA binding domains. The Nova-1 protein has recently been shown to regulate alternative splicing of the α2 glycine receptor subunit pre-mRNA by binding to an intronic element containing repeats of the tetranucleotide UCAU. Here, we have used selection-amplification to demonstrate that the KH3 domain of Nova recognizes a single UCAY element in the context of a 20-base hairpin RNA; the UCAY tetranucleotide is optimally presented as a loop element of the hairpin scaffold and requires protein residues C-terminal to the previously defined KH domain. These results suggest that KH domains in general recognize tetranucleotide motifs and that biological RNA targets of KH domains may use either RNA secondary structure or repeated sequence elements to achieve high affinity and specificity of protein binding.

In eukaryotes, transcription of pre-mRNA is followed by a series of processing steps, including mRNA splicing, polyadenylation, export, localization, translation, and degradation. RNA-protein interactions are central in all of these pathways, and RNA binding proteins are often key regulators of posttranscriptional gene expression. The elucidation of how RNA binding proteins interact with RNA is thus critical to the understanding of these aspects of mRNA processing.

One of the most common protein motifs involved in RNA binding is the K-homology (KH) domain, originally described in the protein hnRNP-K (1). The KH domain consists of approximately 70 amino acids and includes a conserved hydrophobic core, an invariant Gly-X-X-Gly motif, and an additional variable segment; isolated KH domains of approximately this length are competent to bind ribohomopolymers (2–4). Over 50 KH domain-containing proteins have been identified; notably, many of these proteins contain multiple KH domains and/or other known RNA binding domains, such as the RNA recognition motif. NMR structural studies of individual KH domains revealed a conserved $\beta\alpha\alpha\beta\beta\alpha$ fold (3, 5), and high-resolution x-ray structures have additionally permitted the visualization of the Gly-X-X-Gly and variable loops (6).

How KH domain proteins function requires understanding how they interact with RNA and the roles these interactions play in the context of a presumably complex and multicomponent regulatory system. The mammalian KH domain proteins KSRP and SF1, the yeast KH protein MER-1, and the Drosophila KH domain protein PSI have all been implicated in regulating mRNA splicing (7–10). The heterogeneous nuclear RNP proteins hnRNP K and hnRNP E have been shown to play important roles in mRNA stabilization and mRNA translational control (11–13). The proteins ZBP-1 and Vera, which each contain two RNA recognition motif domains and four KH domains, have been shown to play a role in localizing specific mRNAs within the cell cytoplasm (14, 15). The absence of the mammalian KH domain protein FMR-1 is associated with the fragile-x mental retardation disorder, and a single amino acid mutation within its second KH domain impairs RNA binding and leads to a particularly severe form of mental retardation in humans (16–18). The functional roles of several of these KH domain proteins have been linked to their recognition of pyrimidine-rich sequences within their target RNAs, but we have yet to understand in detail how the specificity of RNA recognition is achieved and what other protein factors, if any, are necessary for KH domain protein function.

The Nova-1 and Nova-2 proteins (4, 19, 20) each contain three KH domains and are closely related to the hnRNP-K and hnRNP-E proteins (21). Nova proteins are exclusively expressed in CNS neurons and are the target antigens in the autoimmune neurological disorder paraneoplastic opsoclonus myoclonus ataxia. We have previously used RNA selection-amplification (SELEX) (22, 23) to identify RNA stem loops containing the element $(UCAUY)_3$ as high affinity targets of Nova-1 (24) and have shown that Nova-1 specifically interacts with an intronic element of the same sequence in the pre-mRNA of the α2 glycine receptor subunit (24). Using both genetic and biochemical means, we have demonstrated that Nova-1 binding to this intronic site in the α2 glycine receptor can control the relative utilization of the two downstream, alternatively spliced exons 3A and 3B (25). Additionally, selection-amplification experiments with Nova-2 revealed a consensus of GAG**UCAU** in a stem loop as a high-affinity target of the protein (20).

High-resolution x-ray structures of the third KH domain from Nova-1 and Nova-2 have been solved (6). This particular KH domain of Nova was chosen for structural study because it has been shown to be both necessary and sufficient for binding of the $(UCAUY)_3$ RNA motif (24). For the purposes of obtaining a Nova KH3/RNA co-crystal, we have again used RNA selection-amplification to determine a high-affinity RNA target for the isolated Nova KH3 domain, and we report here the results of that screen. The RNA molecules selected using Nova KH3 contain a single UCAY element that lies within the loop of a 20-base hairpin structure. We have performed extensive mutagenesis on this molecule, confirming the strict requirement of the UCAY and its presentation within the context of a stem loop as necessary for recognition by Nova KH3. We have determined that several amino acids C-terminal to the previously defined KH domain are necessary for RNA binding, and we demonstrate that the specificity and affinity of this RNA:protein interaction is preserved in the full-length Nova protein. These results have led to the co-crystallization of Nova-2 KH3 and one of the RNAs from this selection (26), and this structure shows excellent agreement with the mutagenesis and

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: KH domain, K-homology domain; UTR, untranslated region.

‡Present Address: Department of Biology, Brookhaven National Laboratory, Upton, NY 11973.

¶To whom reprint requests should be addressed. E-mail: darnelr@rockvax.rockefeller.edu.

Article published online before print: *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.090553997.
Article and publication date are at www.pnas.org/cgi/doi/10.1073/pnas.090553997

protein domain mapping studies reported here. Finally, we compare the Nova KH3 RNA ligand to known RNA sequences that interact with KH domain proteins and suggest that KH domains recognize RNA targets based on a minimal four-base RNA sequence element. This interaction is weak, in both the energetic and informational sense, and specificity for targets is likely achieved through the use of multiple KH domains within the same protein, the presence of multiprotein complexes on the RNA target, or, as demonstrated in the case of the Nova KH3 RNA target, an entropically favorable presentation of the core tetranucleotide to the KH domain.

Materials and Methods

KH Domain Expression and Purification. The KH3/C-terminal domains from human Nova-1 (amino acids 423–510) and human Nova-2 (amino acids 406–492) were expressed and purified as in the work by Lewis *et al.* (6). The smaller protein fragments were prepared in an identical fashion. The proteins were assayed by electrospray mass spectrometry for correct molecular weight, circular dichroism for correct protein folding, and dynamic light scattering for monodispersity.

Selection-Amplification. The synthetic oligonucleotide template 5′-TCCCGCTCGTCGTCT [25N] CCGCATCGTCCTCCCT-3′, where "N" indicates random incorporation of all four nucleotides, was prepared for first round transcription by using Klenow fragment (Amersham Pharmacia) and the oligonucleotide primer 5GL, 5′-GAAATTAATACGACTCACTATAGG-GAGGACGATGCGG-3′. RNA was transcribed to yield approximately 8 nmol of full-length product (and approximately 10^{14} unique RNA molecules) for first round selection. Selection-amplification was performed essentially as described by Tuerk and Gold (22). Binding reactions were carried out by using the buffer 1 × BB [200 mM KOAc/50 mM Tris·OAc, pH 7.7/5 mM Mg(OAc)$_2$], and RNA:protein molar ratios varying from 5:1 at the beginning of selection to 250:1 at the end of selection. Partitioning was performed with 0.45-μm nitrocellulose filters (Millipore). Selected RNA was reverse transcribed by using the oligonucleotide primer 3GL, 5′-TCCCGCTCGTCGTCTG-3′, and AMV-RT (Promega). Amplification of the library for the following rounds used mildly mutagenic PCR conditions using primers 5GL and 3GL, 1 mM dNTPs, and 7.5 mM Mg(OAc)$_2$.

Measurement of RNA-Protein Binding. Binding dissociation constants were measured either by a nitrocellulose filter binding assay (27) or by gel shift assay. For filter binding, 50-μl reactions containing 50–100 fmol of RNA internally labeled with ^{32}P and concentrations of Nova-1 or Nova-2 KH3 in 3-fold dilutions typically ranging from 33 μM to 45 nM were mixed in 1 × BB and were incubated at 10 min for 25°C, followed by filtering and washing. For gel shifts, binding reactions were modified as follows: 1 × GS buffer [50 mM KOAc/50 mM Tris·OAc/10 mM DTT/5 mM Mg(OAc)$_2$/30 μg/ml tRNA] was substituted for 1 × BB; reaction volume was 20 μl; reaction temperature was 4°C; and Ficoll 400-DL (Sigma) was added to a final concentration of 2.5%. Samples were run on 8%, 37.5:1 acrylamide:bisacrylamide gels at 4°C and 100 V. Dissociation constants were determined graphically by plotting the fraction of bound RNA versus the log of the protein concentration (28).

Transcription of Oligonucleotide Templates. The template for the 20-mer RNA 10021 was prepared by using the oligonucleotide 5′-GCGGGGTGATCTTAGGTCCGCTATAGTGAGTCG-TATTA-3′. Oligonucleotides for all 10021 mutants are based on this sequence. The 10021 template was annealed to the oligonucleotide 5′-TAATACGACTCACTATAG-3′ for transcription, and RNA synthesis carried out by using either α-^{32}P-UTP or -CTP in standard transcription buffer (Stratagene) supplemented with 80 mg/ml PEG 8000 (Sigma). Transcripts were size-purified by using 20% denaturing PAGE.

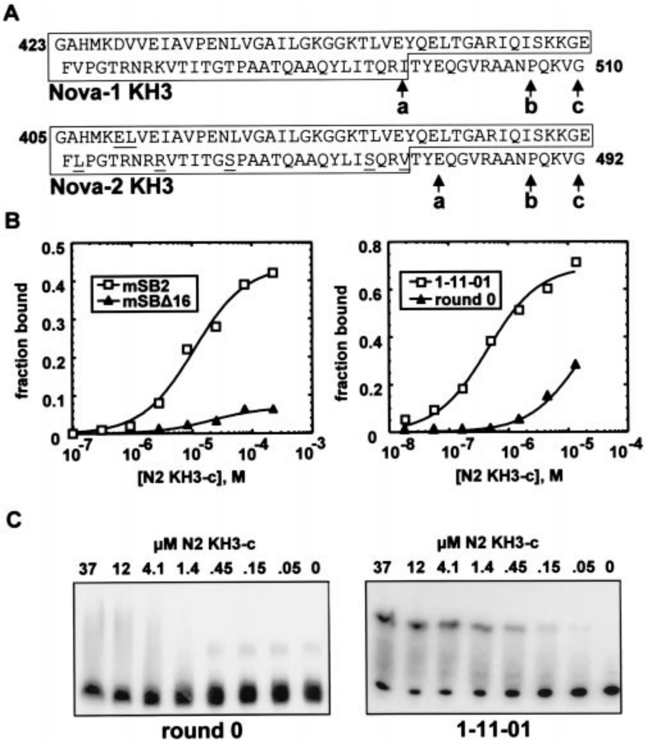
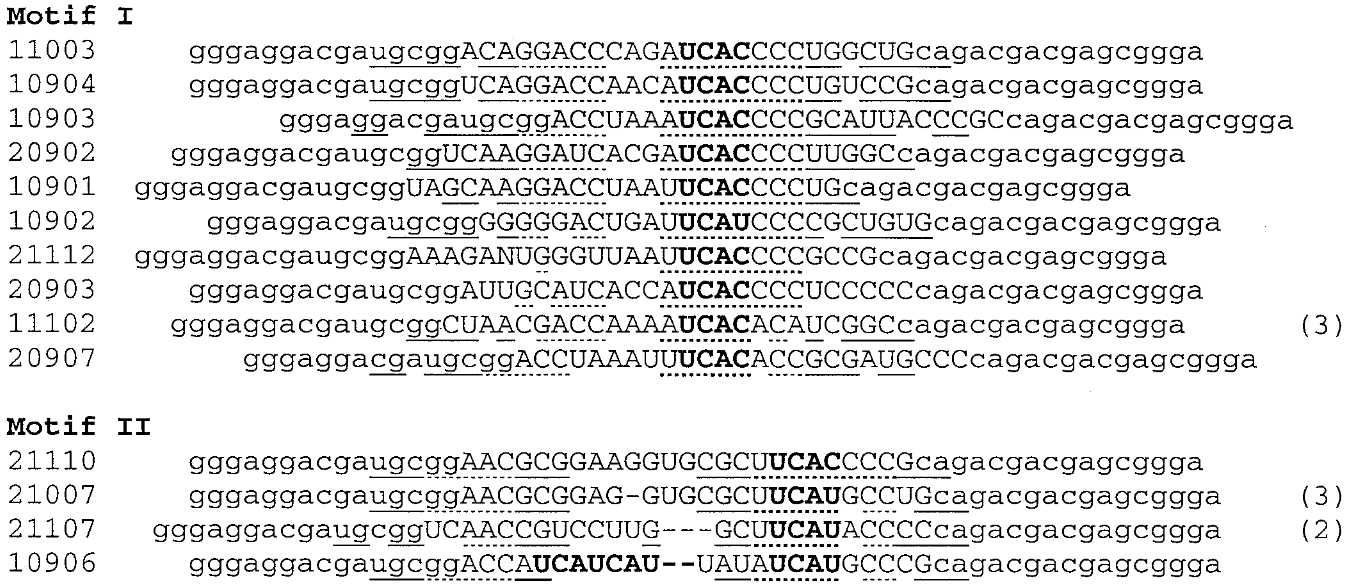


Fig. 1. (*A*) Schematic representation of the Nova-1 and Nova-2 KH3 domain constructs used in this study. The boxed region indicates the KH domain (6). Amino acids that vary between Nova-1 and Nova-2 are underlined in the Nova-2 sequence. Arrows indicate the C-terminal residue in the "a," "b," and "c" forms of each protein. The longest KH3 forms, c, were used for the selection-amplification of RNA. (*B*) Filter binding assays with Nova-2 KH3. The first graph displays the binding of a 21-mer RNA that contains a UCAUY triplet repeat (mSB2) and a mutant RNA (mSBΔ16), which contains a UAAUY triplet. The second graph depicts the binding of the selection-amplification round 0 pool (preselection) and the clone 1-11-01, which was isolated from round 11 of the selection. For clarity, all binding assays depicted employ Nova-2 KH3, although Nova-1 KH3 behaved in an identical fashion in all situations tested (data not shown). (*C*) Gel shift assays using Nova-2 KH3 with the round 0 pool and the selected clone 1-11-01.

Results and Discussion

Selection-Amplification of RNA for Nova KH3. The protein domains of Nova-1 (residues 423–510) and Nova-2 (residues 405–492) used for RNA selection-amplification include the KH3 domain and the remaining C-terminal extension of each Nova protein (the KH3-c constructs in Fig. 1*A*). These two Nova fragments are 81% identical, containing seven conservative amino acid changes, and both proteins bound the round 0 pool of RNA with a K_d of greater than 10 μM. Although the previous x-ray structures of Nova lacked this C-terminal addition (Fig. 1*A*; KH3-a constructs) limited proteolysis of full length Nova-1 in the presence of a (UCAUY)$_3$ RNA indicated that this C-terminal protein fragment was protected by RNA binding, and thus possibly necessary for high-affinity RNA binding to Nova KH3 (6).

The two Nova KH3 protein fragments were assayed for binding to a (UCAUY)$_3$-containing 21-mer (mSB2), previously identified by using selection-amplification with full-length Nova-1, and an identical RNA in which the core recognition sequence had been mutated to (UAAUY)$_3$ (mSBΔ16) (24). Although the isolated Nova KH3 domain retained a clear ability to bind and discriminate the wild-type UCAU-containing sequence from the mutant RNA, the overall K_d of the interaction



```
Motif I
11003      gggaggacgaugcggACAGGACCCAGAUCACCCCUGGCUGcagacgacgagcggga
10904      gggaggacgaugcggUCAGGACCAACAUCACCCCUGUCCGcagacgacgagcggga
10903           gggaggacgaugcggACCUAAAUCACCCCGCAUUACCCGCcagacgacgagcggga
20902     gggaggacgaugcggUCAAGGAUCACGAUCACCCCUUGGCcagacgacgagcggga
10901   gggaggacgaugcggUAGCAAGGACCUAAUUCACCCCUGcagacgacgagcggga
10902       gggaggacgaugcggGGGGACUGAUUCAUCCCCGCUGUGcagacgacgagcggga
21112   gggaggacgaugcggAAAGANUGGGUUAAUUCACCCCGCCGcagacgacgagcggga
20903      gggaggacgaugcggAUUGCAUCACCAUCACCCCUCCCCCcagacgacgagcggga
11102     gggaggacgaugcggCUAACGACCAAAAUCACACAUCGGCcagacgacgagcggga       (3)
20907         gggaggacgaugcggACCUAAAUUUCACACCGCGAUGCCCcagacgacgagcggga

Motif II
21110      gggaggacgaugcggAACGCGGAAGGUGCGCUUCACCCCGcagacgacgagcggga
21007      gggaggacgaugcggAACGCGGAG-GUGCGCUUCAUGCCUGcagacgacgagcggga     (3)
21107    gggaggacgaugcggUCAACCGUCCUUG---GCUUCAUACCCCcagacgacgagcggga    (2)
10906      gggaggacgaugcggACCAUCAUCAU--UAUAUCAUGCCCGcagacgacgagcggga
```

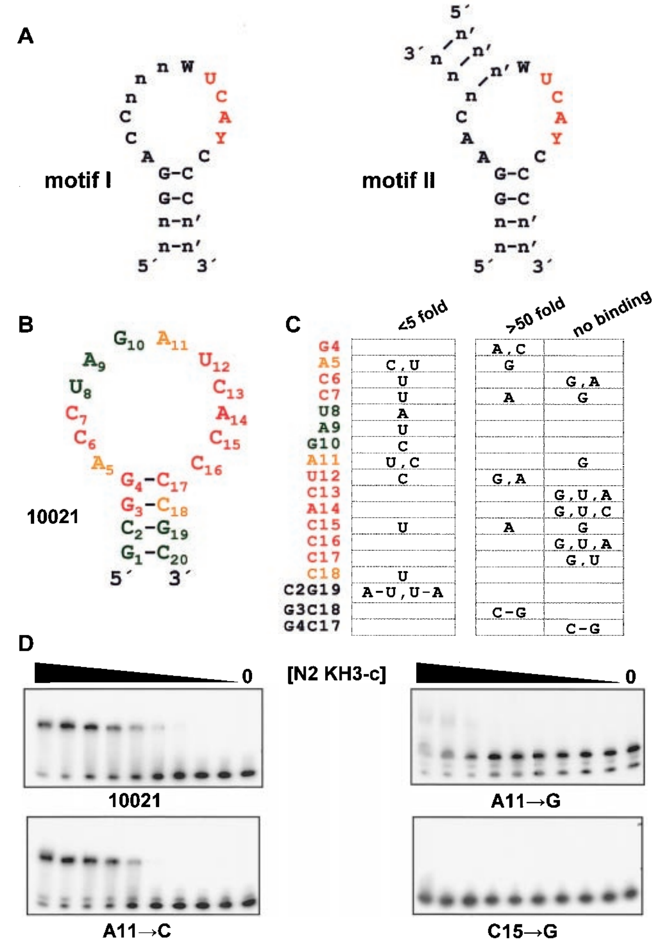
Fig. 2. Results of selection amplification using Nova-1 and Nova-2 KH3 domains. The first digit of each clone number indicates the clone's origin as from the Nova-1 or Nova-2 selection, the second two digits indicate the selection round number (9, 10, or 11), and the last two digits the individual clone number. Sequences are arranged in motifs according to sequence homology. Lowercase letters indicate the fixed sequence of the selection-amplification RNA. The core UCAY Nova KH3 recognition element is indicated by boldface in each sequence. Conserved nucleotides (including the two first base pairs of the conserved motif I and II stem) are underscored with dashes, and potential base pairing interactions are indicated by a solid underscore. Sequences from individual clones that appear to be identical are represented only once, with the number of individual isolates indicated in parentheses after the sequence.

was approximately 10 μM (Fig. 1*B*), over 200-fold less than the K_d of the same RNA with full length Nova, and unsuitable for co-crystallization trials. To select an RNA with a higher affinity for the isolated Nova KH3 domain, we used a library with 25 random positions and performed selection-amplification individually on both Nova-1 KH3 and Nova-2 KH3 for 11 rounds. Fig. 1*B* shows that the affinity of the round 0 pool for Nova-2 KH3 was detectable by filter binding at concentrations of protein approaching 10 μM, but this interaction appeared quite nonspecific as measured by gel-shift assay (Fig. 1*C*).

The selected pools were cloned and sequenced at rounds 9, 10, and 11, and an aligned sequence set of the most abundant clones is displayed in Fig. 2. The largest sequence class (31% of the total), made up of sequences from both the Nova-1 and Nova-2 KH3 selections, is termed motif I. This set is characterized by a conserved bipartite sequence element GGACC[n3]W**UCAY**CCCC in which W is A or U, Y is U or C, and [n3] can be any three bases (29). The "phylogenetic" analysis of this motif's secondary structure suggests that these sequences form a 5- to 10-base pair stem-loop structure shown in Fig. 3*A*. Clone 1-11-01, which contains both the conserved sequence element and a 10-base stem, binds the Nova KH3 domain with a K_d of approximately 500 nM by both filter binding and gel-shift assay (Fig. 1 *B* and *C*). A second class of selected sequences, motif II, contains an almost identical conserved bipartite element GGAAC[n13]W**UCAY**CCCC, differing primarily in the distance between the two conserved elements. Motif II can be base-paired to give a stem-bulge-stem structure that is remarkably similar in secondary structure to that of motif I, with a second stem replacing the three variable positions of the motif I loop (Fig. 3*A*). The full length motif II clone 2-11-10 binds with a similar K_d to that of the motif I clones (data not shown). Another 22% of the cloned sequences contained a core UCAU or UCAC element (sometimes multiple elements), but we have not assessed their secondary structures in detail (data not shown). An additional class appeared not to bind the Nova KH3 domain but show modest affinity for the selection-amplification partitioning matrix. Finally, seven orphan sequences could not be classified into any motif.

Nova KH3 Binds a Hairpin Loop Containing UCAY. We have made a detailed analysis of the interaction of the motif I type clones with the Nova KH3 domain. To determine the minimum RNA element necessary for specific binding to Nova KH3, a series of deletions to the putative 10-base stem of motif I clone 1-11-01 was performed. A 20-mer RNA, identical to 1-11-01 but with only a four-base stem, termed 10021 (Fig. 3*B*), was able to bind to both Nova-1 and Nova-2 KH3 with a K_d of 500 nM, identical to that of the full length parent clone (Fig. 3*D*). This core 20-base RNA was then used as a template for mutational studies. Thirty-five single point mutations in 10021, including all single base changes within the loop region of the conserved bipartite element ACC[n3]A**UCAC**C, were tested. Furthermore, we assayed several changes to the nonconserved loop region of the molecule and within the four-base stem. All mutational analysis was carried out by gel-shift assay using Nova-2 KH3-c (Fig. 1), but identical results were obtained with Nova-1 KH3-c (data not shown).

The results of the mutational study are presented in Fig. 3*C*. The binding characteristics of the mutants are classified into three categories: "tolerated," with a small to nondetectable difference in binding to Nova KH3 (< 5 fold); "poor binders," with significant change in affinity (> 50 fold); and "nonbinders," in which there was no detectable binding at all at the highest concentration of protein used (> 33 μM). Gel-shift autoradiograms of 10021 and a representative mutant from each of the three categories of binding to Nova-2 KH3 are shown in Fig. 3*D*. Within the first half of the bipartite ACC[n3]A**UCAC**C element (A5-C7), a substitution of any of the three positions to a pyrimidine is well tolerated; purine substitutions either severely effect binding (A5→G; C7→A) or eliminate binding altogether (C6→G,A; C7→G). In the second half of the ACC[n3]A**UCAC**C element (A11-C16), mutations as a whole are not tolerated. Within the "core" UCACC sequence, only 2 of 15 point mutations retain significant binding to Nova KH3. At three positions, C13, A14, and C16, all nine possible point mutants completely destroy the interaction between the RNA and the protein. At positions U12 and C15, pyrimidine substitutions are



	<5 fold	>50 fold	no binding
G4		A , C	
A5	C , U	G	
C6	U		G , A
C7	U	A	G
U8	A		
A9	U		
G10	C		
A11	U , C		G
U12	C	G , A	
C13			G , U , A
A14			G , U , C
C15	U	A	G
C16			G , U , A
C17			G , U
C18	U		
C2G19	A-U , U-A		
G3C18		C-G	
G4C17			C-G

Fig. 3. (*A*) Putative secondary structures of the motif I and II sequences. Uppercase letters represent the conserved nucleotides found in each sequence set. (*B*) Predicted secondary structure of 10021. (*C*) Table of the 10021 mutational analysis. The sequence of 10021 is displayed vertically at the left of the table. Point mutations at each nucleotide position are categorized into three classes according to their effect on the RNA's dissociation constant with Nova KH3: less than 5-fold decrease, greater than 50-fold decrease, and no binding at all to Nova KH3. Dissociation constants were measured by gel-shift assay. Listed at the bottom of the table are additional mutational changes to the 10021 stem. (*D*) Autoradiograms of gel-shift assays with the 20-mer RNA 10021 and selected point mutants. Nova-2 KH3-c protein is titrated from left to right by three-fold dilutions starting at 37 μM; additional conditions are described in *Materials and Methods*. A representative of each category of mutational defect is shown. A11→C is a tolerated mutation (3-fold decrease in affinity), A11→G is a poor binder (100-fold decrease in affinity), and C15→G is a nonbinder.

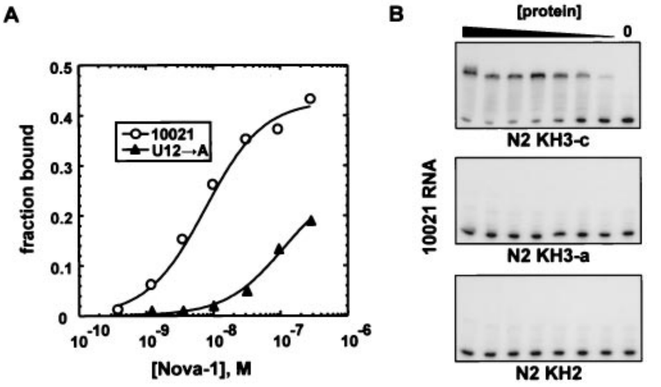


Fig. 4. (*A*) Nitrocellulose filter binding assay of the 10021 RNA and a point mutant RNA against full length Nova-1. The 10021 RNA binds full length Nova-1 with a K_d of approximately 5 nM; the point mutant 20-mer U12→A binds with a K_d of >250 nM. (*B*) Mapping the minimal protein domain of Nova KH3 necessary to bind the 10021 RNA. Protein is titrated from left to right by 3-fold dilutions starting at 37 μM. The top gel shift assay shows the binding of 10021 to the full C-terminal fragment of Nova-2 KH3 (N2 KH3-c). The binding of 10021 to Nova-2 KH3-b (which lacks four C-terminal amino acids) is identical to the c construct. The middle gel-shift autoradiogram is the identical 10021 RNA assayed with the Nova-2 KH3-a protein, which is lacking the 15 C-terminal amino acids. The bottom autoradiogram is the same 10021 RNA assayed with the second KH domain of Nova-2 (N2 KH2).

tolerated, but purines either cause severe binding impairment or complete loss of RNA binding. Thus, the UCACC core is clearly, by mutational criteria, the most important region of the RNA for Nova KH3 recognition, with the C13-A14 dinucleotide being absolutely necessary for binding.

The stem nucleotides of 10021 also play an important role in binding to Nova KH3. The base pair most proximal to the loop, G4C17, was first assayed by reversing the bases to form the pair C4G17. This substitution did not bind Nova KH3. Interestingly, the mutant G4U17, which replaces a G-C pair with a G-U pair, completely eliminated binding. Mismatches in this base pair result in poor binding (G4→A,C), or abolish binding altogether (C17→G). This asymmetry in mutation severity implicates C17 in a more important role in Nova KH3 recognition than G4, but clearly, base pairing of these two nucleotides is required for best binding. The identity of the preceding base pair, G3C18, is also critical, with the reverse pair C3G18 severely affecting binding. However, the C18→U mutation, which substitutes a G-U pair for a G-C pair, does not significantly impair protein binding. The next base pair, C2G19, could be altered to either an A-U or U-A pair with no penalty on protein binding. The most distal base pair was not tested for mutations. We hypothesize that the contribution of the two distal base pairs is for purely structural (entropic) reasons, and that the stem as a whole is important in constraining the conformation of the loop and especially the UCAYC core.

Finally, we find that point mutations within the variable U8-G10 region of the loop do not seem to affect Nova KH3 binding, although we have not exhaustively probed these positions. In contrast, deletion of two or three bases (ΔA9-G10 or ΔU8-G10) within this region abolishes protein binding, and even a single-base deletion (ΔG10) severely affects the protein-RNA interaction. Thus, it is likely that these bases function as a "spacer"—providing the appropriate geometry of the loop for recognition by Nova KH3 (data not shown).

Full-Length Nova Also Specifically Recognizes the 10021 RNA. Previous data from our laboratory has demonstrated that the KH3 domain of Nova-1 is necessary and sufficient for binding to the sequence UCAY in RNA (24). To test whether the RNA ligands raised against Nova KH3 alone are capable of binding to full-length Nova protein, we performed filter binding assays with the complete Nova protein. The 10021 RNA binds full length Nova-1 with a K_d of approximately 8 nM, as shown in Fig. 4*A*. The stronger binding seen with full length protein may be attributable to dimerization of the full-length Nova-1 protein in solution, as suggested by dynamic light scattering experiments (H.A.L. and S.K.B., unpublished observation); the greater apparent K_d would reflect the presence of a second binding site for the 10021 RNA on the same Nova-1/Nova-1 complex combined with the underlying binding equilibrium of the Nova protein dimer. Alternatively, the RNA might in fact bind more strongly to the KH3 domain in the context of the entire protein. To test whether the interaction with the 10021 RNA and full-length Nova-1 is specific, we also assayed the binding of the

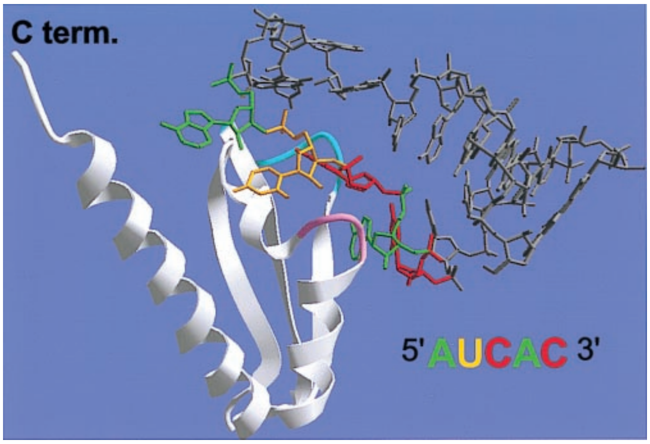
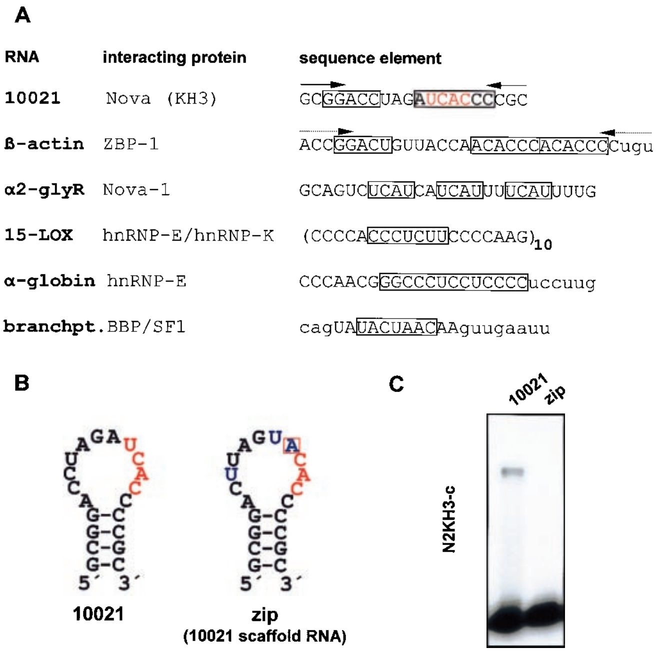


Fig. 5. Structure of the Nova-2 KH3 domain with the 10021 RNA as presented by Lewis *et al.* (26). The RNA (bases 1–10 and 16–20 in gray) assumes the predicted hairpin conformation, with AUCAC (A11-C15) responsible for the majority of the contacts with the protein (in white). The colored AUCAC lies in a cleft formed by the KH domain invariant (in light purple) and variable (in light blue) loops; the bases make extensive contacts with an underlying aliphatic α helix/β sheet binding platform.



RNA	interacting protein	sequence element
10021	Nova (KH3)	GCGGACCUAGAUCACCCCGC
ß-actin	ZBP-1	ACCGGACUGUUACCAACACCCACACCCCugu
α2-glyR	Nova-1	GCAGUCUCAUCAUCAUUUUCAUUUUG
15-LOX	hnRNP-E/hnRNP-K	(CCCCACCCUCUUCCCCAAG)$_{10}$
α-globin	hnRNP-E	CCCAACGGGCCCUCCUCCCCuccuug
branchpt.	BBP/SF1	cagUAUACUAACAAguugaauu

Fig. 6. (*A*) Table of KH domain binding elements from β-actin (15), α2-glyR pre-mRNA (24, 25), 15-lipoxgenase (12, 31), and α-globin 3′ UTRs (35), and a canonical pre-mRNA branchpoint (34, 36). Uppercase letters represent the minimal sequence element necessary for interaction with the cognate KH domain protein. Boxed sequences indicate nucleotides most important for interaction with the protein, as determined by mutagenesis or deletion analysis. The UCAC sequence of 10021, which interacts directly with Nova KH3, is shown in red. Base pairing for 10021 is indicated by arrows; possible base pairing for the β-actin 3′ UTR element is indicated by the dashed arrows. (*B*) Sequence and putative secondary structure of 10021 and the "10021-scaffold" RNA zip, which is based on the β-actin RNA sequence. Changes to the 10021 molecule are indicated in blue; changes that are known to disrupt Nova KH3 binding are boxed in red. (*C*) Gel shift assay with 10021 and the zip 10021-scaffold RNA. Only 10021 forms a shifted complex with Nova KH3-c (protein concentration is 1 μM).

U12→A point mutant RNA, a poor binder for Nova KH3. This RNA shows approximately 50-fold less affinity for the full-length protein than 10021 (Fig. 4*A*). Thus, the specificity of the Nova KH3 interaction for 10021 is intact even in full-length Nova-1. It is unlikely that the higher affinity of the 10021 RNA for full-length Nova-1 is caused by interaction with the KH1 and KH2 domains (see Fig. 4*B* and *Conclusions*).

RNA Binding by Nova KH3 Requires Residues C-Terminal to the KH Domain. Because our proteolysis protection data indicated that a UCAY RNA protected both regions of the KH3 domain and the adjacent C terminus from digestion, we explored what fragment of Nova KH3 was necessary to retain binding to the 10021 RNA. Fig. 4*B* displays gel shifts using the 10021 RNA with two Nova-2 KH3 constructs and the Nova-2 KH2 domain. The 10021 RNA binds the full C-terminal KH3 construct (KH3-c) with a K_d of approximately 500 nM. The dissociation constant using the N2 KH3-b construct (Fig. 1*A*), which is lacking the four terminal amino acids, binds 10021 identically (data not shown). However, when the N2 KH3-a construct is used, which lacks the 15 most C-terminal amino acids, binding to 10021 is lost (Fig. 4*B*). This is an important observation because the Nova-2 KH3-a protein contains the entire KH domain as defined previously by protein sequence alignment (6). Thus, in the 11 amino acids that lie C-terminal to the defined KH domain are residues that directly or indirectly interact with the 10021 RNA, and are possibly necessary for recognition of *in vivo* RNA targets of Nova. The Nova-2 KH2 domain (Fig. 4*B*) also completely fails to bind the 10021 RNA.

The 10021 RNA as a General Tetranucleotide Scaffold for KH Domain Recognition. The Nova KH3 domain is a fully functional RNA binding domain, capable of a highly sequence-specific interaction with RNA. Our selection-amplification experiments demonstrate that the core element of RNA recognition depends on the presence of the tetranucleotide UCAY, and this element is best presented to Nova KH3 in the context of a 12 base loop defined as ACC[n3]W**UCAY**C. An RNA target from this selection was recently co-crystallized with the Nova-2 KH3 domain (26), and precisely confirms the conclusions of this present study. The crystal structure reveals the core UCAY of the RNA gripped in a "molecular vise" formed by the cleft between the invariant and variable loops of the Nova KH3 domain (Fig. 5). The vast majority of the RNA-protein contacts occur within the segment A**UCAC** (A11-C15), the RNA itself is in the predicted hairpin conformation, and there are additional stacking interactions continuing from the four-base stem through the next two nucleotide pairs, A5-C16 (which also forms one hydrogen bond), and C6-C15. A11 contacts an arginine in the 15-residue C-terminal fragment of Nova KH3; this contact explains the requirement of the C-terminal extension for high affinity binding to RNA (26).

The selection-amplification, mutagenesis, and co-crystal studies suggest that the minimal 20-base 10021 RNA hairpin is an optimal "scaffolding" for the presentation of the core UCAY to Nova KH3, one that is likely to be reflected in biologically relevant Nova RNA targets. To examine the biological implications of the selection-amplification results, we examined the published examples of RNA sequences that have been shown to interact with the KH domain proteins Nova-1, hnRNP-E, hnRNP-K, ZBP-1, and branchpoint binding protein/SF1. hnRNP-E1 and E2 are members of a protein complex (the α-complex) that bind to the 3′ untranslated region (UTR) of the α-globin gene and can stabilize the mRNA (13, 30). Similarly, a section of the 15-lipoxygenase 3′ UTR, which contains a 19-base pyrimidine-rich repeated motif, can be crosslinked to hnRNP-E1 and K and is necessary and sufficient to confer translational silencing of a message by the two proteins (12, 31). Also, a 27-base sequence of the 3′ UTR of β-actin has been shown to specifically crosslink to ZBP-1; this "zip-code" sequence element is necessary and sufficient to confer localization of the β-actin message to the leading edge of fibroblasts. (15). The yeast branchpoint binding protein and its mammalian orthologue SF1 (32) bind the conserved pre-mRNA branchpoint sequence UACUA$\underline{A}$C (33) (in which the underlined A is the branchpoint) during commitment-complex formation in splicing.

Fig. 6*A* displays the RNA targets for these KH domain proteins, along with the 10021 RNA. The boxed nucleotides in each figure indicate those bases most critical for KH domain binding and/or biological function. The β-actin RNA contains a GGACU element followed by ACACCC repeats (Fig. 6*A*). The GGACU-ACACCC elements are strikingly similar to the 10021 bipartite loop sequence, and the β-actin RNA can be base paired to form a hairpin that would place these RNA elements in a manner spatially equivalent to the motif I and II sequences for Nova KH3 (Fig. 6*B*). Thus, perhaps the 10021 scaffolding may be used in some biological targets of KH domain proteins. Despite the remarkable overall structural and sequence similarity between the zip element and 10021, Nova KH3 is able to absolutely discriminate between them (Fig. 6*C*). Therefore, the scaffolding may be an optimal strategy for the presentation of the core tetranucleotide to a KH domain. The RNA elements from α2-glyR, α-globin, 15-lipoxygenase, and the pre-mRNA branchpoint do not suggest any obvious secondary structures but do share with 10021 a small pyrimidine/adenine element that is likely the core recognition element for their cognate KH domain proteins. A different strategy might be used by α2-glyR and 15-lipoxygenase targets, which repeat their core sequence elements from 3–10 times (see below).

Conclusions

An unexpected result of this study was the finding that the previously defined KH domain of Nova KH3 was not capable of high affinity binding of the 10021 RNA, requiring the addition of a number of amino acids C-terminal to the KH domain. The co-crystal structure revealed that A11 of 10021 interacts with an arginine residue seven amino acids C-terminal to the KH domain (6). A similar result was found for branchpoint binding protein, in which a construct retaining 31 amino acids C-terminal to the KH domain preserved specificity for the branchpoint sequence whereas the KH domain by itself did not (34). Thus, at least for some KH domain proteins, residues proximal to the defined KH domain may be critical for proper RNA recognition.

The selection-amplification results for Nova KH3 agree well with previous SELEX experiments for the full length Nova-1 and Nova-2 proteins, in which UCAU was found as necessary for high affinity binding to the proteins (20, 24). The protein contacts made by the UCAY sequence in the Nova KH3 co-crystal cannot be duplicated by the Nova KH1 and KH2 domains (26), and biochemical data suggests that Nova KH3 is sufficient for binding of the full length Nova-1 SELEX targets (this study and ref. 24). Our preliminary selection-amplifications results suggest that Nova KH1 and KH2 recognize different RNA targets (K.B.J., K.M., and R.B.D., unpublished observations).

Finally, we have obtained data that implicates Nova-1 in the regulation of alternative splicing of the α2 glycine receptor subunit (25). The target recognized by Nova-1 in the glycine pre-mRNA is an intronic $(UCAU)_3$ element, and, although it contains the absolutely conserved UCAY core necessary for Nova recognition, in secondary structure it resembles more closely the *in vivo* targets of hnRNP-E and hnRNP-K than it does the 10021 hairpin of the *in vitro* selection experiments. It is likely that recognition of small RNA target sequences by KH domains *in vivo* involves several strategies. First might be the coordinate use of several KH domains within the same protein to select several small sequence elements in a longer RNA sequence. High affinity binding would be achieved only if several KH domains are able to bind to their individual core targets. Second, sequences such as the $(UCAU)_3$ repeat of the α2 glycine receptor subunit, or the 19-base element in the 15-lipoxygenase 3′ UTR, might be recognized by a dimer or tetramer of a KH domain protein, which is rendered energetically favorable by multiple KH-RNA interactions, and by protein dimerization. Finally, KH domain proteins might participate in protein-protein interactions with other RNA binding proteins, with recognition achieved coordinately by the two (or several) RNA binding proteins.

We thank B. Kate Dredge for critical reading of the manuscript and all members of our laboratories for help and many useful discussions. This work was supported in part by the National Institutes of Health (R.B.D.). K.B.J. was supported by a Breast Cancer Research Program fellowship DAMD 17-97-1-7097. K.M. was supported by the Tri-Institutional M.D.-Ph.D. program and National Institutes of Health Medical Scientists Training Program Grant GM07739.

1. Siomi, H., Matunis, M. J., Michael, W. M. & Dreyfuss, G. (1993) *Nucleic Acids Res.* **21,** 1193–1198.
2. Dejgaard, K. & Leffers, H. (1996) *Eur. J. Biochem.* **241,** 425–431.
3. Musco, G., Stier, G., Joseph, C., Castiglione Morelli, M. A., Nilges, M., Gibson, T. J. & Pastore, A. (1996) *Cell* **85,** 237–245.
4. Buckanovich, R. J., Yang, Y. Y. & Darnell, R. B. (1996) *J. Neurosci.* **16,** 1114–1122.
5. Baber, J. L., Libutti, D., Levens, D. & Tjandra, N. (1999) *J. Mol. Biol.* **289,** 949–962.
6. Lewis, H. A., Chen, H., Edo, C., Buckanovich, R. J., Yang, Y. Y., Musunuru, K., Zhong, R., Darnell, R. B. & Burley, S. K. (1999) *Structure (London)* **7,** 191–203.
7. Min, H., Turck, C. W., Nikolic, J. M. & Black, D. L. (1997) *Genes Dev.* **11,** 1023–1036.
8. Arning, S., Grüter, P., Bilbe, G. & Kramer, A. (1996) *RNA* **2,** 794–810.
9. Engebrecht, J. A., Voelkel-Meiman, K. & Roeder, G. S. (1991) *Cell* **66,** 1257–1268.
10. Siebel, C. W., Admon, A. & Rio, D. C. (1995) *Genes Dev.* **9,** 269–283.
11. Gamarnik, A. V. & Andino, R. (1997) *RNA* **3,** 882–892.
12. Ostareck, D. H., Ostareck-Lederer, A., Wilm, M., Thiele, B. J., Mann, M. & Hentze, M. W. (1997) *Cell* **89,** 597–606.
13. Kiledjian, M., Wang, X. & Liebhaber, S. A. (1995) *EMBO J.* **14,** 4357–4364.
14. Deshler, J. O., Highett, M. I., Abramson, T. & Schnapp, B. J. (1998) *Curr. Biol.* **8,** 489–496.
15. Ross, A. F., Oleynikov, Y., Kislauskis, E. H., Taneja, K. L. & Singer, R. H. (1997) *Mol. Cell. Biol.* **17,** 2158–2165.
16. Pieretti, M., Zhang, F., Fu, Y., Warren, S., Oostra, B., Caskey, C. & Nelson, D. (1991) *Cell* **66,** 817–822.
17. DeBoulle, K., Verkerk, A., Reyniers, E., Vits, L., Hendrickx, J., Van Roy, B., Van Den Bos, F., de Graaff, E., Oostra, B. & Willems, P. (1993) *Nat. Genet.* **3,** 31–35.
18. Siomi, H., Choi, M., Siomi, M., Nussbaum, R. & Dreyfuss, G. (1994) *Cell* **77,** 33–39.
19. Buckanovich, R. J., Posner, J. B. & Darnell, R. B. (1993) *Neuron* **11,** 657–672.
20. Yang, Y. Y. L., Yin, G. L. & Darnell, R. B. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 13254–13259.
21. Ostareck-Lederer, A., Ostareck, D. H. & Hentze, M. W. (1998) *Trends Biochem. Sci.* **23,** 409–411.
22. Tuerk, C. & Gold, L. (1990) *Science* **249,** 505–510.
23. Ellington, A. & Szostak, J. (1990) *Nature (London)* **346,** 818–822.
24. Buckanovich, R. J. & Darnell, R. B. (1997) *Mol. Cell. Biol.* **17,** 3194–3201.
25. Jensen, K. B., Dredge, B. K., Steffani, G., Zhong, R., Buckanovich, R. J., Okano, H. J., Yang, Y. Y. L. & Darnell, R. B. (2000) *Neuron* **25,** 359–371.
26. Lewis, H. A., Musunuru, K., Jensen, K. B., Edo, C., Chen, H., Darnell, R. B. & Burley, S. K. (2000) *Cell* **100,** 323–332.
27. Carey, J., Cameron, V., de Haseth, P. L. & Uhlenbeck, O. C. (1983) *Biochemistry* **22,** 2601–2610.
28. Irvine, D., Tuerk, C. & Gold, L. (1991) *J. Mol. Biol.* **222,** 739–761.
29. Cornish-Bowden, A. (1985) *Nucleic Acids Res.* **13,** 3021–3030.
30. Wang, X., Kiledjian, M., Weiss, I. M. & Liebhaber, S. A. (1995) *Mol. Cell. Biol.* **15,** 1769–1777.
31. Ostareck-Lederer, A., Ostareck, D. H., Standart, N. & Thiele, B. J. (1994) *EMBO J.* **13,** 1476–1481.
32. Kramer, A. (1992) *Mol. Cell. Biol.* **12,** 4545–4552.
33. Abovich, N. & Rosbash, M. (1997) *Cell* **89,** 403–412.
34. Berglund, J. A., Fleming, M. L. & Rosbash, M. (1998) *RNA* **4,** 998–1006.
35. Holcik, M. & Liebhaber, S. A. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 2410–2414.
36. Berglund, J. A., Chua, K., Abovich, N., Reed, R. & Rosbash, M. (1997) *Cell* **89,** 781–787.