Supporting Information for

NMR-Assisted Prediction of RNA Secondary Structure: Identification of a Probable Pseudoknot in the Coding Region of an R2 Retrotransposon

James M. Hart,[†] *Scott D. Kennedy,*[‡] *David H. Mathews,*[‡] *and Douglas H. Turner*[*†]

[†]Department of Chemistry, University of Rochester, RC Box 270216, Rochester, NY 14627

[‡]Department of Biochemistry and Biophysics, University of Rochester School of Medicine and Dentistry, 601 Elmwood Avenue, Box 712, Rochester, NY 14642

Email: turner@chem.rochester.edu

NAPSS algorithm

READ **sequence** and **dotplot** (generated by RNAstructure)

READ **NMR constraints** (user-created text file that lists base pair connections)

FOR each **NMR constraint**

Iterate through all base pairs of **dotplot** to find possible helical matches

Do not consider nucleotides that have already been matched

At every step except the first or last in an **NMR constraint**:

Allow single-base bulge possibilities

Allow coaxial-stacking possibilities only if the flanking pairs are stable closing pairs for external and/or multibranch loops

STORE nucleotide numbers of **constraint matches**

FOR each **constraint match**

Create a new copy of the original **dotplot**

FOR each base pair in the **match combination**

Remove all dots that represent a base pair to either one of the nucleotides from the pair

Treat all remaining dots as **potential base pairs**

IF **match combination** base pairs force a pseudoknot folding

Remove **potential base pairs** that would create a highly complicated pseudoknot

Temporarily force **match combination** nucleotides to be single-stranded

DETERMINE thermodynamically favorable combinations of remaining **potential base pairs** using existing RNAstructure algorithm and parameters

Re-insert **match combination** base pairs

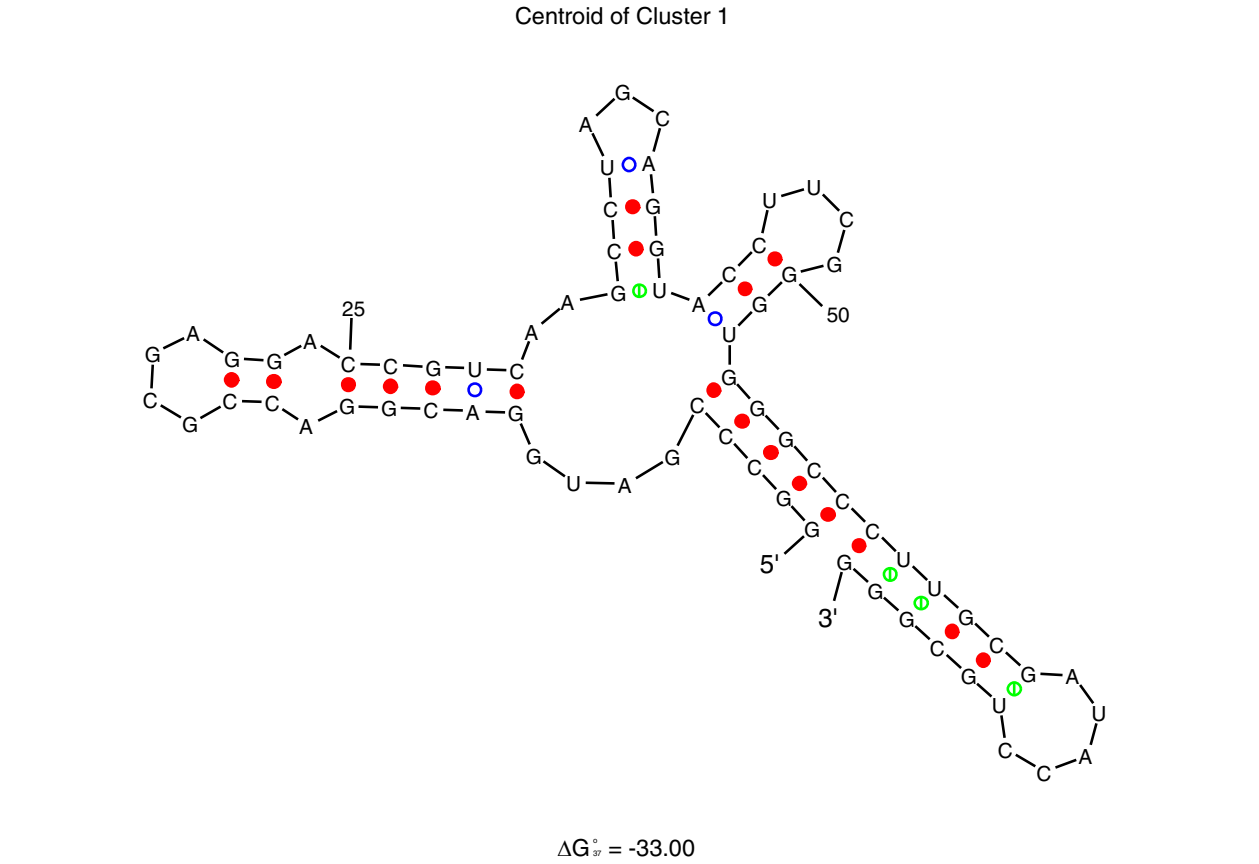
CALCULATE free energy of each **structure**

Discard any **structures** that are less stable than a user-defined cutoff

SORT **structures** according to calculated free energy

OUTPUT **structures** as a .ct file (**c**onnection **t**able format that lists the base pairs of each structure)

Figure S1. Pseudo-code representation of NAPSS.



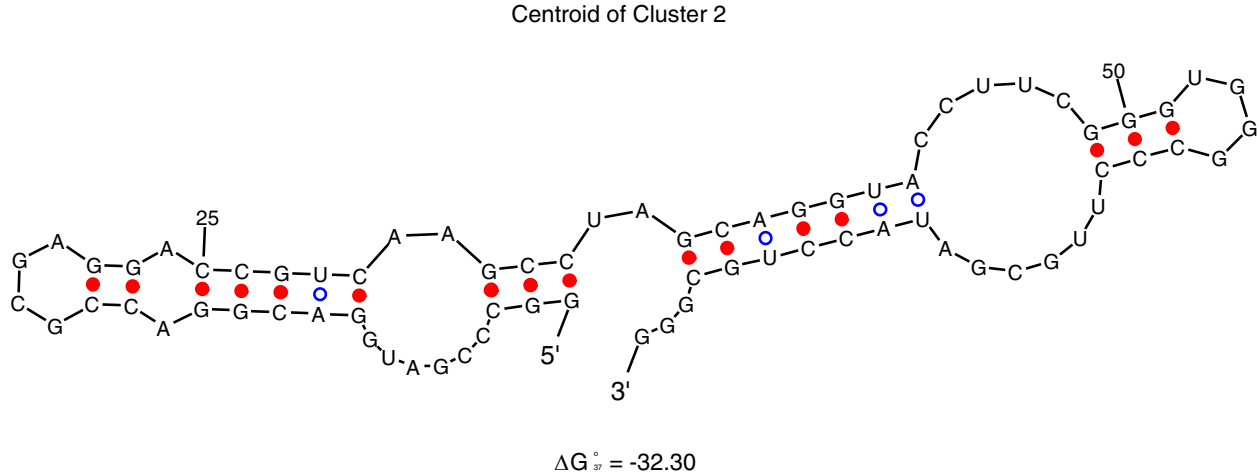


Figure S2. Centroid structures from SFOLD[1] for the *B. mori* R2 74-nt fragment. Neither structure is consistent with all four NMR constraints obtained for this sequence.

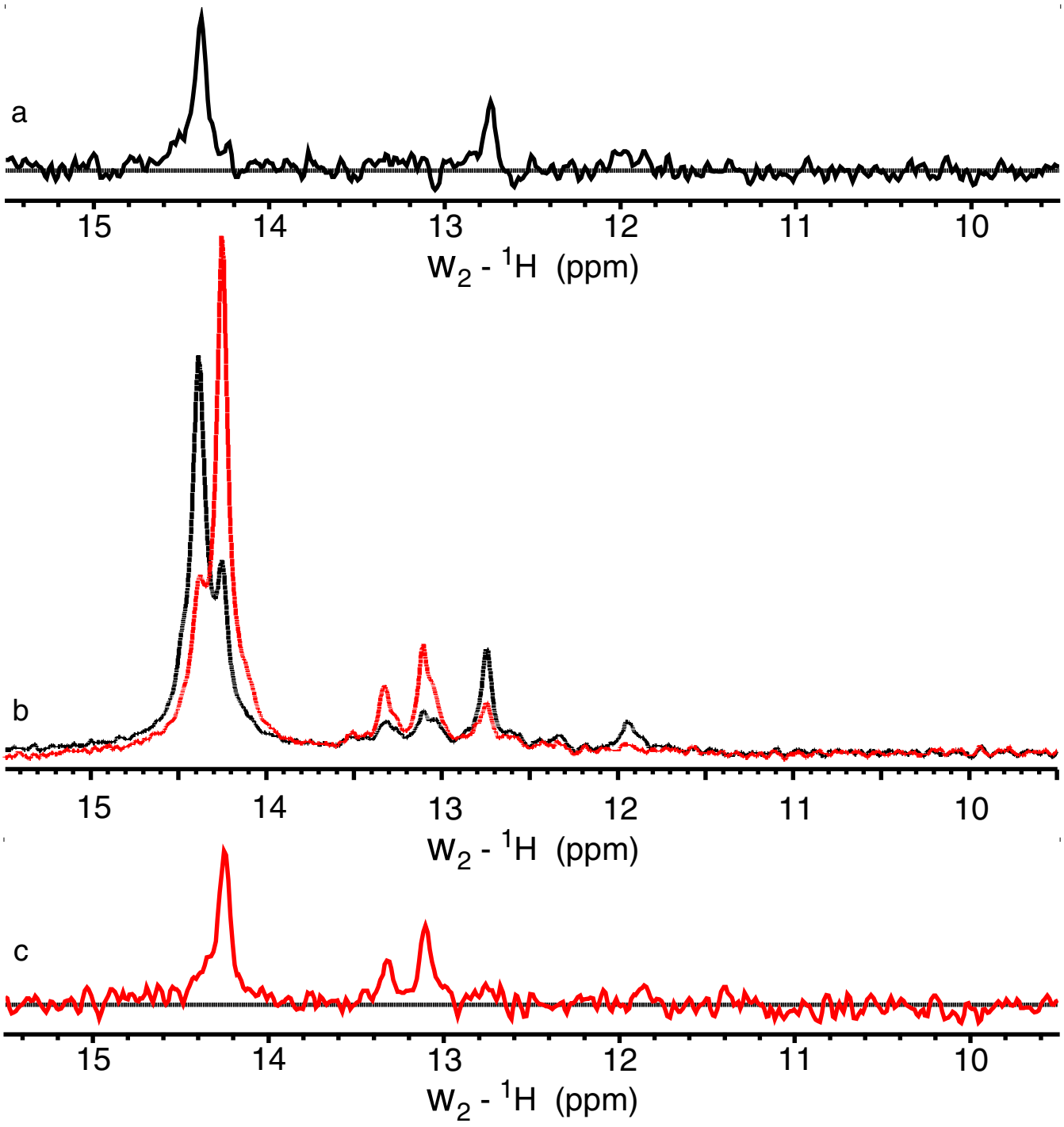


Figure S3. Evidence for an NOE connection between the U imino proton resonance at 14.38 ppm with the G imino proton resonance at 11.95 ppm for the 74-nt fragment of the *B. mori* R2 retrotransposon coding region. (a) 1D slice at 14.38 ppm from a 2D-NOESY experiment at 15ºC. There is some indication of a real signal centered near 12 ppm, but it is too weak to be definitively shown in a 2D

contour plot.

(b) 1D difference NOE experiments at 15ºC with the presaturation pulse set at 14.38 ppm (black) and 14.25 ppm (red). There is some signal intensity in each trace resulting from the incomplete selectivity of the presaturation pulse, but a comparison of the spectra clearly demonstrates the NOE connections of the 14.38 ppm resonance with the 12.75 ppm and 11.95 ppm resonances (black), as well as the connections between the 14.25 ppm, 13.33 ppm, and 13.10 ppm resonances (red). (c) 1D slice at 14.25 ppm from a 2D-NOESY experiment at 15ºC. Taken altogether, these experiments validate the least intense crosspeaks for the imino walks depicted in Figure 1 and Figure S4.

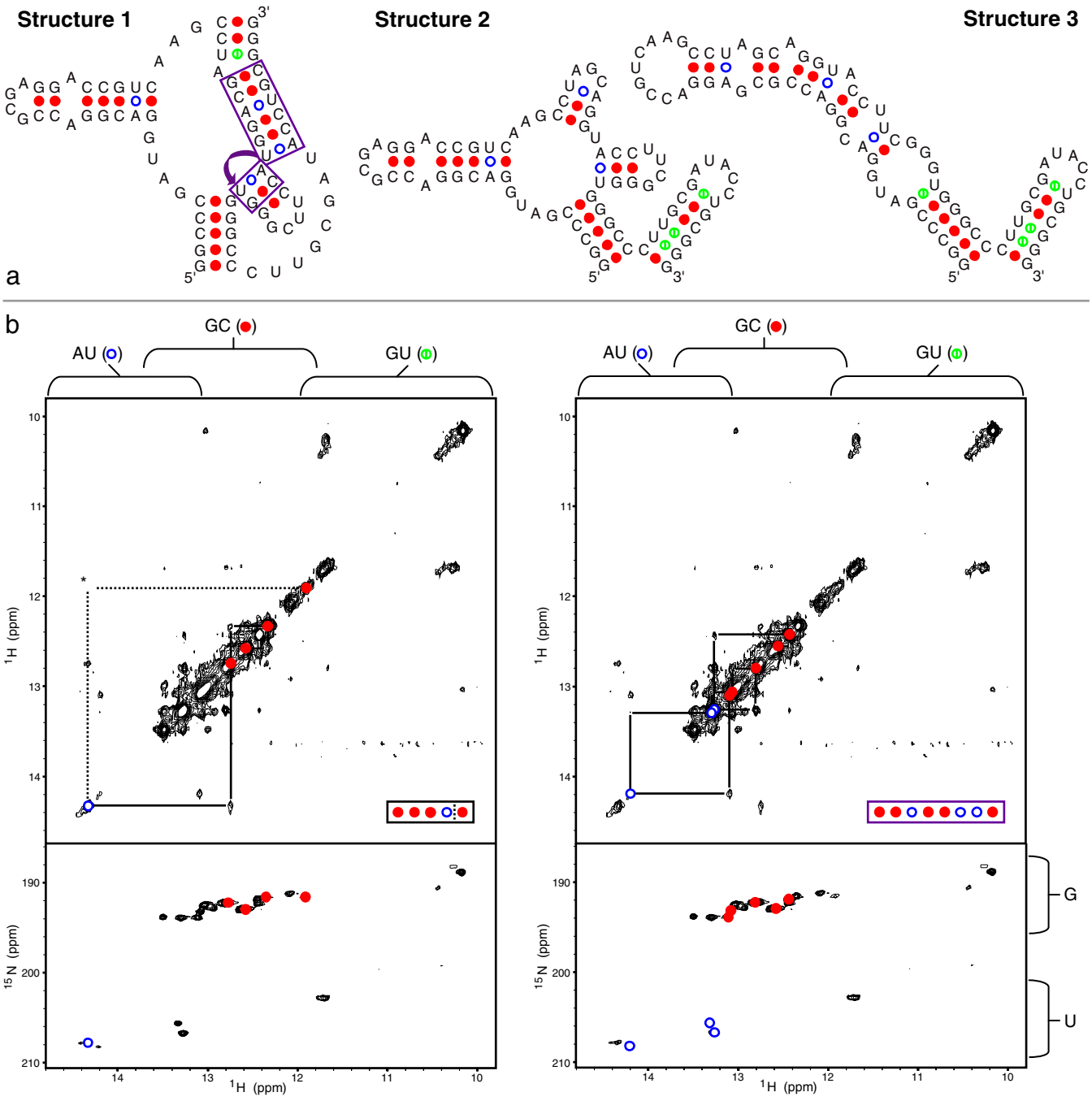
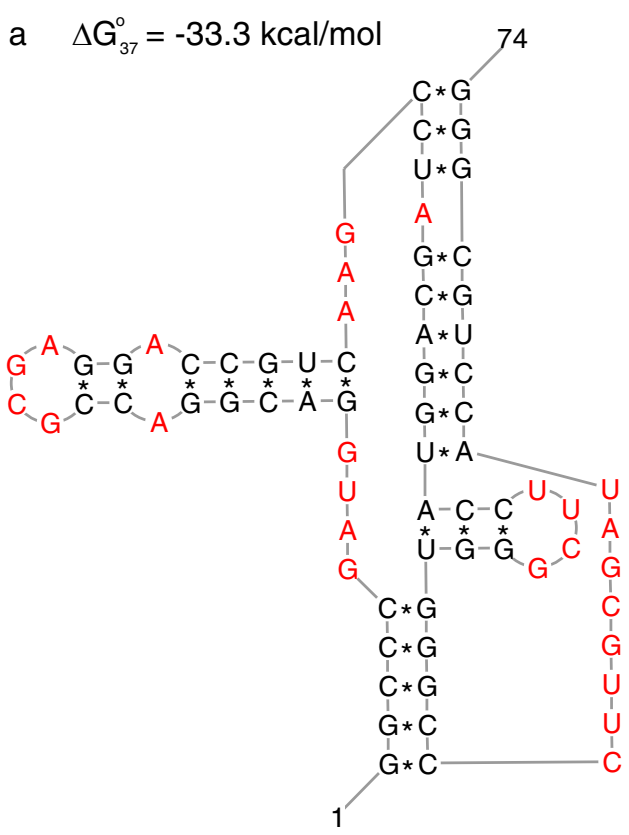


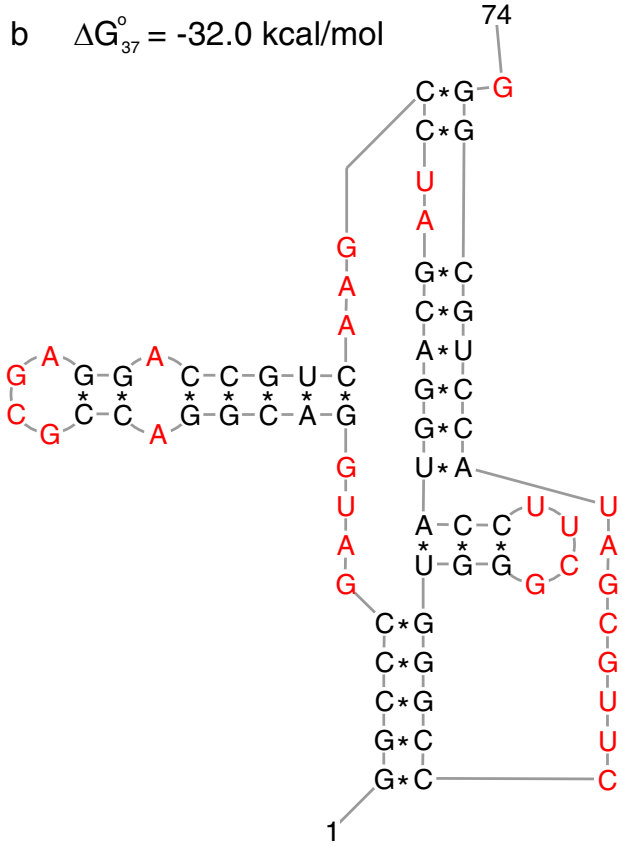
Figure S4. Comparison of predicted structures and NMR evidence at 25ºC for a 74-nt fragment of the *B. mori* R2 retrotransposon coding region. (a) Possible secondary structures; colored dots indicate base pair type (blue=AU, red=GC, green=GU). (b) NMR spectral data. Upper frames depict the imino region from a 200 msec 2D-NOESY spectrum of an unlabeled sample; lower frames are taken from a 2D-NHSQC spectrum of a G- and U-labeled sample. Typical chemical shift ranges are indicated by brackets

above and to the right of the spectra. Dashed lines and asterisk indicate an imino proton connection that is not apparent from this plot alone (see Figure S3 from Supporting Information). The connections in the upper left frame depict a helical walk that rules out Structure 3. Similarly, the upper right frame depicts a helix that does not match any in Structure 2 or 3, but does coincide with the helices in Structure 1 indicated by purple boxes (including a connection across a flush coaxial stack). Additional helical walks connecting two and five consecutive GC pairs are also observed but not depicted here. Data were acquired on a 600 MHz spectrometer at 25ºC in NMR buffer of 150 mM NaCl, 2 mM $MgCl_2$, 10 mM NaH_2PO_4, 0.05 mM EDTA, pH 6.0. The horizontal band of small peaks near 13.6 ppm is caused by a large upfield impurity peak around 1 ppm that appears here due to spectral wrapping in this dimension that was incorporated to reduce the overall experimental time.

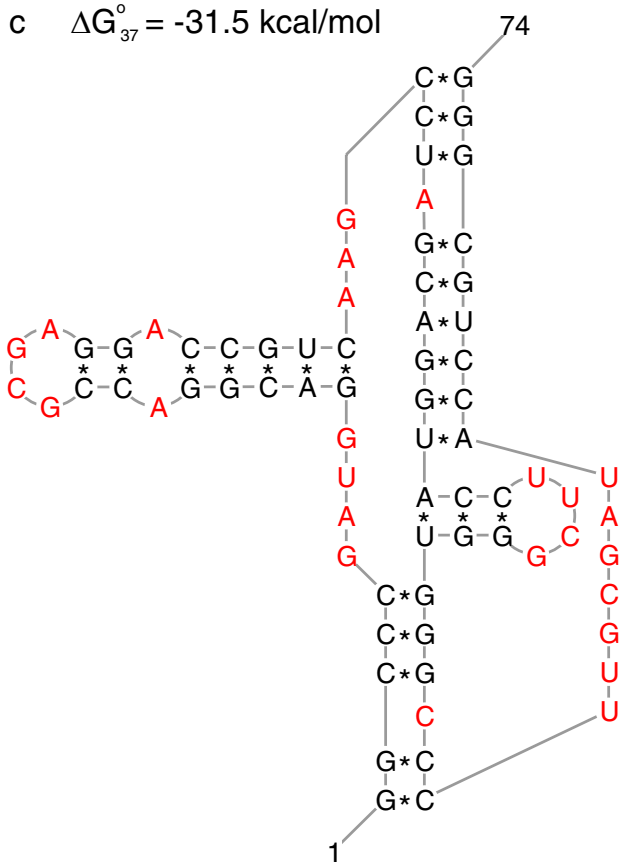
a ΔG°_{37} = -33.3 kcal/mol



b ΔG°_{37} = -32.0 kcal/mol



c ΔG°_{37} = -31.5 kcal/mol



d ΔG°_{37} = -30.2 kcal/mol

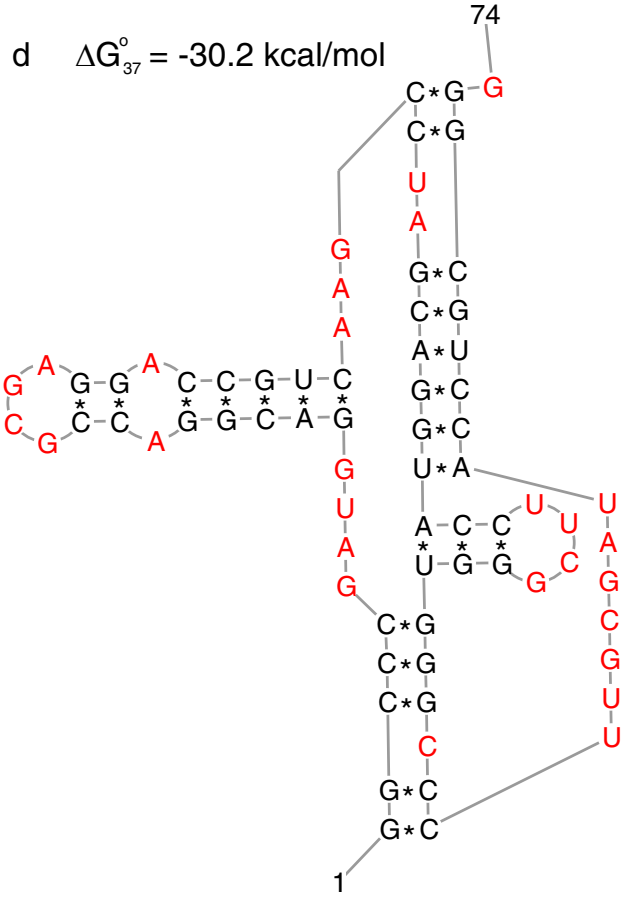


Figure S5. (a-d) The four most stable secondary structures output from NAPSS for the *B. mori* R2 74-nt fragment. These structures are the most stable predictions regardless of whether the walk on the left side of Figure 1b extends four or five base pairs. The remaining secondary structures from both predictions all contain the same major pseudoknotted helices as these structures, but with minor variations involving a few peripheral base pairs.

References

(1) Ding, Y.; Chan, C. Y.; Lawrence, C. E. *RNA* **2005,** *11*, 1157.