Published in final edited form as:

J Mol Biol. 2009 July 17; 390(3): 428–442. doi:10.1016/j.jmb.2009.04.048.

Secondary structures for 5' regions of R2 retrotransposon RNAs reveal a novel conserved pseudoknot and regions that evolve under different constraints

Elzbieta Kierzek^{1,2}, Shawn M. Christensen^{3,4}, Thomas H. Eickbush³, Ryszard Kierzek², Douglas H. Turner^{1,5}, and Walter N. Moss¹

¹Department of Chemistry, University of Rochester, RC Box 270216, Rochester, NY 14627-0216, USA ²Institute of Bioorganic Chemistry, Polish Academy of Sciences, 60-714 Poznan, Noskowskiego 12/14, Poland ³Department of Biology, University of Rochester, RC Box 270211, Rochester, NY 14627, USA ⁴Department of Biology, University of Texas at Arlington, Arlington, TX 76019 ⁵Department of Pediatrics, University of Rochester School of Medicine and Dentistry, Rochester, NY 14642, USA

Summary

Sequences from the 5' region of R2 retrotransposons of four species of silk moth are reported. In *Bombyx mori*, this region of the R2 messenger RNA contains a binding site for R2 protein and mediates interactions critical to R2 element insertion into the host genome. A model of secondary structure for this RNA fragment is proposed on the basis of binding to oligonucleotide microarrays, chemical mapping, and comparative sequence analysis. Five regions of conserved secondary structure are identified, including a novel pseudoknot. There is an apparent transition from an entirely RNA structure coding function in most of the 5' segment of the fragment to a protein coding function in the 3' segment. This suggests that regions evolved under separate functional constraints (structural, coding, or both).

Keywords

R2 element; RNA secondary structure; microarray; comparative structure analysis; silk moth

Introduction

R2 retrotransposons are widely distributed throughout arthropods and have been a persistent component of insect genomes since pre-Cambrian times. The R2 element reproduces itself by site specific insertion into the host genome 28S rRNA locus.¹ Insertion occurs via target primed reverse transcription, which has been extensively studied in the silk moth, *Bombyx mori*.^{2,3} Insertion proceeds by an ordered set of strand cleavage and templated DNA synthesis mediated by an RNA-protein complex composed of R2 mRNA with two copies of R2 protein.^{4,5} The mRNA-protein complex recognizes the insertion site whereupon the R2 protein bound to the

Correspondence to: Ryszard Kierzek; Douglas H. Turner.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

3' end of the R2 RNA template nicks one DNA strand. The freed 3'-OH group provides a primer for the reverse-transcriptase domain of the same R2 protein to generate cDNA.² The R2 protein bound to the 5' end of the RNA template then cleaves the opposing target strand⁶ and functions as a DNA-dependent DNA polymerase, using the newly synthesized cDNA as template.⁴

The role of RNA secondary structure in this fascinating and complex mechanism of transposition has yet to be elucidated. The 3' UTR binding site for the R2 protein has a conserved secondary structure.⁷ A secondary structure for the 5' R2 protein binding site, which occurs 650 bases downstream of the 5' terminus in *Bombyx mori* R2 RNA (R2Bm), has been proposed on the basis of binding to an isoelectric oligonucleotide microarray, chemical mapping, and free energy minimization.⁸ It was suggested that a segment of 74 nucleotides could fold into a complex pseudoknot and this was subsequently supported by NMR spectra.⁹ Pseudoknots are rare structural motifs that often play important roles in biological processes such as gene expression and replication.¹⁰⁻¹⁴ Here, we report further study of the pseudoknot and other aspects of the RNA secondary structure in the 5' R2 protein binding site. Comparison of alignments for R2 RNA and predicted R2 open reading frames suggests a transition between regions that evolved to preserve only RNA secondary structure and only protein sequence. The alignments also suggest an unusual mode of translation initiation.

Results

Structure probing

R2 5' RNAs from *Samia cynthia* (R2Sc), *Coscinocera hercules* (R2Ch), *Callosamia promethea* (R2Cpr), and *Saturnia pyri* (R2Spy) were probed with isoelectric microarrays designed and built as described in Materials and Methods. Typical data are shown in Figure 1, and detailed results are provided in Supporting Information. Each probe is identified by a number corresponding to the number of the R2 5' RNA nucleotide in the middle of the sequence Watson-Crick complementary to the first five nucleotides of the probe. In Supporting Information, a second binding site with equal complementarity to a probe is listed in parentheses. Hexanucleotide probes all have a 3' terminal LNA G.

R2 5' RNAs were chemically probed in 0.2 Na⁺/5 Mg²⁺/10T buffer at room temperature (see Materials and Methods and Supporting Information). Mapping with N-methylisatoic anhydride (NMIA), which identifies flexible sugars, is largely in agreement with results from DMS and CMCT.

Structure modeling

Minimum free energy (MFE) structures satisfying experimental microarray and chemical mapping constraints were computed with *RNAstructure 4.6*¹⁵ for the four new R2 sequences (Supporting Information). Five hairpins are conserved between each structure except for R2Ch, where only two of the conserved hairpins are predicted. The five conserved features are also present in the R2Bm secondary structure published previously.⁸ Outside of these hairpins, there is no apparent conservation of pairing in these MFE structures.

In subsequent modeling steps, the five common features persist with slight modifications and R2Ch was refolded to possess the same set of hairpins. Figure 2 presents the structure alignment (see Supporting Information for common secondary structure). Figures 3 through 7 show the final models along with constraints used in the two stages of modeling. Partition function calculations with constraints from chemical and microarray mapping (Figures 3-7) provide a measure of the certainty of each base pair. The optimized structural models have from 59.0% to 64.9% of the nucleotides in R2 5' RNA involved in base pairs. The percentage of nucleotides involved in conserved pairing ranges from 43.8% to 44.6%, which accounts for on average

72.3% of all base pairs (Table 1). This conserved secondary structure is arranged as four hairpins and a pseudoknot. There is evolutionary support for this structure: on average 20.1% of nucleotides are in conserved base pairs that have compensatory mutations, which accounts for 33.4% of all base pairs (Table 1).

Comparison of microarray and chemical mapping results with generated structures and structure alignment

With the exceptions of probe 246 for R2Spy and probes 227 and 228 for R2Bm, there is good agreement between generated structures for all R2 5' RNAs and oligonucleotide binding sites (Figures 2-7 and Supporting Information) revealed by isoenergetic microarray mapping. Probe 246 (DC^LUC^LCg^L) for R2Spy can bind to alternative sites 46 and 72 because it is a hexamer with a 5'-terminal 2,6 diaminopurine, D, rather than a 5'-terminal A as in the pentamer probe (AC^LCC^LC) for 46 and 72, which does not bind 5' R2 RNA. Probes 227 and 228 for R2Bm may bind by slipping the helix shown in Figure 3. Interestingly, on the structure alignment (Figure 2), the regions in alignment positions: 10-23 and 256-257 have similar accessibility to oligonucleotide probes. This suggests these regions are open to solvent and not involved in tertiary interactions.

With the exception of NMIA reactivity for nt 65 in R2Sc and nt 245 in R2Spy, modeled structures are consistent with the chemical modification data, i.e. strongly reactive nucleotides are not involved in Watson-Crick pairs flanked by Watson-Crick pairs. DMS, CMCT and NMIA mapping show similarity in several fragments. Strongly modified sites are in alignment regions: 59-63 (R2Bm, R2Spy, R2Sc, R2Cpr); 81-83 (R2Sc, R2Ch, R2Cpr); 94 (R2Sc, R2Cpr, R2Spy); 124-125 (R2Sc, R2Cpr, R2Spy); 221 (R2Sc, R2Ch, R2Spy); 222 (R2Sc, R2Ch, R2Cpr); 240-241 (R2Bm, R2Sc, R2Cpr); and 302-303 (R2Ch, R2Cpr, R2Spy). Many nucleotides were modified for all RNAs in the region spanning 219-230, especially for R2Sc and R2Ch.

Discussion

Little is known about the molecular details for site specific insertion of R2 retrotransposon sequences into the genomes of hosts. One copy of R2 protein binds tightly to the region of the R2 RNA studied here, and this copy of R2 protein is responsible for second strand cleavage of the host DNA.⁶ A model for the secondary structure of this region of R2 RNA in *B. mori* has been previously deduced on the basis of chemical modification and microarray binding.⁸ Here, the model is refined by comparison to sequences, chemical modification, and microarray binding for four additional 5' R2 RNAs. Comparisons between secondary structures for all five sequences suggest that different regions evolve under differing constraints and that initiation of R2 protein synthesis may be unusual. The results also provide insights into methods for determining secondary structures of new RNAs.

The alignment of primary and secondary structure shown in Figure 2 reveals several features (a consensus structure drawing is also provided in Supporting Information). The conserved regions of secondary structure are separated by stretches of sequence where structure is not conserved. Primary sequence is least conserved in the regions where structure is not conserved. The regions outside conserved structures are also prone to large sequence insertions and deletions (indels). Such indels are especially common in alignment positions 1 to 48 and 276 to 302 (Figure 2). When mutations occur in the structured regions they are almost always compensatory or consistent mutations (i.e., double or single point mutations that preserve canonical pairing). Out of a total of 75 conserved pairs, 34 pairs are observed to undergo compensatory mutations that preserve Watson-Crick base pairing. In another 11 of the paired sites there are consistent mutations, where a point mutation occurs on one RNA strand such that canonical base pairing is preserved (Figure 2). The number and distribution of

compensatory and consistent mutations give good support for the modeled secondary structures. Three compensatory mutations per helix is one common criterion for proving a helix phylogenetically.¹⁶ At least one helix in each of the conserved structures meets this condition (Figure 2), and on average 33.4% of all base pairs have support from compensatory mutations. Where this condition is not met, it is typically because there is no variation in the sequence (the pair is absolutely conserved) or mutations are observed to occur only on one strand (consistent mutations).

The R2Bm modeled structure is consistent with the secondary structure previously proposed.⁸ Only two differences appear: nucleotides 50 to 124 are folded into a pseudoknot and nucleotides 221 and 226 form a GC pair in the revised model. The possibility of R2Bm having a pseudoknot was suggested previously⁸ and has been supported by NMR on a 74 nucleotide fragment equivalent to nucleotides 50-123 of *B. mori* 5' R2 RNA (Figure 3) but with base pair 51-105 flipped from CG to GC.⁹ Sequence comparison confirms the pseudoknot and extends it to include a G81 to C124 base pair.

Some trends in susceptibility to chemicals and/or oligonucleotide binding can be discerned. Alignment positions 3 to 33 show great susceptibility to oligonucleotide binding (Figure 2 and Supporting Information). Oligonucleotide binding sites are dispersed across the sequence and do not occur in the same aligned sequence positions. Chemical reactivity was often observed between alignment positions 194 to 234. These reactive sites are dispersed in positions 194 to 234 and do not appear to be conserved. In the regions of conserved structure, there are apparent similarities in reactivity profiles; for instance, the hairpin loop of P1 shows chemical modifications in two sequences at alignment position 60 and in four sequences at position 62. Similar conservation of reactivity or binding is also observed in two of the loop regions of the pseudoknot, J2/3 and the hairpin loop of P3 (alignment positions 81-83 and 93-96), and in the hairpin loop of P7b (alignment positions 256 and 257). UUCG loops in the structures typically react with NMIA, usually at the C ribose. The hairpin loop of P6 (alignment positions 173-175) does not show reactivity. Surprisingly few nucleotides modeled as single stranded are strongly reactive, suggesting a compact three-dimensional structure.

The partition function for the secondary structure of an RNA provides a measure of the confidence of forming a base pair.¹⁷ For example, comparing partition function calculations to a database of known secondary structures showed that considering structures formed from pairs with greater than 99% probability increased the average positive predictive value (PPV) from 65.8% to 91.0%. In the set of R2 RNAs, there is a strong correlation between the level of structure conservation and the base pair probability calculated for each sequence individually with constraints from strong chemical mapping and microarray binding (Figures 3-7). The partition function calculation found the majority of conserved stems to be composed of pairs with pairing probability typically greater than 99%. When the conserved pairs drop below this level, they are still of a higher probability than non-conserved pairs. This was true even when partition function calculations were run without experimental constraints (data not shown). There are some regions, however, where conserved pairs do not have high probability. The base pairs in the internal helices of the conserved pseudoknot are mostly found to have below 50% probability, except for several helices in R2Sc, R2Bm, and R2Spy. The partition function calculation is expected to under-predict probabilities for base pairs in the pseudoknot, as the algorithm forbids non-nested pairs from appearing in the same structure. In R2Ch, the conserved hairpin stem formed between nucleotides 38-58 has base pair probabilities below 50% as are those in nucleotides 281-312 of R2Cpr.

Base pairs that were not conserved typically had probabilities below 50%. Some regions of the non-conserved structure show high probability in specific RNAs, however. One example is the extended basal stem of the terminal 3' conserved hairpin. In four of the five species (all but

R2Spy), the 3' conserved hairpin sits atop a long basal stem, P8.0, of 11 to 15 base pairs. These extended stems do not align, but the structure appears in at least four R2 elements, is consistent with experiment, and is composed of highly probable base pairs. These may be species specific structural elements or it could be that these structures are composed of nucleotides that are not homologous. That is, each of the four R2 elements has utilized a different set of bases to form these structures.

There are limits to the prediction improvement possible by experimentally restraining MFE structures. Using strong chemical mapping constraints as well as strong and moderate microarray constraints led to structures that were more consistent with the final model in two out of five cases (Table 2). In the other three cases, these constraints had essentially no effect on the structure (with slight improvement for R2Sc and slight decrease in R2Spy). Expanding the number of chemical modification constraints to include medium strength hits typically had no effect, except in R2Cpr, where including medium strength constraints decreased the improvement from only using strong chemical modification constraints, although there was still improvement vs the MFE structure. A major limitation is that *RNAstructure* does not predict pseudoknotted helices, which are forbidden in most dynamic programming algorithms. As shown in Table 2, however, chemical mapping can sometimes improve predictions even when a large pseudoknot is present. NMR is able to identify pseudoknotted pairs⁹ but cannot currently be applied to RNAs of the size of these R2 element fragments. As for phylogenetic prediction methods, there were too few sequences to model the structure of the roughly 320 nucleotide RNA by sequence comparison alone. Sequence comparison does provide the evidence showing that the pseudoknot forms as a domain in the large RNAs. Figure 8 summarizes this evidence. The pseudoknot is unusual because loops 1 and 2 each contain hairpins with tetraloops having UNCG, GNRA, and UUUU sequences that are often seen in other types of RNA.¹⁸⁻²¹

While no single method was sufficient for finding the secondary structure, the combination of methods provided tools sufficient to deduce the secondary structures of these RNAs. Constraints from chemical mapping, NMR, and microarray binding gave an initial set of structures that were thermodynamically feasible and consistent with experimental data. These results were essential as a guide for the phylogenetic modeling steps that led to the final structures. Chemical mapping and microarray binding delimited regions that were single stranded or paired in a weak context, while NMR data identified paired nucleotides in the pseudoknotted region of R2Bm 5' RNA; the phylogenetic modeling "locked in" the conserved pairs in the RNA structures.

In addition to encoding RNA secondary structure, these molecules also encode the R2 protein and evolve under this constraint as well. Figure 9 shows an alignment of the R2 open reading frames (ORFs) with conserved secondary structure and primary sequence annotation for the RNA. Earlier experiments, as in Figure 9, used alignments to predict the start of R2 open reading frames by scanning for potential initiation codons upstream of the highly conserved zinc finger domain.^{22,23} Potential start codons are present in alignment region 208 to 231 in four of the five sequences (Figure 9).²² The first conservation of amino acid sequences is in alignment region 244 to 246, a region that is also involved in conserved RNA secondary structure. As evidence that the encoded amino acid sequence is being conserved, and not just a conservation of RNA structure, five synonymous substitutions have accumulated in this region and only two non-synonymous substitutions.

Immediately upstream (i.e. alignment region 199-243) of the conserved protein region, the reading frame is conserved but the protein sequence is not (16 synonymous substitutions and 42 non-synonymous substitutions). Upstream of alignment position 199 there are frequent stop codons and changes in reading frames, suggesting that this region is no longer protein encoding.

Thus, it would appear that the protein encoding region begins in the region from alignment position 199 to 244. Methionine (Met) initiation codons can be found in this region in four species but they are not conserved in position. The first Met codon of the R2Bm ORF is upstream of the sequence listed in Figure 9 and is preceded by non-productive start and stop codons (data not shown).

This ORF data is similar to earlier reports in that non productive start and stop codons often precede the presumptive start of the ORF, requiring the translational machinery to bypass these non-productive codons.^{22,23} R2 elements lacking Met initiation codons have also been described.²³ In addition, it is believed that the R2 RNA is co-transcribed with ribosomal RNA by RNA Pol I, and then processed to element sized RNA.²⁴ For this reason, R2 transcripts are unlikely to behave as canonical Pol II mRNAs during translation. Taken together, the data suggests the possibility that R2 elements utilize a non-canonical translation initiation mechanism involving some type of internal ribosome capturing/initiation mechanism as has recently been observed in certain viruses.^{25,26} Perhaps the RNA structures in region 244 to 267, or elsewhere, serve “double duty” in binding to the R2 protein and interacting with the translational machinery to initiate translation. It is possible that the pseudoknot is involved in a non-canonical initiation of translation. Pseudoknots often form tRNA-like structures (TLS).²⁷⁻²⁹ Such structures can interact with the ribosome.^{25,26,30,31}

In the region spanning positions 276 to 332, a single alignment is not able to align both the RNA secondary structure and amino acid sequence. The alignment of this region shown in Figure 9 can be explained by two events: a deletion of one amino acid in R2Spy (277-279) and an insertion of valine in R2Bm (316-318); however, conservation of the RNA hairpin requires a more complicated pattern of insertions and deletions. This can be interpreted as a reorganization of the secondary structure of R2Bm 5' RNA in response to perturbations of sequence that primarily manifest themselves on the protein level. The RNA secondary structure is conserved in the five R2 elements, but it appears that stem P8 is shifted to utilize a different set of bases. Indeed, this hairpin has an extended basal stem (P8.0) in R2Bm, R2Sc, R2Ch, and R2Cpr that is well represented in the partition function, but is not alignable. These could be homologous structures that are composed of non-homologous nucleotides.

There is no conserved RNA structure beyond alignment position 332 in Figure 9 (position 335 in Figure 2). This region codes for the cysteine motif, which is universally conserved across all known R2 elements. This protein motif is a known zinc finger motif that binds DNA.³² Across the alignment, conservation patterns transition from acting solely on the level of RNA secondary structure, to also include the amino acid sequence of the encoded R2 protein. This pattern of conservation is illustrated in Figure 10, which shows the five conserved RNA structures as well as the coding region. The putative ORF start site begins after the third conserved RNA structure (the hairpin formed from stem P6). The next two conserved hairpins overlap with two conserved regions in the amino acid sequence (A and B).

Additional evidence for the importance of this fragment of the R2 element comes from the discovery of small antisense RNAs expressed in *B. mori* ovarian tissue.³³ One antisense RNA complementary site spans the longer of the two pseudoknotted helices (P4 in Figure 3) as well as part of P3. The binding of this antisense RNA to the R2 transcript would completely abolish the pseudoknot structure. Because small RNAs, particularly of the piwi class are proposed to be a defense system against transposons, here is a feasible mechanism by which it may act on the level of RNA secondary structure. A pseudoknot may be an attractive target for antisense because pseudoknots often fold slowly.³⁴⁻³⁷

Conclusions

A secondary structure model of the 5' R2 protein binding site of five silk moth R2 elements has been constructed from free energy minimization, isoenergetic microarray binding, chemical mapping, NMR spectroscopy, and comparative sequence analysis. The final structural model is robust and yields insight into the biological function and molecular evolution of these sequences.

Analyses of these sequences with respect to the RNA secondary structure model and ORF sequence suggest an unusual mechanism of translation that may bypass in-frame stop codons. The sequence comparisons also suggest that during evolution different segments of 5' R2 RNA sequences had to satisfy different constraints, either individually or in synchronization.

Materials and Methods

Materials

The 5' end of the silkmoths *S. cynthia* (R2Sc), *C. hercules* (R2Ch), *C. promethea* (R2Cpr), and *S. pyri* (R2Spy) were obtained by PCR using a common primer complementary to the 28S gene sequences upstream of the R2 element (GGGTAAACGGCGGGAGTAAGTATGACTC) and specific primers to regions near the 3' end of R2 from each species. These 3' end sequences had been previously obtained by Ruschak et al. (2004). The primers used were GCCCGAGAAACCAACAGGATGATCGG for R2Sc, GGCAACCCACTCACAGGATCTTCGG for R2Cpr, GCAGCCCACACAGGATCTTCGG for R2Spy and GGGTAAACGGCGGGAGTAAGTATGACTC for R2Ch. The PCR fragments were band purified, TA cloned and sequenced. The *B. mori* sequence has been reported (R2Bm; pubmed ID: M16558).

A, C, G, and U 2'-O-methyl RNA phosphoramidites and C6-aminolinker for oligonucleotide synthesis were purchased from Glen Research and Prologo. Phosphoramidites of LNA, LNA 2,6-diaminopurine riboside, and 2'-O-methyl-2,6-diaminopurine riboside were synthesized as described previously.³⁸

Restriction enzymes, XcmI and SacI, were purchased from New England BioLabs. Taq polymerase was a product of Promega. AmpliScribe T7-Flash transcription kit was from Epicentre. DNA oligonucleotides for PCR and *in vitro* reverse transcription were purchased from Integrated DNA Technologies.

The γ ³²P-ATP was purchased from Perkin Elmer. Reverse transcriptase SuperScript III, T4 polynucleotide kinase, agarose, and HybriSlip hybridization covers were from Invitrogen; dNTPs and ddNTPs were from Amersham Biosciences. Dimethyl sulfate (DMS) and 1-cyclohexyl-3-(2-morpholinoethyl)carbodiimide metho-p-toluenesulfonate (CMCT) were from Aldrich. N-methylisatoic anhydride (NMIA) was from Molecular Probes. Silanized slides were purchased from Sigma. Humidity-temperature chambers for single hybridization were manufactured in-house.

Chemical synthesis of modified oligonucleotide probes

Oligonucleotide probes for microarrays were synthesized with a C6-aminolinker on the 5'-end by the phosphoramidite method on an ABI 392 synthesizer. Probes were de-protected and purified according to published procedures.^{8,39} Molecular weights were confirmed by mass spectrometry (LC MS Hewlett Packard series 1100 MSD with API-ES). Concentrations of all oligonucleotides were determined from predicted extinction coefficients for RNA and measured absorbance at 260 nm at 80 °C. It was assumed that 2'-O-methyl RNA-LNA chimeras and RNA strands with identical sequences have identical extinction coefficients.

Preparation of isoenergetic microarrays

To take advantage of sequence homology between R2 5' RNAs, a semi-universal isoenergetic microarray was designed that contains probes complementary to the four new R2 5' RNAs. Probes were 2'-O-methyl oligonucleotide pentamers and hexamers with incorporated LNA nucleotides and 2,6-diaminopurine riboside (LNA or 2'-O-methyl). Modifications were chosen such that the vast majority are predicted to have free energies for binding at 37°C (ΔG°_{37}) to unstructured RNA of between -8.0 and -10.5 kcal/mol (Supporting Information) with a distribution similar to that used to study the *B. mori* 5' RNA.⁸ There are 468 specific probes and 12 negative control probes printed in 48 blocks on the microarray. Each probe was spotted in triplicate, with a spot distance of 750 μ m. The sequence UUUUU and spotting buffer were used as negative controls. Microarrays were prepared according to the method described earlier.⁸ Silanized slides were coated with 2% agarose activated by NaIO₄ and probes were spotted by a microarray printer (MicroGridII TAS Arrayer at the Microarray and Genomics Facility, Roswell Park Cancer Institute, Buffalo, NY). Printed microarrays were incubated for 12 h at 25 °C in 50% humidity. The remaining aldehyde groups on microarrays were reduced with 35 mM NaBH₄ solution in PBS buffer (137 mM NaCl, 2.7 mM KCl, 4.3 mM Na₂HPO₄, 1.4 mM KH₂PO₄, pH 7.5) and ethanol (3:1 v/v). Then slides were washed in water at room temperature (3 washes for 30 min each), and in 1% SDS solution at 55 °C for 1 h, and finally in water at room temperature (3 washes for 30 min each) and dried at room temperature overnight.

Synthesis and purification of 5' RNA from silk moth R2 RNA

R2 5' RNA from *S. cynthia*, *C. hercules*, *C. promethea*, and *S. pyri* were obtained by *in vitro* transcription. DNA templates were generated by PCR using primers containing a T7 promoter sequence. Primer sequences are given in Supporting Information. PCR reactions of 50 μ l were run with 70 ng of DNA template, 10 pmol of each primer and 2.5 units of Taq DNA polymerase, following the Promega protocol. R2 5' RNAs were transcribed from 1 μ g of PCR template using an AmpliScribe T7-Flash transcription kit.

Buffers, folding method and native gel electrophoresis

Native gel electrophoresis indicated that R2 5' RNAs form a single structure under various folding conditions. R2 5' RNAs labeled with ³²P on the 5'-end were purified on an 8% polyacrylamide denaturing gel. Each RNA was subsequently incubated for 5 min at 65°C and slowly cooled to room temperature in folding buffers: (1) 200 mM NaCl, 5 mM MgCl₂, 10 mM Tris-HCl, pH 8.0 (0.2 Na⁺/5 Mg²⁺/10T) or (2) 1 M NaCl, 5 mM MgCl₂, 10 mM Tris-HCl, pH 8.0 (1 Na⁺/5 Mg²⁺/10T). After folding, samples were analyzed (15,000 cpm per lane) on a 27 cm, 4% polyacrylamide non-denaturing gel containing Tris-borate-EDTA, pH 8.0 with a running buffer of Tris-borate-EDTA (89 mM Tris, 89 mM boric acid, 2 mM sodium EDTA, pH 8.0). The gel was pre-electrophoresed at 20 W for 0.5 h. Electrophoresis was performed at 20 W and 4 °C until bromophenol blue migrated within 18 cm of the base of the gel. Dried gels were analyzed by exposing to X-ray film and also to a phosphorimager screen. For each R2 5' RNA, both folding conditions yield a single band.

Binding to microarrays

For hybridization, ~10 nM of ³²P labeled R2 5' RNA in 250 μ l of folding buffer was placed on a microarray slide and covered with a HybriSlip. Each microarray slide was placed in a humidity-temperature chamber with 100% humidity and incubated for 18 h at 4 °C. After hybridization, buffers with RNA were poured out and slides were washed in buffers with the same salt concentrations for 1 min at 0 °C. Longer washing times (3-10 min) did not change relative strength of binding signals. Slides were dried by centrifugation in a clinical centrifuge (2000 rpm, 2 min) and covered with plastic wrap. Hybridization was visualized by exposure

to a phosphorimager screen (Dynamics 840 Storm Phosphorimager). Quantitative analysis was performed with ImageQuant 5.2. Binding was considered strong, medium, or weak when the integrated intensity was $\geq 1/3$, $\geq 1/9$, or $\geq 1/27$ of the strongest integrated intensity, respectively. The integrated intensity was calculated as the intensity of a probe spot minus the local background intensity. Experiments were repeated at least three times and the average of the data is presented. Generally, more probes bind strongly in 1 Na⁺/5 Mg²⁺/10T than in 0.2 Na⁺/5 Mg²⁺/10T and usually the same probes bind in both 0.2 Na⁺/5 Mg²⁺/10T and 1 Na⁺/5 Mg²⁺/10T.

Chemical mapping

Dimethyl sulfate (DMS) was used to modify adenosine and cytidine, CMCT to modify uridine and guanosine⁴⁰, NMIA to modify flexible riboses.⁴¹ For each reaction, 1 pmol of RNA was folded in 0.2 Na⁺/5 Mg²⁺/10T buffer, as described above. Then tRNA carrier was added to give a total RNA concentration of 8 μ M and the solution was incubated for 10 min at room temperature. To a 9 μ l sample, 1 μ l of 600 mM DMS in ethanol was added to give a final concentration of 60 mM DMS. For modification with CMCT, 9 μ l of CMCT solution was added to the 9 μ l of RNA sample. CMCT was diluted in an appropriate buffer to give a final concentration of 625 mM in the reaction mixture. Chemical modification reactions were performed for 20 min at room temperature. Reactions were stopped by ethanol precipitation on dry ice. Chemical mapping with NMIA was done as described earlier with some changes.⁴¹ For each reaction, 1 pmol of RNA was taken and refolded as described above. To a 9 μ l sample, 1 μ l of NMIA solution (1 mg NMIA/42 μ l DMSO) was added. Samples were incubated for 3 h at room temperature. Reactions were stopped by ethanol precipitation on dry ice.

Primer extension reactions

Modification sites were identified by primer extension. Primers were 5' end labeled with γ ³²P-ATP according to standard procedures. For each reaction, 1 pmol of primer was used. Primer extension was performed at 55 °C with SuperScript III reverse transcriptase following Invitrogen's protocol. Reactions were stopped by adding loading buffer containing formamide and 10 mM EDTA, then chilling to 0 °C. Products were heated for 5 min at 95 °C and then separated on a 12% polyacrylamide denaturing gel. Products were identified by comparing to sequencing lanes and to control lanes. Modifications were initially identified by visual inspection of autoradiograms and were considered strong or medium when the band corresponding to chemical modification had at least 6 times, or 2-6 times, respectively, the integrated intensity of the equivalent band in the control lane, as quantified with the ImageQuant 5.2 program. DNA sequences complementary to the R2Spy sequence 55-75 primed reverse transcription significantly less efficiently than other primers used for this RNA. All primers used for reverse transcription are listed in the Supporting Information.

Structure modeling with experimental constraints and sequence alignment

For each of the four new R2 sequences, minimum free energy (MFE) structures were computed with *RNAstructure* 4.6 with and without constraints from chemical modification and microarray binding in 0.2 Na⁺/5 Mg²⁺/10T buffer.^{15,42} Only nucleotides that were strongly modified by CMCT, DMS, and/or NMIA were used as constraints in the prediction of the secondary structures. Microarray data were screened to eliminate probes that could have alternative binding sites. For probes providing signal intensities $\geq 1/9$ of the maximum signal intensity, potential alternative binding sites were predicted with the bimolecular mode of *RNAstructure* 4.5 assuming no intramolecular base pairing in the R2 RNA or probe. Probes with more than one site exactly complementary to the first five nucleotides or with alternative binding sites having predicted free energy more favorable than -4.0 kcal/mol at 37 °C and that bind their exact complement with at least one third the signal of the probe being considered

were not used as constraints. Probes were also eliminated as constraints if they were hexamers and probes to alternative sites were pentamers whose binding would be increased by formation of a GC or GU pair if the probe was a hexamer with a 3' terminal G. A probe was also not used as a constraint if the bimolecular binding mode of *RNAstructure* predicted an alternative site less favorable than -4.0 kcal/mol but within 2 kcal/mol of the complementary site. Microarray constraints were used as if they were chemical modifications: the center nucleotide of strong ($\geq 1/3$ maximum intensity) and medium ($\geq 1/9$ maximum intensity) binding sites were forbidden to be in a Watson-Crick pair flanked by Watson-Crick pairs. For R2Bm, the chemical modification and microarray constraints were taken from a previous work⁸ and used to recalculate the MFE structure.

The four new R2 sequences and that of R2Bm were aligned with *ClustalX* v. 1.81 with the default nucleic acid parameters.⁴³ A PERL script was written to map secondary structures, in dot bracket notation, onto the sequence alignment and to format it for viewing in a spreadsheet. The initial structure alignment was manually optimized to maximize conservation of sequence and structure, which could be displayed in “real time” in the spreadsheet by counting the characters in a column. Several conserved structural features with the potential to form in all sequences were identified by inspection. Alignment positions 75 to 153 were modeled on the basis of the *B. mori* pseudoknotted structure consistent with imino proton NMR data.⁹ Potential structures were modified to contain conserved features and these helices were extended if additional canonical pairs could be made. Regions that were not apparently conserved were folded with *RNAstructure* with constraints on non-pseudoknotted conserved base pairs along with single strand constraints from microarray and chemical mapping data.^{17,44} Optimization of structure and sequence alignment continued through multiple iterations, until no further conserved features were found.

Unconstrained and experimentally constrained partition function calculations were run on each of the molecules using *RNAstructure* to calculate the probability of each base pair.¹⁷ These data were mapped onto the secondary structure models to provide an indication of the certainty for each base pair.

Sequences were translated in three open reading frames (ORFs) *in silico* and mapped onto the alignment. ORFs used were those that gave protein motifs known to be conserved in R2 elements.²² Additionally, protein alignments were performed with *ClustalX* v. 1.81 with the default protein parameters.⁴³

Representations of the pseudoknot were generated with the program *PseudoViewer*.^{45,46}

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH grants GM 22939 (D.H.T) and GM42790 (T.H.E) and the Ministry of Science and Higher Education grants N N301 3383 33 (E.K.) and 2 P04A 03729 (R.K.). We thank the Microarray and Genomics Facility, Roswell Park Cancer Institute, Buffalo, NY for printing microarrays. We also give gratitude to Biao Liu for assistance with figures.

References

1. Eickbush, TH. R2 and Related Site-specific non-LTR Retrotransposons. In: Craig, N.; C, R.; Gellert, M.; Lambowitz, A., editors. Mobile DNA II. American Society of Microbiology Press; Washington D.C: 2002. p. 813-835.

2. Luan DD, Korman MH, Jakubczak JL, Eickbush TH. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* 1993;72:595–605. [PubMed: 7679954]
3. Yang J, Malik HS, Eickbush TH. Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. *Proc Natl Acad Sci U S A* 1999;96:7847–52. [PubMed: 10393910]
4. Kurzynska-Kokorniak A, Jamburuthugoda VK, Bibillo A, Eickbush TH. DNA-directed DNA polymerase and strand displacement activity of the reverse transcriptase encoded by the R2 retrotransposon. *J Mol Biol* 2007;374:322–33. [PubMed: 17936300]
5. Christensen SM, Eickbush TH. R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. *Mol Cell Biol* 2005;25:6617–28. [PubMed: 16024797]
6. Christensen SM, Ye J, Eickbush TH. RNA from the 5' end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site. *Proc Natl Acad Sci U S A* 2006;103:17602–7. [PubMed: 17105809]
7. Ruschak AM, Mathews DH, Bibillo A, Spinelli SL, Childs JL, Eickbush TH, Turner DH. Secondary structure models of the 3' untranslated regions of diverse R2 RNAs. *RNA* 2004;10:978–87. [PubMed: 15146081]
8. Kierzek E, Kierzek R, Moss WN, Christensen SM, Eickbush TH, Turner DH. Isoenergetic penta- and hexanucleotide microarray probing and chemical mapping provide a secondary structure model for an RNA element orchestrating R2 retrotransposon protein function. *Nucleic Acids Res* 2008;36:1770–82. [PubMed: 18252773]
9. Hart JM, Kennedy SD, Mathews DH, Turner DH. NMR-assisted prediction of RNA secondary structure: identification of a probable pseudoknot in the coding region of an R2 retrotransposon. *J Am Chem Soc* 2008;130:10233–9. [PubMed: 18613678]
10. Brierley I, Pennell S, Gilbert RJ. Viral RNA pseudoknots: versatile motifs in gene expression and replication. *Nat Rev Microbiol* 2007;5:598–610. [PubMed: 17632571]
11. Wang Y, Wills NM, Du Z, Rangan A, Atkins JF, Gesteland RF, Hoffman DW. Comparative studies of frameshifting and nonframeshifting RNA pseudoknots: a mutational and NMR investigation of pseudoknots derived from the bacteriophage T2 gene 32 mRNA and the retroviral gag-pro frameshift site. *RNA* 2002;8:981–96. [PubMed: 12212853]
12. van Batenburg FH, Gulyaev AP, Pleij CW, Ng J, Oliehoek J. PseudoBase: a database with RNA pseudoknots. *Nucleic Acids Res* 2000;28:201–4. [PubMed: 10592225]
13. ten Dam EB, Pleij CW, Bosch L. RNA pseudoknots: translational frameshifting and readthrough on viral RNAs. *Virus Genes* 1990;4:121–36. [PubMed: 2402881]
14. Chen X, Chamorro M, Lee SI, Shen LX, Hines JV, Tinoco I Jr, Varmus HE. Structural and functional studies of retroviral RNA pseudoknots involved in ribosomal frameshifting: nucleotides at the junction of the two stems are important for efficient ribosomal frameshifting. *Embo J* 1995;14:842–52. [PubMed: 7882986]
15. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A* 2004;101:7287–92. [PubMed: 15123812]
16. Noller HF, Kop J, Wheaton V, Brosius J, Gutell RR, Kopylov AM, Dohme F, Herr W, Stahl DA, Gupta R, Woese CR. Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Res* 1981;9:6167–89. [PubMed: 7031608]
17. Mathews DH. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* 2004;10:1178–90. [PubMed: 15272118]
18. Woese CR, Winker S, Gutell RR. Architecture of ribosomal RNA: constraints on the sequence of “tetra-loops”. *Proc Natl Acad Sci U S A* 1990;87:8467–71. [PubMed: 2236056]
19. Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Muller KM, Pande N, Shang Z, Yu N, Gutell RR. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* 2002;3:2. [PubMed: 11869452]

20. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. *Nucleic Acids Res* 2003;31:439–41. [PubMed: 12520045]
21. Thomas J, Lea K, Zucker-Aprison E, Blumenthal T. The spliceosomal snRNAs of *Caenorhabditis elegans*. *Nucleic Acids Res* 1990;18:2633–42. [PubMed: 2339054]
22. Burke WD, Malik HS, Jones JP, Eickbush TH. The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. *Mol Biol Evol* 1999;16:502–11. [PubMed: 10331276]
23. George JA, Eickbush TH. Conserved features at the 5' end of *Drosophila* R2 retrotransposable elements: implications for transcription and translation. *Insect Mol Biol* 1999;8:3–10. [PubMed: 9927169]
24. Eickbush DG, Ye J, Zhang X, Burke WD, Eickbush TH. Epigenetic regulation of retrotransposons within the nucleolus of *Drosophila*. *Mol Cell Biol* 2008;28:6452–61. [PubMed: 18678644]
25. Stupina VA, Meskauskas A, McCormack JC, Yingling YG, Shapiro BA, Dinman JD, Simon AE. The 3' proximal translational enhancer of Turnip crinkle virus binds to 60S ribosomal subunits. *RNA* 2008;14:2379–93. [PubMed: 18824512]
26. McCormack JC, Yuan X, Yingling YG, Kasprzak W, Zamora RE, Shapiro BA, Simon AE. Structural domains within the 3' untranslated region of Turnip crinkle virus. *J Virol* 2008;82:8706–20. [PubMed: 18579599]
27. Pleij CW, Rietveld K, Bosch L. A new principle of RNA folding based on pseudoknotting. *Nucleic Acids Res* 1985;13:1717–31. [PubMed: 4000943]
28. Kolk MH, van der Graaf M, Wijmenga SS, Pleij CW, Heus HA, Hilbers CW. NMR structure of a classical pseudoknot: interplay of single- and double-stranded RNA. *Science* 1998;280:434–8. [PubMed: 9545221]
29. Hammond JA, Rambo RP, Filbin ME, Kieft JS. Comparison and functional implications of the 3D architectures of viral tRNA-like structures. *RNA* 2009;15:294–307. [PubMed: 19144910]
30. Barends S, Bink HH, van den Worm SH, Pleij CW, Kraal B. Entrapping ribosomes for viral translation: tRNA mimicry as a molecular Trojan horse. *Cell* 2003;112:123–9. [PubMed: 12526799]
31. Pflingsten JS, Kieft JS. RNA structure-based ribosome recruitment: lessons from the Dicistroviridae intergenic region IRESes. *RNA* 2008;14:1255–63. [PubMed: 18515544]
32. Berg JM. Potential metal-binding domains in nucleic acid binding proteins. *Science* 1986;232:485–7. [PubMed: 2421409]
33. Kawaoka S, Hayashi N, Katsuma S, Kishino H, Kohara Y, Mita K, Shimada T. Bombyx small RNAs: Genomic defense system against transposons in the silkworm, *Bombyx mori*. *Insect Biochem Mol Biol* 2008;38:1058–1065. [PubMed: 18801438]
34. Banerjee AR, Turner DH. The time dependence of chemical modification reveals slow steps in the folding of a group I ribozyme. *Biochemistry* 1995;34:6504–12. [PubMed: 7756281]
35. Pan J, Woodson SA. Folding intermediates of a self-splicing RNA: mispairing of the catalytic core. *J Mol Biol* 1998;280:597–609. [PubMed: 9677291]
36. Treiber DK, Rook MS, Zarrinkar PP, Williamson JR. Kinetic intermediates trapped by native interactions in RNA folding. *Science* 1998;279:1943–6. [PubMed: 9506945]
37. Zarrinkar PP, Williamson JR. Kinetic intermediates in RNA folding. *Science* 1994;265:918–24. [PubMed: 8052848]
38. Pasternak A, Kierzek E, Pasternak K, Turner DH, Kierzek R. A chemical synthesis of LNA-2,6-diaminopurine riboside, and the influence of 2'-O-methyl-2,6-diaminopurine and LNA-2,6-diaminopurine ribosides on the thermodynamic properties of 2'-O-methyl RNA/RNA heteroduplexes. *Nucleic Acids Res* 2007;35:4055–63. [PubMed: 17567607]
39. Xia T, SantaLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* 1998;37:14719–35. [PubMed: 9778347]
40. Ehresmann C, Baudin F, Mougél M, Romby P, Ebel JP, Ehresmann B. Probing the structure of RNAs in solution. *Nucleic Acids Res* 1987;15:9109–28. [PubMed: 2446263]
41. Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J Am Chem Soc* 2005;127:4223–31. [PubMed: 15783204]

42. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 1999;288:911–40. [PubMed: 10329189]
43. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22:4673–80. [PubMed: 7984417]
44. Kierzek E, Kierzek R, Turner DH, Catrina IE. Facilitating RNA structure prediction with microarrays. *Biochemistry* 2006;45:581–93. [PubMed: 16401087]
45. Byun Y, Han K. PseudoViewer: web application and web service for visualizing RNA pseudoknots and secondary structures. *Nucleic Acids Res* 2006;34:W416–22. [PubMed: 16845039]
46. Han K, Byun Y. PSEUDOVIEWER2: Visualization of RNA pseudoknots of any type. *Nucleic Acids Res* 2003;31:3432–40. [PubMed: 12824341]

Figure 1.

Example of typical microarray data. Hybridization results for *C. hercules* in 200 mM NaCl, 5 mM MgCl₂, 10 mM Tris-HCl, pH 8.0. Bar graph shows relative intensity of probes complementary to sites 3 to 327 in *C. hercules*. Full information about probes and hybridization results are in Supporting Information.

Figure 2. Optimized structure alignment of R2 5' RNAs. Conserved structures are highlighted in different colors, except in the pseudoknot, where the individual helices are colored for clarity. Base pairs are indicated with matching brackets, while unpaired nucleotides are indicated by "dots". Sites in each RNA that are strongly modified by DMS or CMCT are colored green and blue, respectively, in the nucleotide numbering, while sites that are strongly modified by NMIA are in red. Sites strongly modified by DMS or CMCT and by NMIA are orange. For R2Bm, sites strongly modified by kethoxal are also colored green. Strongly modified sites have ≥ 6 times the control intensity. Unambiguous microarray binding centers exhibiting medium ($\geq 1/9$ maximum intensity) or strong ($\geq 1/3$ maximum intensity) binding are boxed. Below the alignment, the consensus structure is indicated by "<" and ">" characters. Below that, sites with compensatory mutations are given by the boxed nucleotides and consistent mutations are

indicated by the unboxed nucleotides. Note that numbering for alignment positions is at the top and differs from the numbering of each RNA. For R2Sc, R2Ch, R2Cpr and R2Spy, the nucleotide labeled 2 corresponds to nucleotides 472, 734, 482, and 433, respectively, of the full-length R2 sequence (the first G was added for increased transcription efficiency). For R2Bm, the nucleotide labeled 1 is position 650 on the full-length R2 sequence.

Figure 3.

Structure model for *B. mori* 5' RNA (R2Bm). Structure is annotated for initial structural modeling constraints generated from strong ($\geq 1/3$ maximum intensity) or medium ($\geq 1/9$ maximum intensity) binding to oligonucleotide microarrays (boxed nucleotides), and from chemical mapping (green, blue, red, and orange circles, respectively, corresponding to strong reactivity, i.e. ≥ 6 times control intensity, with DMS, CMCT, NMIA, and CMCT or DMS overlapping with NMIA). Also annotated are base pair conservation (open boxes between nucleotides), compensatory mutations (filled boxes between nucleotides) and partition function probabilities (colored boxes or dashes between nucleotides). An annotation key appears on the figure.

Figure 4.
Structure model for *S. cynthia* 5' RNA (R2Sc). Structure is annotated as in Figure 3.

Figure 5.
Structure model for *C. hercules* 5' RNA (R2Ch). Structure is annotated as in Figure 3.

Figure 6.
Structure model for *C. promethea* 5' RNA (R2Cpr). Structure is annotated as in Figure 3.

Figure 7.
Structure model for *S. pyri* 5' RNA (R2Spy). Structure is annotated as in Figure 3.

Figure 8.

Secondary structure of *S. pyri* pseudoknot annotated with mutations from the other four R2 pseudoknots. Mutations appear next to their respective bases and are color coded according to the type of mutation: dark and light blue are compensatory and consistent mutations, respectively, red are inconsistent mutations (ones that lead to non-canonical pairs), grey are loop mutations, green are insertions, and finally green "X" characters indicate deletions. Also shown is a cartoon of the pseudoknot that combines the annotation of conserved features (i.e. P2–P4) from Figures 5–7 with standard pseudoknot nomenclature: S1 (stem 1), L1 (loop 1), etc. The tetraloop capping P3 on L1 is boxed and the three alternative tetraloop sequences are given in grey.

Figure 9.

Alignment of R2 open reading frames annotated with primary and secondary structure of RNA. Conserved helices are colored as in Figure 2; conserved codons are boxed and bolded. Green boxes indicate potential methionine start codons while red boxes indicate potential stop codons.

Figure 10.

Organization of conserved elements on the R2 element 5' segment. Features I and III-V indicate four conserved hairpin loops, while feature II represents the conserved pseudoknot. Features A-C in the open reading frame (ORF) are conserved fragments within the amino acid sequence. The putative region containing the ORF start site is based on amino acid conservation.

Table 1Statistics on the secondary structure and its conservation in R2 RNAs.^a

	nt	bps	cbps	cbpcm
R2Bm	323	103 (63.7%)	72 (44.6%)	34 (21.1%)
R2Sc	339	110 (64.9%)	75 (44.2%)	34 (20.1%)
R2Ch	329	100 (60.8%)	72 (43.8%)	33 (20.1%)
R2Cpr	336	100 (59.5%)	74 (44.0%)	33 (19.6%)
R2Spy	329	97 (59.0%)	72 (43.8%)	32 (19.5%)

^aThe number of nucleotides (nt), number of base pairs (bps) with percentage of nucleotides paired in parentheses, conserved base pairs (cbps) with percentage of nucleotides in conserved pairs in parentheses, and the number of conserved base pairs with compensatory mutations (cbpcm) with percentage of nucleotides in conserved pairs with compensatory mutation in parentheses are listed.

Table 2
Comparison of minimum free energy (MFE) and restrained MFE structures to final model structure including pseudoknotted base pairs.^a

5' R2 RNA	MFE		MFE + CM + MA	
	Sensitivity (%)	PPV (%)	Sensitivity (%)	PPV (%)
R2Bm	52.7	57.0	90.3	87.3
R2Sc	67.7	64.3	67.7	66.4
R2Ch	37.8	48.5	37.8	48.5
R2Cpr	51.1	57.4	67.8	64.8
R2Spy	74.7	72.1	72.4	71.7

^aThe sensitivity is the percentage of base pairs in the model structure that also appear in the predicted structure and PPV (positive prediction value) is the percentage of base pairs in the predicted structure that match the model structure. A correct pair could include “slippage” of the calculated paired nucleotides +/- one nucleotide. Chemical modifications (CM) include only strong hits from DMS, CMCT, NMIA, and kethoxal (kethoxal for R2Bm only) and microarray constraints (MA) include strong and moderate binding sites.