BIOCHEMISTRY
including biophysical chemistry & molecular biology

Article

pubs.acs.org/biochemistry

Nuclear Magnetic Resonance-Assisted Prediction of Secondary Structure for RNA: Incorporation of Direction-Dependent Chemical Shift Constraints

Jonathan L. Chen,[†] Stanislav Bellaousov,[‡] Jason D. Tubbs,[†] Scott D. Kennedy,[‡] Michael J. Lopez,[†] David H. Mathews,[‡,§] and Douglas H. Turner*[,†,§]

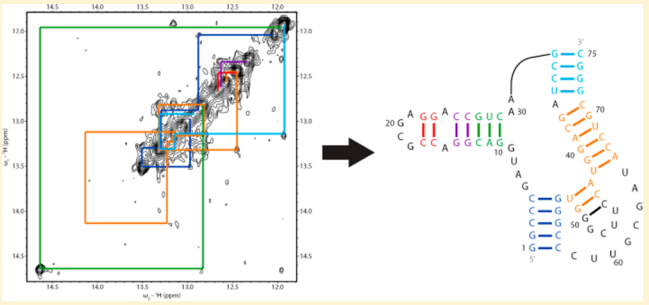
[†]Department of Chemistry, University of Rochester, Rochester, New York 14627, United States

[‡]Department of Biochemistry and Biophysics, University of Rochester School of Medicine and Dentistry, Rochester, New York 14642, United States

[§]Center for RNA Biology, University of Rochester, Rochester, New York 14642, United States

Supporting Information

ABSTRACT: Knowledge of RNA structure is necessary to determine structure−function relationships and to facilitate design of potential therapeutics. RNA secondary structure prediction can be improved by applying constraints from nuclear magnetic resonance (NMR) experiments to a dynamic programming algorithm. Imino proton walks from NOESY spectra reveal double-stranded regions. Chemical shifts of protons in GH1, UH3, and UH5 of GU pairs, UH3, UH5, and AH2 of AU pairs, and GH1 of GC pairs were analyzed to identify constraints for the 5′ to 3′ directionality of base pairs in helices. The 5′ to 3′ directionality constraints were incorporated into an NMR-assisted prediction of secondary structure (NAPSS-CS) program. When it was tested on 18 structures, including nine pseudoknots, the sensitivity and positive predictive value were improved relative to those of three unrestrained programs. The prediction accuracy for the pseudoknots improved the most. The program also facilitates assignment of chemical shifts to individual nucleotides, a necessary step for determining three-dimensional structure.



RNA is a central biomolecule involved in many cellular functions, including synthesizing proteins, regulating gene expression, catalyzing reactions, and storing genetic data in many viruses.[1] Therefore, knowledge of RNA structure can lead to the discovery of causes and cures of many diseases. Once sequence is known, the first step in determining RNA structure is defining the secondary structure, i.e., base pairing. This level of structure is used for many applications, including using nuclear magnetic resonance (NMR) to determine 3D structure,[2−4] designing therapeutics,[5] and providing insights into functional mechanisms[6] and evolution.[7−10]

Methods for predicting RNA secondary structure include comparative sequence analysis[7,11−13] and minimization of free energy predicted on the basis of thermodynamic parameters.[14−18] The latter method is ∼70% accurate when predicting secondary structure from a single sequence,[19,20] but prediction methods that integrate sequence comparison typically give better than 85% average accuracy.[21−26]

A pseudoknot is a type of RNA secondary structure in which nucleotides in a loop region pair with nucleotides outside of where the loop was closed. Formally, a pseudoknot occurs if there are two base pairs with indices i paired to j and k paired to l, with positions $i < k < j < l$. Base pairs forming pseudoknots are the most difficult to predict, and most popular programs do not allow pseudoknots. Pseudoknots, however, have important biological roles.[27−33] The lack of thermodynamic data for pseudoknots compounds the difficulty of accurately predicting them by calculating free energy changes.[34,35]

To improve the accuracy of predicting secondary structure, techniques that employ experimental data have been developed. Chemical mapping constraints[20] and/or restraints[36−38] can be used to identify nucleotides not in Watson−Crick base pairs. NMR constraints identify nucleotides in Watson−Crick and GU pairs.[39]

Crystallization of RNA is challenging.[40,41] Therefore, many RNA 3D structures have been determined by NMR, which also provides insight into dynamics. NMR-Assisted Prediction of Secondary Structure (NAPSS) utilizes NMR-derived constraints in conjunction with a thermodynamic folding algorithm to reduce the potential folding space of an RNA molecule and provide some initial chemical shift assignments.[39] Imino proton 2D NOE spectroscopy (NOESY) is used to identify helices of stacked canonical base pairs. The user enters constraints from sequential imino proton walks into a program that recursively

Received: July 23, 2015
Revised: October 6, 2015
Published: October 9, 2015

DOI: 10.1021/acs.biochem.5b00833
Biochemistry 2015, 54, 6769−6782

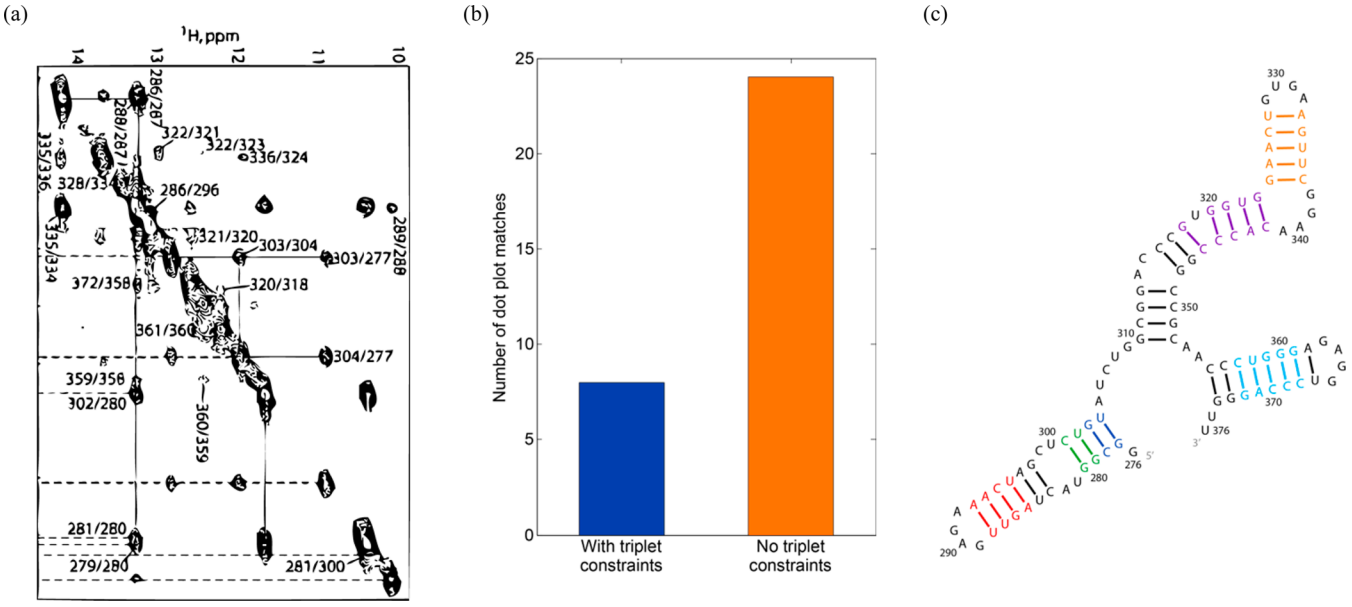


Figure 1. NAPSS-CS with direction-dependent constraints reduces the folding space of an RNA sequence. (a) A 2D NOESY spectrum of the Moloney murine leukemia virus core encapsidation signal multibranch loop showing imino proton walks.[63] (b) Applying triplet constraints to folding of this sequence by NAPSS-CS reduces the number of dot plot matches by 16. (c) Secondary structure of this sequence. Colored base pairs correspond to helixces identified by imino proton walks.

searches for structures containing matching helical walks. For the set of structures in which all constraints are matched, the program refolds unconstrained regions and calculates the folding free energy change of the entire structure. The putative structures are ranked by predicted folding free energy change. The helical walk constraints, however, do not provide any information about the 5′ to 3′ strand direction of the base pair stacks. That is, the directionality of the set of stacks along the sequence is unknown. Thus, NAPSS often identifies a large set of possible structures that satisfy all of the stacking constraints.

Chemical shifts of protons in base pairs are affected by the identity of neighboring bases.[42−44] Direction-dependent trends for chemical shifts of base pair stacks were mostly identified from the Biological Magnetic Resonance Data Bank (BMRB)[45] and by measuring spectra for some model duplexes with GU pairs. Here, we report direction-dependent trends for GH1, UH3, and UH5 of GU wobble pairs, AH2, UH3, and UH5 of AU pairs, and GH1 of GC pairs. An expanded NAPSS algorithm, NAPSS-CS, is also provided that uses these spectroscopic properties to constrain predictions of RNA secondary structure (Figure 1).

While the NAPSS-CS algorithm is intended for RNA secondary structure prediction, the approach can be generalized to self-assembling polymers in which structure allows NMR walks and thermodynamic rules are known. This includes self-assembling DNA structures and nucleic acid mimics with different backbones and "bases" for which chemical modification reagents may not be available.[46−49] The algorithm may also facilitate determination of secondary structure when more than one structure is present and the structures are in slow exchange on the NMR time scale.

METHODS

Oligoribonucleotide Preparation and NMR Spectroscopy. To further explore the dispersion of resonances for GU pairs, whose chemical shifts are underrepresented in the literature compared to AU and GC pairs, NOESY and TOCSY spectra were measured for the following sequences to determine GH1, UH3, UH5, and AH2 chemical shifts: r(AGGCUU)$_2$, r(AGUCGAUU)$_2$, r(CUGGCUAG)$_2$, r(CAGUCGAUUG)$_2$, r(CCGAAUUUGG)$_2$, r(CGGAAUUUCG)$_2$, r(CGGAUAUUCG)$_2$, r(CGUGAUUACG)$_2$, r(CUGGAUUCAG)$_2$, r(GAGAGCUUUC)$_2$, r(GAGGAUCUUC)$_2$, and r(GUGAAUUUAC)$_2$. Imino proton 1D spectra for these sequences have been published.[50]

All oligoribonucleotides were purchased from Integrated DNA Technologies, Inc., except for r(GUGAAUUUAC)$_2$, which was synthesized by M. Serra (Allegheny College, Meadville, PA) using standard phosphoramidite chemistry. Oligoribonucleotides were dissolved in 300 μL of 80 mM NaCl, 18.8 mM NaH_2PO_4, 1.16 mM Na_2HPO_4, and 0.02 mM EDTA at pH 6.0. To provide a lock signal, 15 μL of D_2O was then added. NMR spectra were acquired with a Varian Inova 500 MHz spectrometer. A 35 ms mixing time was used for 2D TOCSY spectra, and 100 and 400 ms mixing times were used for 2D ^{1}H−^{1}H NOESY spectra. For 2D spectra, water signals were suppressed with a WATERGATE-style pulse with flipback.[51,52] 2D spectra were processed with NMRPipe[53] and resonances assigned with SPARKY.[54]

Proton chemical shifts were referenced to a temperature-dependent water shift (eq 1), where T is temperature in kelvin measured at pH 5.5.[55] 2,2-Dimethylsilapentate-5-sulfonic acid was used as an external reference standard for water.

$$\delta = 7.83 - T/96.9 \quad (1)$$

Quantification of Secondary Structure Accuracy. The accuracy of prediction of secondary structures was quantified by sensitivity and positive predictive value (PPV):

$$\text{sensitivity} = \frac{\text{number of correctly predicted base pairs}}{\text{total number of known base pairs}} \quad (2)$$

$$\text{PPV} = \frac{\text{number of correctly predicted base pairs}}{\text{total number of predicted base pairs}} \quad (3)$$

DOI: 10.1021/acs.biochem.5b00833
Biochemistry 2015, 54, 6769−6782

Table 1. Averages of Measured Distances between Imino Protons of Adjacent Base Pairs in X-ray and NMR Structures Obtained from the PDB[a]

base pair doublet	sets of bases	average distance in X-ray structures (Å)	occurrence in X-ray structures	average distance in NMR structures (Å)	occurrence in NMR structures
5′AA3′	U-U	3.71	17	3.88	49
3′UU5′					
5′AG3′	G×U	3.38	1	3.63	12
3′UU5′	U-U	3.85	1	4.20	10
5′AU3′	U×U	3.52	5	3.61	14
3′UA5′					
5′AU3′	G-U	N/A		3.67	7
3′UG5′	U×U			3.77	7
5′CA3′	G-U	4.85	41	4.78	60
3′GU5′					
5′CG3′	G×G	4.35	35	4.35	30
3′GC5′					
5′CG3′	G×G	4.70	10	4.64	18
3′GU5′	G-U	5.48	10	4.84	18
5′CU3′	G×U	3.46	35	3.68	75
3′GA5′					
5′CU3′	G-G	4.26	13	4.17	26
3′GG5′	G×U	3.57	13	3.93	26
5′GA3′	G×U	4.48	17	4.57	86
3′CU5′					
5′GC3′	G×G	3.66	41	3.71	55
3′CG5′					
5′GG3′	G-G	4.02	53	4.03	94
3′CC5′					
5′GG3′	G-G	4.32	14	4.21	9
3′CU5′	G×U	4.63	14	4.64	9
5′GG3′	G-G	4.36	1	4.69	2
	G×U	4.46	2	4.75	4
3′UU5′	U-U	3.39	1	3.94	2
5′GU3′	G-U	3.34	29	3.57	81
3′CA5′					
5′GU3′	G×G	3.86	12	4.06	20
3′CG5′	G-U	3.47	12	3.69	20
5′GU3′	G×G	3.79	1	3.55	4
3′UG5′	G-U	3.75	2	3.65	8
	U×U	3.83	1	3.78	4
5′UA3′	U×U	5.45	7	5.18	16
3′AU5′					
5′UG3′	G-U	N/A		5.11	3
3′AU5′	U×U			5.54	3
5′UG3′	G×G	5.64	1	5.13	2
	G-U	5.91	2	5.68	4
3′GU5′	U×U	5.44	1	5.45	2
5′UU3′	G×U	4.95	1	4.82	6
3′AG5′	U-U	3.64	1	3.83	6

[a]-, Same strand distance; ×, cross-strand distance.

A predicted pair, *i* to *j*, was considered correct if the accepted structure contained a pair from *i* to *j*, *i* ± 1 to *j*, or *i* to *j* ± 1. This accounts for the fact that comparative analysis cannot always resolve the exact pairing and that pairs can be dynamic.[19]

RNA Preparation and NMR Spectroscopy. A plasmid construct containing the sequence for human accelerated region 1 (HAR1) used by Beniaminov et al.[9] was provided by A. Krol of the Institut de Biologie Moléculaire et Cellulaire.

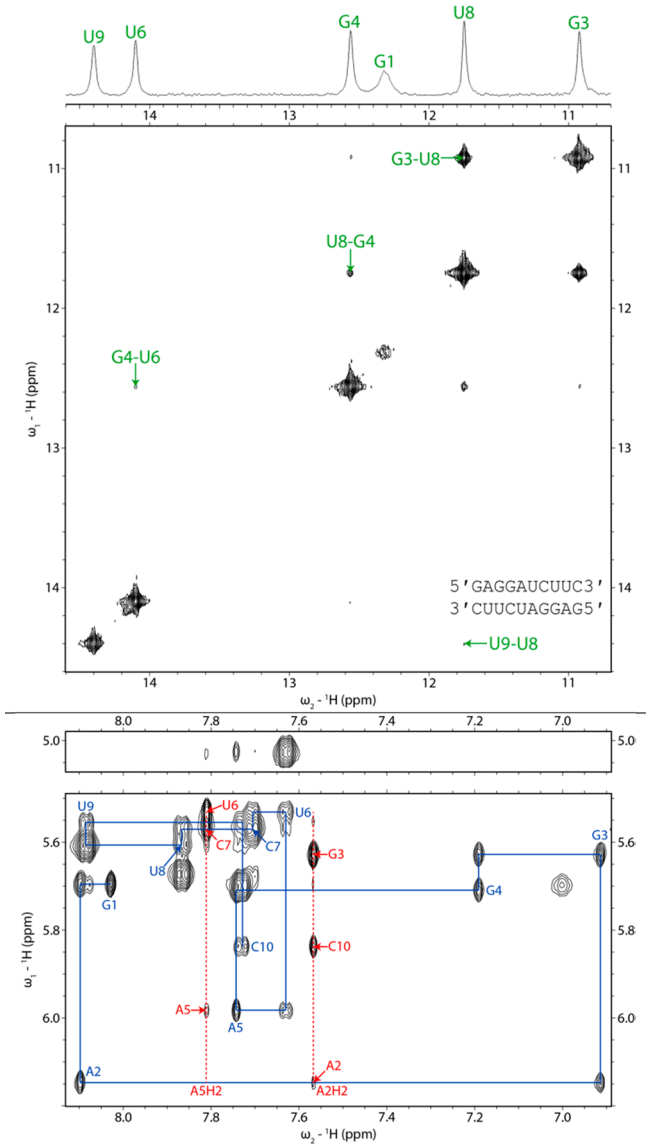


Figure 2. 2D imino (top) and NOESY walk (bottom) region spectra of r(GAGGAUCUUC)$_2$. The top spectrum was acquired with a 100 ms mixing time at 5 °C, and the bottom spectrum was acquired with a 400 ms mixing time at 25 °C. Both spectra were acquired with a WATERGATE pulse to suppress water.

The plasmid was amplified in and purified from *Escherichia coli* competent cells using standard plasmid preparation protocols. The purified plasmid was linearized overnight by *Sma*I restriction endonuclease (New England BioLabs) at 25 °C. RNA was transcribed from the linearized plasmid with T7 High Yield RNA Synthesis Kits (New England BioLabs), with 1 μg of plasmid per 20 μL reaction mixture. After transcription mixtures had been incubated for 12 h at 37 °C, 5 μL of 100 mM EDTA was added to stop reactions. A sample with ^{15}N- and ^{13}C-labeled rGTP and rUTP (Sigma-Aldrich) was synthesized with a similar protocol.

HAR1 samples were purified with FPLC.[56] FPLC fractions were concentrated and exchanged into NMR buffer with Amicon Ultra-15 Centrifugal Filter Units (EMD Millipore). NMR buffer consisted of 100 mM NaCl, 20 mM NaH_2PO_4/$NaHPO_4$, and 0.05 mM Na_2EDTA (pH 6.25). Each sample was placed in a susceptibility-matched Shigemi NMR

DOI: 10.1021/acs.biochem.5b00833
Biochemistry 2015, 54, 6769−6782

Table 2. Chemical Shift Ranges Used as Direction-Dependent Constraints for Secondary Structure Prediction[a]

base pair	AH2/GH1 range	UH3 range	UH5 range	triplet
GU	11.30−11.90	11.60−12.80	5.00−5.75	+RGY
GU	9.70−10.35	11.10−12.00	5.50−6.00	+YGR
GU	10.60−11.20	11.45−12.30	5.20−5.50	+YGY
GU	11.30−11.90	11.60−12.80	–	+RGY
GU	9.70−10.35	11.10−12.00	–	+YGR
GU	10.35−10.60	11.10−12.30	5.20−6.00	−RGY
GU	9.70−10.35	12.00−12.30	5.20−6.00	−RGY
GU	9.70−10.35	11.10−12.00	5.20−5.50	−RGY
GU	10.60−11.20	11.10−11.45	5.20−6.00	−RGY
GU	10.60−11.20	11.45−12.30	5.50−6.00	−RGY
GU	9.70−10.35	12.00−12.30	–	−RGY/−YGY
GU	10.35−10.60	11.10−12.30	–	−RGY/−YGY
GU	10.60−10.80	11.10−12.30	–	−RGY
GU	10.80−11.20	11.10−11.45	–	−RGY
GU	10.80−11.30	11.45−12.30	–	−RGY/−YGR
AU	7.75−8.10	14.00−14.80	4.70−5.20	+RAY
AU	6.30−6.75	12.90−13.40	5.35−5.85	+YAR
AU	6.75−7.25	12.80−13.50	4.70−5.20	+YAY
AU	7.70−8.10	14.00−14.80	–	+RAY
AU	6.30−6.75	12.90−13.40	–	+YAR
AU	–	14.25−14.80	4.70−5.10	+RAY
AU	–	12.80−13.50	4.70−5.20	+YAY
AU	6.30−6.75	12.80−12.90	–	−RAY
AU	6.30−6.75	13.40−13.50	–	−RAY
AU	6.75−7.40	12.80−13.50	–	−RAY
AU	–	12.90−13.50	5.20−5.80	−RAY
GC	11.50−12.10	–	–	−YGY
GC	13.30−13.90	–	–	−YGR

[a]For each triplet type, R and Y represent purines (G or A) and pyrimidines (C or U), respectively. The algorithm incorporates "+" triplets into and excludes "−" triplets from potential matches.

tube (Shigemi, Inc.). Final NMR samples were 300 μL, including 10 μL of D_2O to provide a lock signal. The final RNA concentration of each sample was approximately 0.2 mM.

A pUC19 plasmid containing an insert for a 75 nt sequence for the *Bombyx mori* R2 retrotransposon pseudoknot was constructed and amplified with standard plasmid purification protocols (see the Supporting Information for the plasmid construct). This construct differs from the 74 nt construct previously studied with NMR[39] by addition of the wild-type 3′ terminal C. The plasmid was linearized with NheI restriction endonuclease at 37 °C prior to *in vitro* transcription. Unlabeled and ^{13}C- and ^{15}N-labeled rGTP and rUTP samples were transcribed, purified with polyacrylamide gel electrophoresis, and extracted from gels with electroelution. Purified RNAs were exchanged into NMR buffer (see ref 39), concentrated with centrifugal filter units, and added to a Shigemi NMR tube. D_2O was added to provide a lock signal.

NMR spectra were recorded on a Varian Inova 600 MHz NMR spectrometer. NOESY spectra were acquired for the unlabeled HAR1 sample with mixing times of 125 and 250 ms at 25 °C and 60 ms at 15 °C. A $^{15}N-^{1}H$ HSQC spectrum was acquired for the labeled HAR1 sample at 25 °C. Relevant acquisition parameters are listed in Table S1. NOESY spectra were acquired for the unlabeled 75 nt *B. mori* R2 retrotransposon pseudoknot with a mixing time of 200 ms at 25 °C. A $^{15}N-^{1}H$ HSQC spectrum was acquired for the labeled *B. mori* R2 retrotransposon pseudoknot sample at 25 °C.

RESULTS

Imino Proton Distances in PDB Helices. Adjacent canonical base pairs and therefore potential helices can be identified from NOEs between imino protons separated by <5 Å. A database of imino proton distances between doublets of canonical base pairs was assembled from 3D structures in the Protein Data Bank (PDB).[57] Distances were obtained from nonredundant X-ray structures with a resolution no larger than 2 Å,[58] to which hydrogens were added with the REDUCE program[59] and then averaged for each doublet of canonical pairs. For doublets with at least one GU pair, G/G, G/U, and/or U/U, imino proton distances between adjacent base pairs were averaged separately, although the 5′G-5′U and 3′G-3′U distances were averaged together for 5′GG3′/3′UU5′. A separate set of imino proton distances for each doublet was obtained from NMR structures. Imino distances between doublets adjacent to loops, bulges, and helix termini were not considered because the terminal or closing base pairs may be dynamic.

Most average distances from the X-ray and NMR structures agreed within 0.3 Å (Table 1). In X-ray structures, 26 interbase average imino proton distances were <5 Å, thus allowing "imino proton walks" to identify helices.[60] The exceptions were 5′UA/3′AU, 5′UG/3′AU, 5′UG/3′GU, and G to U in X-ray structures of 5′CG/3′GU. Thus, helices with these doublets may not be completely constrained by NMR data. In published NMR spectra, however, an NOE between H3 of uracils of 5′UA/3′AU was observed for a human R/G stem-loop[61] and a hairpin stem from *Yersinia enterocolitica*.[62] In contrast, a

DOI: 10.1021/acs.biochem.5b00833
Biochemistry 2015, 54, 6769−6782

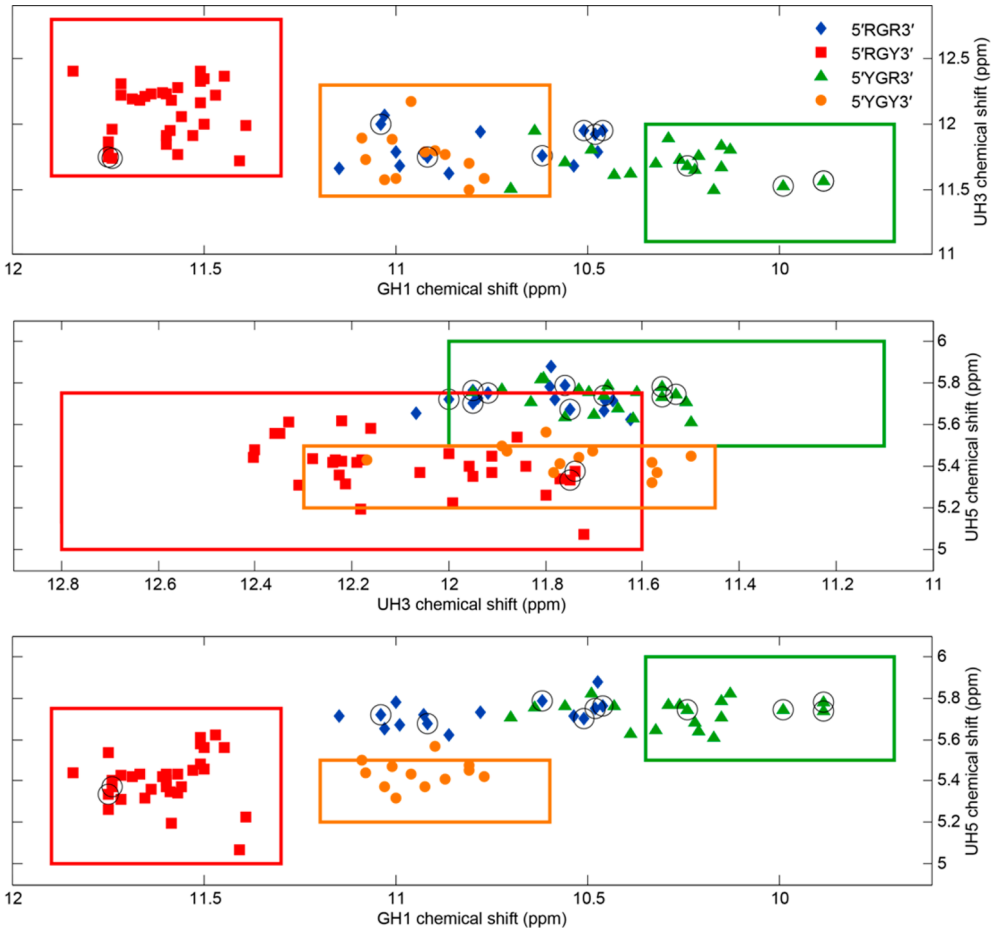


Figure 3. Chemical shift patterns for GU pairs. UH3 vs GH1 (top), UH5 vs UH3 (middle), and UH5 vs GH1 (bottom) colored according to triplet type. For each triplet type, R is purine (A or G) and Y is pyrimidine (C or U). Circled points represent chemical shifts not previously reported in the literature. Colored boxes are chemical shift ranges used as direction-dependent constraints for triplets with the same color points. Apparent overlaps in these plots are resolved by including a third chemical shift (Figure 4).

5′UA/3′AU imino NOE was not observed for the core encapsidation signal RNA from Moloney MLV,[63] likely because of overlap between the H3 signals. Similarly for 5′UG/3′AU, NMR spectra of the human R/G stem-loop[61] and the loop E region of *Spinacia oleracia* 5S rRNA[64] have an NOE between either imino proton of the GU pair and UH3 of the AU pair, but these NOEs were not observed in the spectrum of r(GUGAAUUUAC)$_2$ reported here. For the 5′CG/3′GU doublet, an NOE was observed between H3 of U and H1 of the 5′ G in NMR spectra of a mutant of an element of the 3′ UTR of turnip crinkle virus[65] and a domain of medaka telomerase CR4/5.[66] Sometimes spin diffusion will produce imino−imino proton NOEs even if the imino protons are separated by >5 Å.[60] Thus, complete helical walks are more likely to be observed than expected for the distances in Table 1.

Spectral Analysis of Duplexes with GU Pairs. To expand the database for GU pairs, exchangeable protons of short duplexes were assigned from NOESY spectra and the nonexchangeable H5 and H6 protons of C and U were assigned from TOCSY spectra (Tables S2−S13). CH5−H41/42 cross-peaks in NOESY spectra differentiated C from U in the TOCSY spectra and initiated assignment of proton resonances. Imino protons were identified as those of GU, GC, or AU pairs based largely on their positions in the imino fingerprint regions, from 10 to 12 ppm for GH1 and UH3 with a strong GH1/UH3 cross-peak for GU pairs, from 11 to 13.5 ppm for GH1 in GC pairs, and from 13 to 15 ppm for UH3 in AU pairs.[60]

Aromatic proton assignments were confirmed with sequential aromatic walks. In A-form RNA, H1′ of a residue forms NOE contacts with its own H6 or H8 base proton and that of its 3′ base, allowing sequential assignments.[60] H2 of adenosine has an intrastrand cross-peak to H1′ of its 3′ residue and an interstrand cross-peak to H1′ of the residue 3′ of the base to which it is paired. While NMR spectra for all duplexes used in this study were assigned with similar methods, an example of the process for r(GAGGAUCUUC)$_2$ is given below. Two others are given in the Supporting Information.

2D NOESY spectra for r(GAGGAUCUUC)$_2$ were obtained with mixing times of 100 and 400 ms at 5 and 20 °C, respectively. The G3 H1 and U8 H3 protons were assigned according to their positions in the imino region of the 100 ms spectrum (Figure 2). G4 H1 was identified by its cross-peaks to U8 H3 and G3 H1, while G1 H1 displays a weak resonance due to exchange with water. H3 of U6 was differentiated from U9 by its cross-peak to G4 H1. H5 to H6 peaks corresponding to C7 and C10 were differentiated by cross-peaks from the former's H41 and H42 to G4H1. Remaining aromatic proton and H1′ assignments were made with the 400 ms spectrum as described above (Figure 2). The H1′ assignments of U6 and C7 were confirmed by NOE contacts to A5 H2, while those of G3 and C10 were confirmed by NOE contacts to A2 H2.

DOI: 10.1021/acs.biochem.5b00833
Biochemistry 2015, 54, 6769−6782

Database. NOEs from imino proton resonances provide identification of intrabase pair resonances for H5 of C and U and H2 of A.[60] A database of such NMR chemical shifts was assembled for 78 GU wobble pairs, 292 AU pairs, and 490 GC pairs flanked on both sides by at least one canonical base pair and categorized according to the identities and orientations of the flanking canonical pairs. Most of the structures were 10−40 nt long and contained stem-loops, while a few were completely double-stranded RNA. Literature data were mostly obtained from the BMRB.[45]

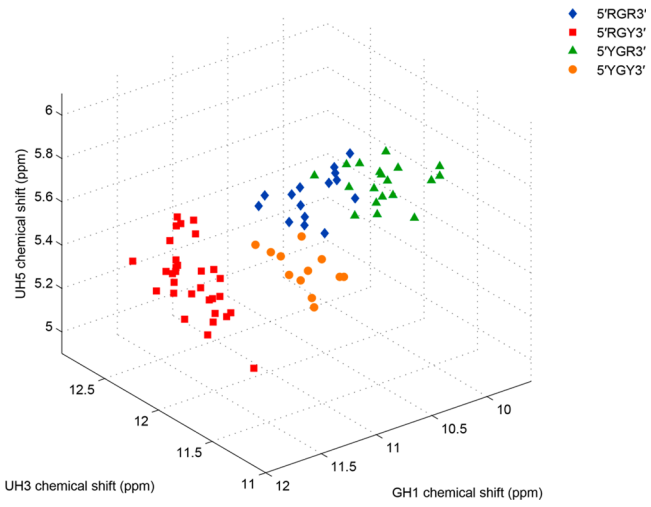


Figure 4. Chemical shift patterns for GU pairs. UH5 vs UH3 vs GH1 chemical shifts colored according to triplet type.

Chemical shifts determined for GH1, UH3, and UH5 protons of GU pairs, AH2, UH3, and UH5 protons of AU pairs, and GH1 protons of GC pairs on the basis of oligoribonucleotide spectra reported here were consistent with existing data. The expanded database showed strong sequence-dependent patterns, which were used to determine chemical shift ranges that identify the 5′ to 3′ direction of base pair triplets centered on GU, AU, and GC pairs (Table 2).

For GU pairs, a 2D plot of UH3 versus GH1 chemical shifts (Figure 3) revealed well-defined clusters for 5′RGY3′ and 5′YGR3′, but 5′RGR3′ and 5′YGY3′ regions overlap, where R and Y represent purines (G or A) and pyrimidines (C or U), respectively. A 2D plot of UH5 versus GH1 chemical shifts contained well-defined clusters corresponding to 5′RGY3′, 5′YGR3′, and 5′YGY3′, but not 5′RGR3′ due to overlap with the 5′YGR3′ region (Figure 3). Thus, if triplets have a central GU pair, then GH1, UH3, and UH5 chemical shifts often reveal the direction of the helix (Figures 3 and 4).

For AU pairs, 2D plots of AH2 versus UH3 chemical shifts (Figure 5) revealed defined clusters corresponding to 5′RAY3′

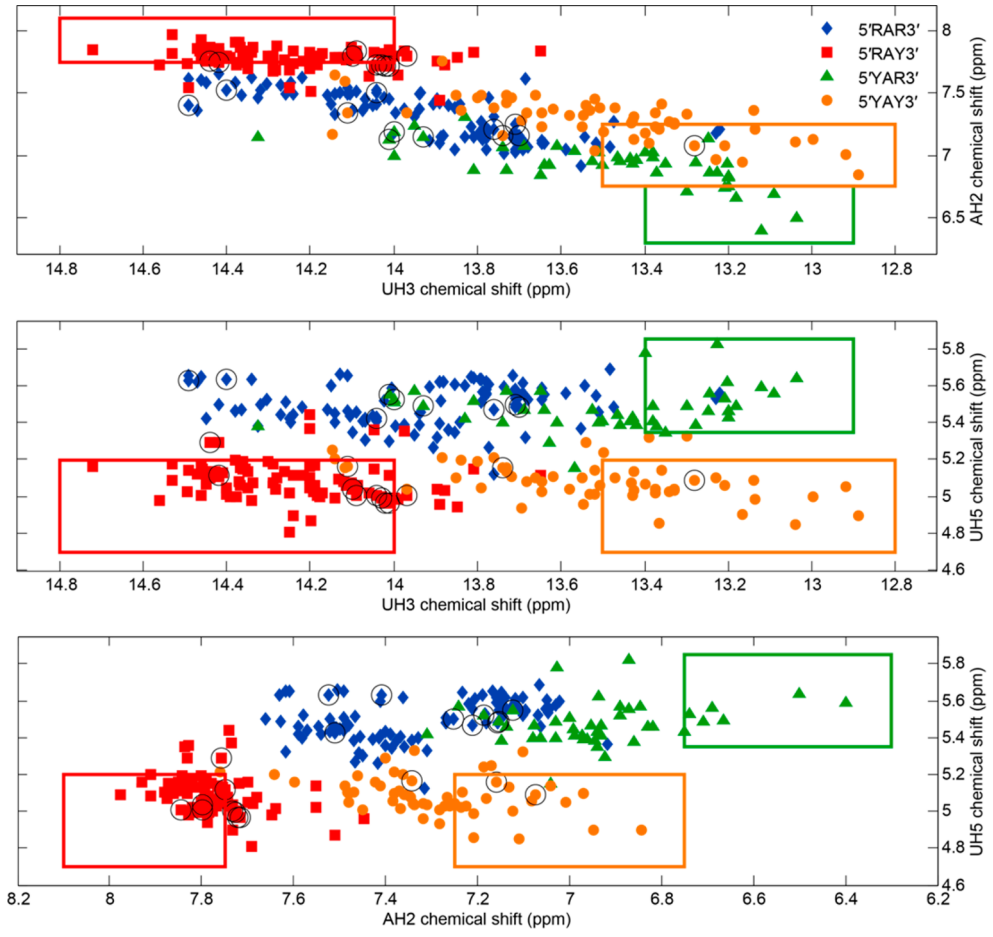


Figure 5. Chemical shift patterns for AU pairs. AH2 vs UH3 (top), UH5 vs UH3 (middle), and UH5 vs AH2 (bottom) colored according to triplet type. For each triplet type, R is purine (A or G) and Y is pyrimidine (C or U). Circled points represent chemical shifts not previously reported in the literature. Colored boxes are chemical shift ranges used as direction-dependent constraints for triplets with the same color points. Apparent overlaps in these plots are resolved by including a third chemical shift (Figure 6).

DOI: 10.1021/acs.biochem.5b00833
Biochemistry 2015, 54, 6769−6782

and 5′YAR3′, but the 5′RAR3′ region overlaps with 5′YAY3′. A 2D plot of UH3 versus UH5 chemical shifts contained well-defined clusters corresponding to 5′RAY3′ and 5′YAY3′, but the 5′RAR3′ region overlaps with 5′YAR3′. Thus, if triplets have a central AU pair, then UH3, AH2, and UH5 chemical shifts often reveal the direction of the helix (Figures 5 and 6).

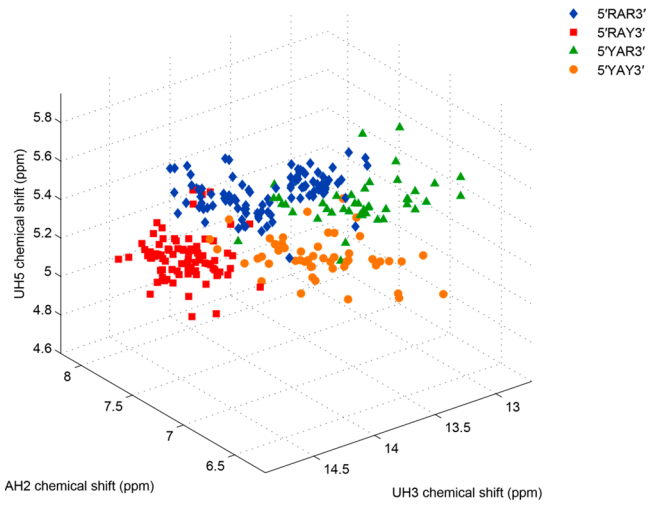


Figure 6. Chemical shift patterns for AU pairs. UH5 vs UH3 vs AH2 chemical shifts colored according to triplet type.

For GC pairs flanked by canonical base pairs, a plot of the distribution of GH1 chemical shifts (Figure 7) showed that no

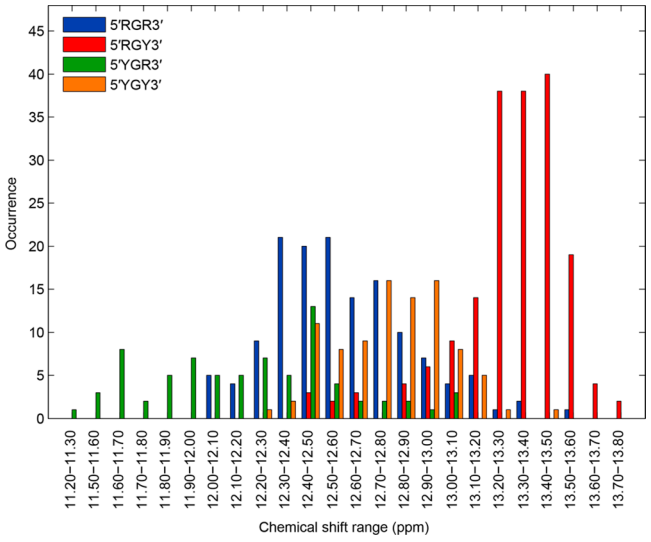


Figure 7. Distribution of GH1 chemical shifts for GC pairs colored according to triplet type. For each triplet type, R is purine (A or G) and Y is pyrimidine (C or U). No GH1 chemical shifts from 11.31 to 11.50 ppm were found for GC pairs.

GC pairs in the 5′RGY3′ direction have a GH1 chemical shift below 12.40 ppm, whereas no GC pairs in the 5′YGR3′ direction have a chemical shift above 13.10 ppm. Thus, if triplets have a central GC pair, then a GH1 chemical shift below 12.40 ppm eliminates the 5′RGY3′ direction and a shift above 13.10 ppm eliminates the 5′YGR3′ direction.

NAPSS-CS Algorithm. The NAPSS-CS program identifies helices from a sequence by applying NMR constraints to a

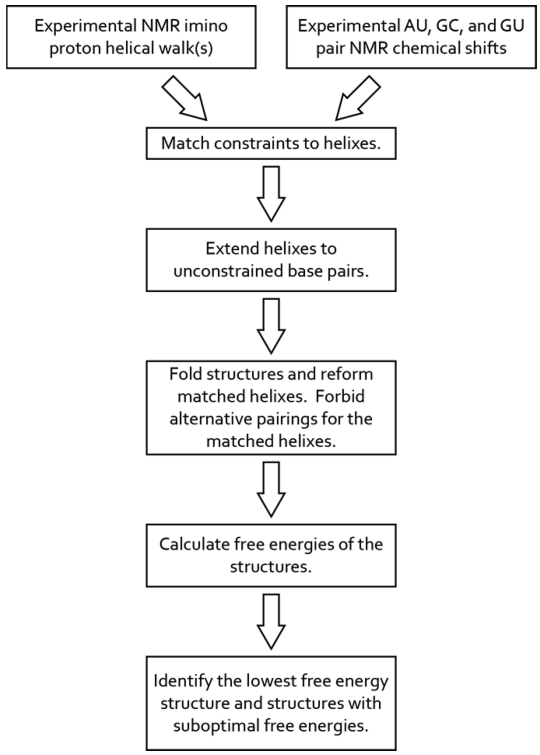


Figure 8. Flow diagram of the NAPSS-CS algorithm.

dynamic programming algorithm (Figure 8). This algorithm generates a free energy dot plot for a given sequence and recursively searches it for helices or sets of helices that fully match helical walk constraints from NMR. As in the previous version,[39] NAPSS-CS considers imino walks across single-nucleotide bulges and also coaxial stacks if the helices are not in separate pseudoknots. In NAPSS-CS, experimental chemical shifts for GH1, UH3, and UH5 in GU pairs, AH2, UH3, and UH5 in AU pairs, and GH1 in GC pairs can be input, which may define the orientation of base pairs in helices (Table 2).

To reduce the search space, NAPSS-CS looks for the dot plot with the minimal number of pairs needed to find at least one complete match to all the helices revealed by NMR. Starting with the dot plot with pairs in the predicted most stable structure based only on thermodynamics, the number of pairs in the dot plot is increased by allowing base pairs found in structures generated at less favorable free energy changes. The allowed free energy difference between the predicted most favorable free energy change and that used as a cutoff for finding new potential base pairs is gradually increased in 1% increments until there is at least one match for all of the helices revealed by NMR.

Chemical exchange with water may prevent detection of imino protons of terminal base pairs and/or NOEs from those protons to imino protons of adjacent base pairs from being identified in NMR spectra and included in imino walk constraints. Therefore, the program extends matched helices to canonical base pairs that can stack on the matched helices. For every extended match set, NAPSS-CS forbids alternative pairings for the nucleotides in the set, uses a dynamic programming algorithm to fold the RNA into a secondary structure by free energy minimization, and then re-forms the base pairs from the match set if they have not already been re-formed by the dynamic programming algorithm. Because nucleotides are not allowed in matched base pairs to base pair

DOI: 10.1021/acs.biochem.5b00833
Biochemistry 2015, 54, 6769−6782

to any nucleotide other than its matched partner, helices that are pseudoknotted once the matched base pairs are added are allowed to form.[67] Similar to the ShapeKnots algorithm,[38] NAPSS-CS fills in single or tandem mismatches and removes isolated pairs. The latter eliminates possible nonsensical pseudoknots. NAPSS-CS then calculates the folding free energy for every structure with the extended matches still intact. The free energy change for nonpseudoknotted secondary structures is calculated with the INN-HB model.[68,69] Free energy changes for structures with pseudoknots are calculated by adding three energy terms: (a) the free energy change of the structure with the pseudoknot removed (maximizing the number of remaining base pairs)[70] using the INN-HB model, (b) the energy of the helix(es) that was removed when breaking the pseudoknot, calculated with the INN-HB model (without intermolecular initiation or any loop free energy changes), and (c) a pseudoknot energy penalty introduced in the ShapeKnots algorithm.[38,71]

If available, additional constraints or restraints can be used to further reduce the folding space. These include restraints from SHAPE and other mapping experiments or pairing constraints that can be derived, for example, from sequence comparison.

Application of NAPSS-CS to HAR1 and *B. mori* R2 Retrotransposon RNAs Illustrates the Method. HAR1 is a rapidly evolving noncoding RNA discovered in the brains of chimpanzees and humans.[8] Two secondary structures were initially proposed for HAR1, but NMR spectra[10] confirmed the secondary structure proposed by Beniaminov et al.[9] (Figure 9). To expand on published NMR spectra, new spectra were measured for HAR1 (Figure 10 and Figures S3–S5). NOESY

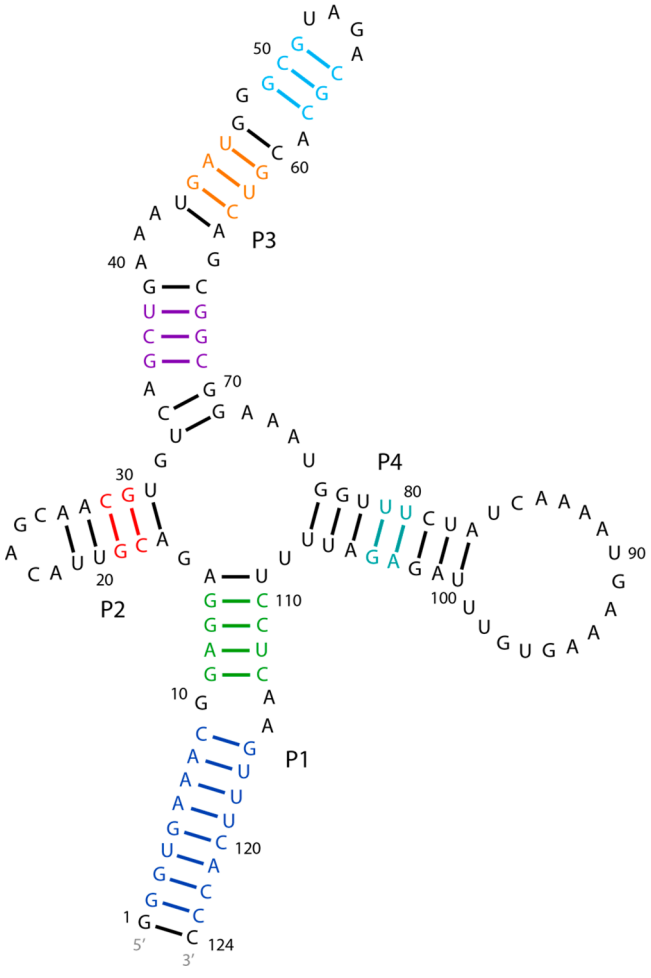


Figure 9. Secondary structure of human HAR1. Colored base pairs correspond to imino proton walks identified from NMR spectra (Table 3). Residue numbering in the structure reported in this work is shifted by three residues relative to the consensus sequence.[9]

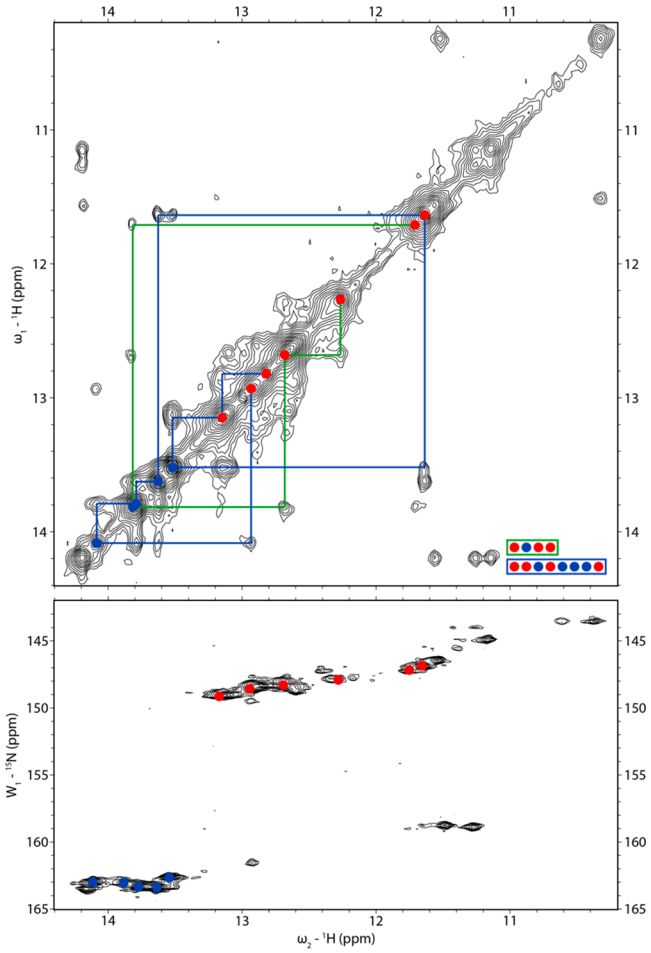


Figure 10. NMR spectra of human HAR1 showing imino proton walks for helix P1 (see Figure 9 for the secondary structure). The top spectrum is a ^{1}H–^{1}H NOESY spectrum acquired at 25 °C with a 125 ms mixing time. Colored lines depict imino proton walks for helix P1 and correspond to base pairs in Figure 9 with the same colors. Colored dots represent base pair type (red for GC and blue for AU). Sequences of dots for imino proton walks are in the bottom right of the spectrum. Blue and green boxes around sequences of dots correspond to blue and green base pairs in Figure 9. The bottom spectrum is a ^{15}N–^{1}H HSQC spectrum acquired at 25 °C.

and ^{15}N–^{1}H HSQC spectra (Figure 10 and Figures S3 and S4) acquired at 25 °C revealed 17 G and 8 U imino peaks, corresponding to 15 GC pairs, 6 AU pairs, and 2 GU pairs. An additional AU and GU pair was identified in a NOESY spectrum acquired at 15 °C (Figure S5). These data are consistent with seven imino proton walks (Figures 9 and 10 and Figures S3–S5). The imino proton chemical shift data from this work (Table S14) agree well with those reported by Ziegeler et al.[10] Imino proton resonances not identified in this work that Ziegeler et al.[10] assigned only in fragments of the full-length construct are for G1, G48, and U109. These resonances were likely difficult to assign in the full-length construct because

DOI: 10.1021/acs.biochem.5b00833
Biochemistry 2015, 54, 6769–6782

Table 3. NMR Constraints Used To Predict the Structure of HAR1[a]

Constraints	Sequence of base pairs predicted by NAPSS-CS
66(13.15 0 0)5(7.335 13.52 0)6(11.64 0 0)5(6.986 13.63 0)5(7.079 13.79 0)5(7.717 14.09 0)6	G2:C123, G3:C122, U4:A121, G5:C120, A6:U119, A7:U118, A8:U117, C9:G116
65(7.353 13.82 0)6(12.68 0 0)6	G11:C113, A12:U112, G13:C111, G14:C110
66	C18:G30, G19:C29
66(13.18 0 0)7	G36:C69, C37:G68, U38:G67
65(7.677 14.20 0)7	G44:C63, A45:U62, U46:G61
66(12.90 0 0)6	G49:C58, C50:G57, G51:C56
75	U79:G104, U80:A103

[a]Colors correspond to those in the secondary structure of Figure 9 and the imino proton walks shown in Figure 10. Chemical shifts are submitted to NAPSS-CS as shown on the left, and base pairs predicted by NAPSS-CS are shown on the right. Integers 5, 6, and 7 represent AU, GC, and GU pairs, respectively. For GU pairs, numbers in parentheses are chemical shifts of GH1, UH3, and UH5, respectively. For AU pairs, numbers in parentheses are chemical shifts of AH2, UH3, and UH5, respectively. For GC pairs, the first number in parentheses is the chemical shift of GH1. For GU and AU pairs, zero means an unavailable chemical shift. For GC pairs, the second two zeros are placeholders for absent chemical shift constraints. Chemical shifts (Table S14) are from ref 10 and this work.

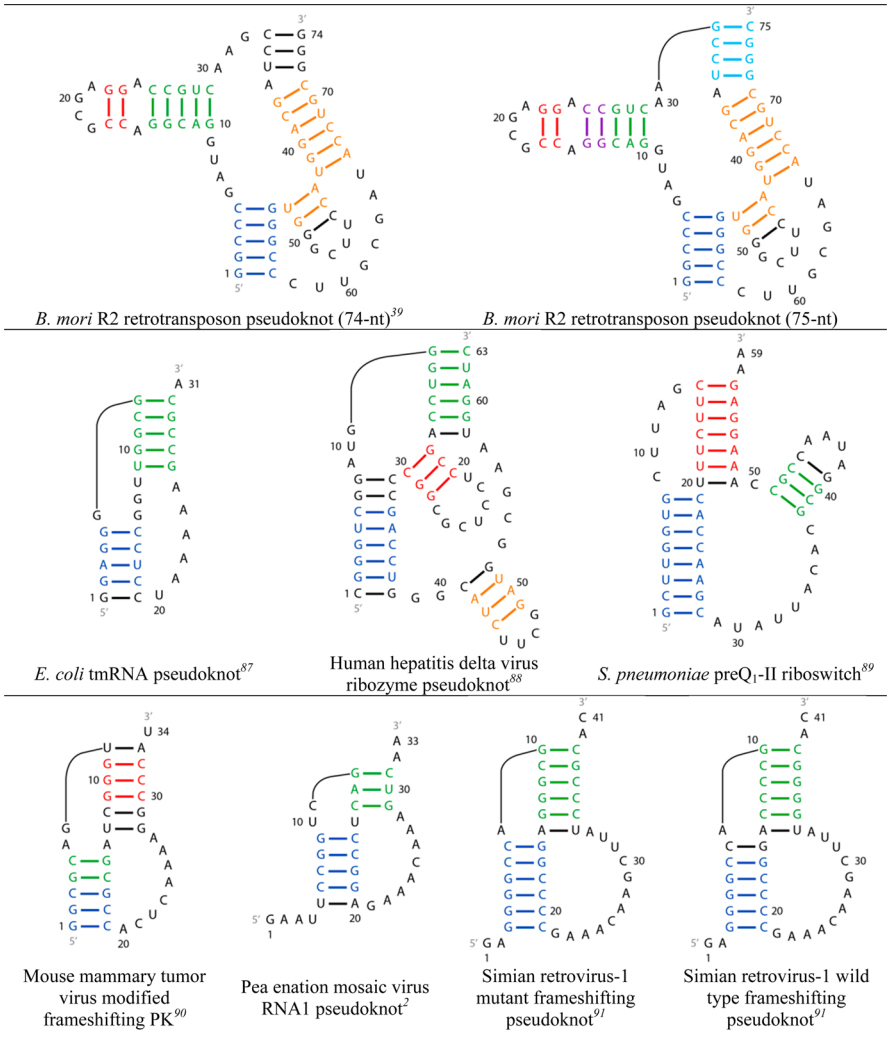


Figure 11. Pseudoknotted structures used to benchmark NAPSS-CS. Colored base pairs correspond to imino proton walks identified from literature NMR spectra (Tables S16–S24).

they exchange with solvent protons in solvent-accessible environments: G1 and U109 are located at the end of helices, and G48 is in a GA pair. Imino proton resonances that Ziegeler et al.[10] did not assign in the full-length construct but were identified in this work are for G19, G30, G44, G49, G51, U79, U80, and G104 (Table S14). On the basis of the secondary structure shown in Figure 9, P2 and P4 were missing expected imino signals and the two pairs assigned to P4 could be observed only at a reduced temperature (15 °C) and a reduced mixing time (60 ms). These observations suggest that P2 and P4 may be unstable and/or dynamic. Constraints used for HAR1 are listed in Table 3.

The 75 nt *B. mori* R2 retrotransposon pseudoknot (Figure 11) has NMR spectra that agree well with those of a 74 nt construct (Figure 11) originally used to identify the same pseudoknot. NOESY and $^{15}N-^{1}H$ HSQC spectra (Figures S6–S9) acquired at 25 °C revealed imino resonances corresponding to 20 GC pairs, 4 AU pairs, and 1 GU pair. These data provided six imino

DOI: 10.1021/acs.biochem.5b00833
Biochemistry 2015, 54, 6769–6782

Table 4. Prediction Accuracies (in percentage of known base pairs)[a]

structure	length (nt)	no. of base pairs in accepted structure	Fold[20]		ProbKnot[72]		ShapeKnots[38]		NAPSS-CS	
			Sens	PPV	Sens	PPV	Sens	PPV	Sens	PPV
Bacillus subtilis pbuE adenine riboswitch aptamer mutant[92]	67	18	100	100	100	78	100	100	100	100
B. subtilis Trp tRNA[93]	74	21	95	100	100	81	95	83	90	86
bovine Trp tRNA[94]	75	21	95	100	100	88	95	100	100	100
human HAR1 MBL	124	41	100	100	95	98	100	100	100	98
influenza A segment 7 MBL[95]	61	19	100	100	100	100	100	100	100	100
Medaka telomerase RNA CR4/5 domain MBL[66]	53	20	85	89	95	90	85	89	90	89
Moloney MLV core encapsidation signal MBL[63]	101	34	100	97	100	94	100	97	100	94
Saccharomyces cerevisiae group II intron *Sc*.ai5γ domain 1 κ–ζ MBL[96]	49	16	88	100	88	78	88	100	100	100
TRSV adenine-dependent hairpin ribozyme[97]	80	23	100	77	100	77	100	77	100	77
B. mori R2 retrotransposon PK (74 nt)[39]	74	24	58	58	42	45	58	58	83	91
B. mori R2 retrotransposon PK (75 nt)	75	25	56	58	40	45	56	58	88	79
E. coli tmRNA PK[87]	31	10	0	0	30	30	100	91	100	91
human HDV ribozyme PK[88]	63	21	19	21	19	20	19	21	76	73
MMTV-modified frameshifting PK[90]	34	11	45	71	45	71	100	92	100	92
PEMV RNA1 PK[2]	33	8	63	83	63	83	63	83	100	80
Streptococcus pneumoniae preQ$_1$-II riboswitch[89]	59	19	63	100	63	100	63	100	100	100
SRV-1 mutant frameshifting PK[91]	41	12	50	75	83	100	100	100	100	100
SRV-1 wild-type frameshifting PK[91]	41	12	0	0	0	0	0	0	100	100
average, all structures			68	74	70	71	79	81	96	92
paired *t* test *p* values			0.001	0.017	0.001	0.003	0.009	0.046		

[a]Only NAPSS-CS was constrained by experimental data. Pseudoknotted structures are denoted by "PK".

proton walks (Figure 11 and Figures S6–S9). One walk is new because G32 is able to pair with C75 in the 75-mer (Figure 11 and Figure S8). Unlike in the spectra of the 74 nt construct,[39] an NOE from G13 to G27 was not observed in the spectrum of the 75 nt construct, possibly because its sample concentration was lower than for the 74 nt construct. This reduced a previous walk of 5 bp to two separate walks of 3 and 2 bp. Constraints used for the 75-mer *B. mori* R2 pseudoknot are provided in Table S17. While not used as a constraint, an imino proton resonance at 12.31 ppm was assigned to G9 in the 75 nt construct, consistent with formation of an imino G9-A30 pair.

Accuracy of Structure Prediction. The NAPSS-CS algorithm was tested on a total of 18 structures, nine of which contain pseudoknots (Figures 9 and 11 and Figure S10). NMR constraints for these sequences are provided in Table 3 and Tables S16–S32. The average sensitivity and PPV were calculated over all 18 structures. A one-tailed, paired *t* test was performed to test the significance of the improvement in sensitivities and PPVs from constrained NAPSS-CS as compared to those from three other algorithms without restraints (Table 4). The null hypothesis (H_0), stating that NAPSS-CS is not more accurate than an alternative program, was tested. Constrained NAPSS-CS was found to have a sensitivity ($p < 0.05$) significantly higher than those of Fold,[20] ProbKnot,[72] and ShapeKnots,[38] and a PPV significantly higher than that of ProbKnot (Table 4).

For structures with pseudoknots, the sensitivity and PPV for pseudoknotted base pairs were calculated separately with eqs 2 and 3. Pseudoknotted base pairs are those that form between single-stranded and loop bases and close the loop involved in formation of those base pairs (see Methods). Pseudoknotted base pairs were predicted considerably better with the constrained NAPSS-CS algorithm than with unrestrained programs (Table 5). The average pseudoknot sensitivity and PPV for NAPSS-CS are 95 and 88%, respectively, compared to ≤33% for each of the other programs when unrestrained. Six of the nine sequences were predicted to be pseudoknotted only by constrained NAPSS-CS.

Both SHAPE[73] and NMR[39] data are only available for the 74 nt *B. mori* R2 retrotransposon pseudoknot. When two strong and two moderate SHAPE hits[73] were used as restraints with reactivities of 0.8 and 0.4, respectively, ShapeKnots predicted a structure with 58% sensitivity and PPV for all base pairs, but with no pseudoknotted base pairs. In contrast, NAPSS-CS with only NMR constraints predicted a structure with a sensitivity and PPV of 83 and 91% of all base pairs (Table 4) and 79 and 85% of pseudoknotted base pairs (Table 5), respectively.

Aside from the 18 structures whose prediction accuracies are reported, NAPSS-CS failed to give results for two additional structures (Figure S11). The *Kluyveromyces lactis* telomerase RNA pseudoknot[74] contains a helical walk across a single base pair (C22-G36), which is not allowed by the dynamic programming algorithm. The MLV recording signal pseudoknot[75] has a helical walk across coaxially stacked pseudoknots, which is not supported by NAPSS-CS.[39] Aside from the two structures, little evidence of walks across these arrangements of base pairs exists in the literature.[39] For that reason, NAPSS-CS is designed not to support these structures.

DISCUSSION

High-throughput sequencing methods reveal many new RNAs with unknown structure.[76−80] Determination of secondary structure usually involves sequence comparison[7,11−13] and/or a

DOI: 10.1021/acs.biochem.5b00833
Biochemistry 2015, 54, 6769–6782

Table 5. Prediction Accuracies (in percentage of pseudoknotted base pairs) for Pseudoknots[a]

			ProbKnot[72]		ShapeKnots[38]		NAPSS-CS	
structure	length (nt)	no. of PK base pairs in accepted structure	PK Sens	PK PPV	PK Sens	PK PPV	PK Sens	PK PPV
B. mori R2 retrotransposon PK (74 nt)[39]	74	14	0	0	0	0	79	85
B. mori R2 retrotransposon PK (75 nt)	75	15	0	0	0	0	87	77
E. coli tmRNA PK[87]	31	10	0	0	100	91	100	91
human HDV ribozyme PK[88]	63	14	0	0	0	0	86	67
MMTV-modified frameshifting PK[90]	34	11	0	0	100	92	100	92
PEMV RNA1 PK[2]	33	8	0	0	0	0	100	80
S. pneumoniae preQ$_1$-II riboswitch[89]	59	15	0	0	0	0	100	100
SRV-1 mutant frameshifting PK[91]	41	12	83	100	100	100	100	100
SRV-1 wild-type frameshifting PK[91]	41	12	0	0	0	0	100	100
average, all structures			9	11	33	31	95	88

[a]Only NAPSS-CS was constrained by experimental data. The pseudoknot sensitivity and PPV were not calculated for structures predicted with Fold12 because it does not allow pseudoknots. The pseudoknot sensitivity and PPV were calculated according to the method of Hadjin et al.[38] If predicted structures have pseudoknots, then the sensitivity and PPV are calculated with eqs 2 and 3 only for base pairs involved in the pseudoknots. If the predicted structure contains no pseudoknot, then the sensitivity and PPV are defined as 0%. For structures predicted with Fold, the pseudoknot sensitivity and PPV are 0% because it does not allow prediction of pseudoknots.

combination of chemical mapping and free energy minimization.[20,36,38,81,82] Chemical mapping limits folding space by revealing nucleotides that are not in Watson-Crick base pairs. In contrast, NMR Assisted Prediction of Secondary Structure limits folding space by revealing nucleotides in canonical base pairs.[39] The results presented here show that adding chemical shifts for imino, UH5, and AH2 protons can further reduce folding space by indicating the 5′ to 3′ direction of base pairs in helices.

Pseudoknots are an important motif because they usually indicate functional significance. Pseudoknots, however, are particularly difficult to prove, partially because of the limited knowledge of the sequence dependence of thermodynamics.[34,35] The results in Figure 12 and Table 5 show that constraints from NMR dramatically improve evidence for and predictions of pseudoknots. Similar improvement has been shown for 13 sequences when ShapeKnots is restrained with SHAPE data.[38]

An NMR and biochemical study of an adenine riboswitch revealed that temperature-independent function required two slowly exchanging secondary structures when ligand was not bound.[6] *In vivo* chemical mapping of *Arabidopsis thaliana* seedlings suggests that stress-expressed RNAs may have multiple structures.[80] Even an RNA duplex of 22 nucleotides has been shown by NMR to have two slowly exchanging secondary structures.[83,84] When secondary structures are slowly exchanging so that imino cross-peaks are observed for the different conformations, the NAPSS-CS algorithm may facilitate determination of individual secondary structures.

Compared to proteins, relatively few 3D structures are known for RNA despite the importance of 3D structure for deducing structure−function relationships. In addition to providing base pairing information, NAPSS-CS provides assignments for many imino proton, UH5, and AH2 resonances for bases in canonical base pairs. Assignment of UH5 also allows rapid assignment of UH6.[60] These assignments then permit assignment of other nonexchangeable resonances and thus for many assignments required to determine 3D structure.

The approach presented here should become more important as the RNA database of NMR spectra, structures, and dynamics grows. For example, additional patterns for chemical shifts and NOE connections may be found and incorporated into the NAPSS-CS algorithm to improve

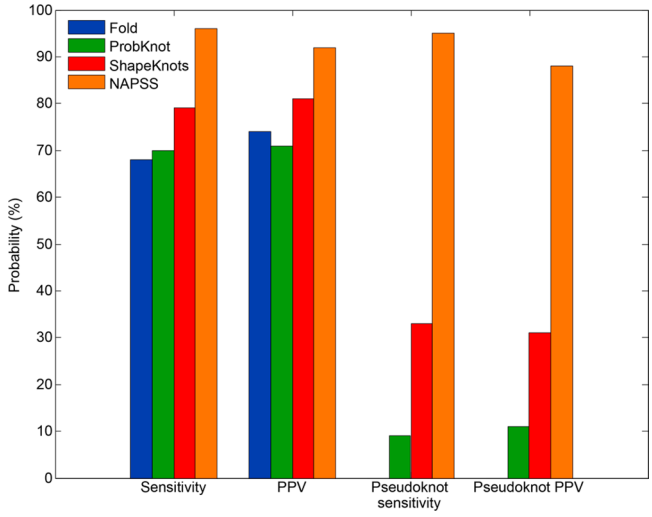


Figure 12. Average accuracy of prediction of RNA secondary structure with NAPSS-CS constrained by NMR data compared with Fold,[20] ProbKnot,[72] and ShapeKnots[38] when the last three are not restrained by experimental data. Accuracy was measured as sensitivity and positive predictive value (PPV). On the left are shown the average sensitivity and PPV for canonical base pairs of structures in a database of nine nonpseudoknotted and nine pseudoknotted structures (Table 4). On the right are shown the average sensitivity and PPV for predicting the presence of a pseudoknot (Table 5). Pseudoknot sensitivity and PPV were not calculated for structures predicted with Fold because it does not allow pseudoknots.

accuracy. Conversely, local environments may produce exceptions to the constraints currently used. A recent apparent exception is the U5 primer binding site in Moloney MLV,[85] where a 5′RGY3′ triplet is flanked by a GU pair on both sides. As more data become available, NMR restraints rather than constraints could be used with chemical mapping data to further improve the overall prediction accuracy.[38] The existing database for pseudoknots, however, is too small to train well the parameters required for an approach using restraints.

Because RNAs are often difficult to crystallize and packing interactions can affect structure,[86] it is likely that NMR will continue to be an important method for determining RNA structure. The approach can also be applied to other

DOI: 10.1021/acs.biochem.5b00833
Biochemistry 2015, 54, 6769−6782

self-assembling polymers that would generate NOESY walks and have known thermodynamics. This includes DNA and a wide variety of nucleic acid mimics, many of which cannot be interrogated by all chemical mapping methods because backbones and/or base equivalents may not be natural.[46−49]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.biochem.5b00833.

(I) Spectral analysis of r(CUGGCUAG)$_2$ and r(AGGCUU)$_2$. (II) 2D imino and NOESY walk region spectra of r(CUGGCUAG)$_2$ and r(AGGCUU)$_2$. (III) NMR spectra of HAR1 depicting imino proton walks for helices P2−P4 and of the 75 nt *B. mori* R2 retrotransposon. (IV) Structures without pseudoknots used to benchmark NAPSS-CS. (V) Structures that NAPSS-CS failed to predict. (VI) Acquisition parameters used in NMR experiments on HAR1. (VII) Chemical shift assignments for duplexes on which NMR spectra were acquired for this work. (VIII) Assigned ^{1}H and ^{15}N chemical shifts for HAR1 and the 75 nt *B. mori* R2 retrotransposon. (IX) NMR constraints used to predict the structures of RNAs used to benchmark NAPSS-CS. (X) Plasmid insert design for *in vitro* transcription of the 75 nt *B. mori* R2 retrotransposon pseudoknot by T7 RNA polymerase. (XI) Accession codes and primary references from which direction-dependent chemical shift constraints were derived. (XII) Accession codes and primary references of PDB structures in which distances between imino protons of adjacent base pairs were measured (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: turner@chem.rochester.edu. Phone: (585) 275-3207. Fax: (585) 276-0205.

Author Contributions

J.L.C. and S.B. contributed equally to this work.

Funding

This work was supported by National Institutes of Health (NIH) Grant GM076485 (D.H.M.) and NIH Grant GM22939 (D.H.T.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the following individuals for their contributions to this work: Prof. Marty Serra for synthesizing r(GUGAAUUUAC)$_2$, Prof. Alain Krol for supplying a plasmid containing the sequence for HAR1, Dr. Jesse Kleingardner and Prof. Kara Bren for assistance with purifying HAR1 transcripts, Min Hur for NMR spectra of r(CGUGAUUACG)$_2$, and Prof. Steven Van Doren for chemical shift assignments of the *Bacillus* RNase P P5.1 hairpin.

■ ABBREVIATIONS

1D, one-dimensional; 2D, two-dimensional; 3D, three-dimensional; BMRB, Biological Magnetic Resonance Data Bank; CR4/5, conserved regions 4 and 5; CS, chemical shift; FPLC, fast protein liquid chromatography; HAR1, human accelerated region 1; HDV, hepatitis delta virus; HSQC, heteronuclear single-quantum coherence; INN-HB, individual nearest neighbor hydrogen bond; MBL, multibranch loop; MLV, murine leukemia virus; MMTV, mouse mammary tumor virus; NAPSS, NMR-assisted prediction of secondary structure; NOE, nuclear Overhauser effect; NOESY, nuclear Overhauser effect spectroscopy; nt, nucleotide; PBS, primer binding site; PDB, Protein Data Bank; PEMV, pea enation mosaic virus; PK, pseudoknot; PPV, positive predictive value; Sens, sensitivity; SHAPE, selective 2′-hydroxyl acylation analyzed by primer extension; SRV, simian retrovirus; TCV, turnip crinkle virus; TOCSY, total correlation spectroscopy; TRSV, tobacco ringspot virus.

■ REFERENCES

(1) Atkins, J. F., Gesteland, R. F., and Cech, T. R. (2011) *RNA Worlds: From Life's Origins to Diversity in Gene Regulation*, Cold Spring Harbor Laboratory Press, Plainview, NY.

(2) Nixon, P. L., Rangan, A., Kim, Y. G., Rich, A., Hoffman, D. W., Hennig, M., and Giedroc, D. P. (2002) Solution structure of a luteoviral P1-P2 frameshifting mRNA pseudoknot. *J. Mol. Biol. 322*, 621−633.

(3) Sripakdeevong, P., Cevec, M., Chang, A. T., Erat, M. C., Ziegeler, M., Zhao, Q., Fox, G. E., Gao, X., Kennedy, S. D., Kierzek, R., Nikonowicz, E. P., Schwalbe, H., Sigel, R. K. O., Turner, D. H., and Das, R. (2014) Structure determination of noncanonical RNA motifs guided by ^{1}H NMR chemical shifts. *Nat. Methods 11*, 413−416.

(4) Keane, S. C., Heng, X., Lu, K., Kharytonchyk, S., Ramakrishnan, V., Carter, G., Barton, S., Hosic, A., Florwick, A., Santos, J., Bolden, N. C., McCowin, S., Case, D. A., Johnson, B. A., Salemi, M., Telesnitsky, A., and Summers, M. F. (2015) Structure of the HIV-1 RNA packaging signal. *Science 348*, 917−921.

(5) Disney, M. D., Yildirim, I., and Childs-Disney, J. L. (2014) Methods to enable the design of bioactive small molecules targeting RNA. *Org. Biomol. Chem. 12*, 1029−1039.

(6) Reining, A., Nozinovic, S., Schlepckow, K., Buhr, F., Furtig, B., and Schwalbe, H. (2013) Three-state mechanism couples ligand and temperature sensing in riboswitches. *Nature 499*, 355−359.

(7) Woese, C. R., and Pace, N. R. (1993) Probing RNA Structure, Function, and History by Comparative Analysis. In *The RNA World*, (Gesteland, R. F., and Atkins, R. F. Eds.) pp 91−117, Cold Spring Harbor Laboratory Press, Plainview, NY.

(8) Pollard, K. S., Salama, S. R., Lambert, N., Lambot, M.-A., Coppens, S., Pedersen, J. S., Katzman, S., King, B., Onodera, C., Siepel, A., Kern, A. D., Dehay, C., Igel, H., Ares, M., Jr., Vanderhaeghen, P., and Haussler, D. (2006) An RNA gene expressed during cortical development evolved rapidly in humans. *Nature 443*, 167−172.

(9) Beniaminov, A., Westhof, E., and Krol, A. (2008) Distinctive structures between chimpanzee and human in a brain noncoding RNA. *RNA 14*, 1270−1275.

(10) Ziegeler, M., Cevec, M., Richter, C., and Schwalbe, H. (2012) NMR studies of HAR1 RNA secondary structures reveal conformational dynamics in the human RNA. *ChemBioChem 13*, 2100−2112.

(11) Woese, C. R., Magrum, L. J., Gupta, R., Siegel, R. B., Stahl, D. A., Kop, J., Crawford, N., Brosius, R., Gutell, R., Hogan, J. J., and Noller, H. F. (1980) Secondary structure model for bacterial 16S ribosomal RNA: Phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res. 8*, 2275−2294.

(12) Noller, H. F., and Woese, C. R. (1981) Secondary structure of 16S ribosomal RNA. *Science 212*, 403−411.

(13) Gutell, R. R., Lee, J. C., and Cannone, J. J. (2002) The accuracy of ribosomal RNA comparative structure models. *Curr. Opin. Struct. Biol. 12*, 301−310.

(14) Tinoco, I., Jr., Uhlenbeck, O. C., and Levine, M. D. (1971) Estimation of secondary structure in ribonucleic acids. *Nature 230*, 362−367.

DOI: 10.1021/acs.biochem.5b00833
Biochemistry 2015, 54, 6769−6782

(15) Pipas, J. M., and McMahon, J. E. (1975) Method for predicting RNA secondary structure. *Proc. Natl. Acad. Sci. U. S. A. 72*, 2017−2021.
(16) Nussinov, R., and Jacobson, A. B. (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. U. S. A. 77*, 6309−6313.
(17) Zuker, M., and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res. 9*, 133−148.
(18) Zuker, M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science 244*, 48−52.
(19) Mathews, D. H., Sabina, J., Zuker, M., and Turner, D. H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol. 288*, 911−940.
(20) Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. U. S. A. 101*, 7287−7292.
(21) Mathews, D. H., and Turner, D. H. (2002) Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol. 317*, 191−203.
(22) Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R., and Stadler, P. F. (2008) RNAalifold: Improved consensus structure prediction for RNA alignments. *BMC Bioinf. 9*, 474.
(23) Bernhart, S. H., and Hofacker, I. L. (2009) From consensus structure prediction to RNA gene finding. *Briefings Funct. Genomics Proteomics 8*, 461−471.
(24) Mathews, D. H., Moss, W. N., and Turner, D. H. (2010) Folding and finding RNA secondary structure. *Cold Spring Harbor Perspect. Biol. 2*, a003665.
(25) Harmanci, A. O., Sharma, G., and Mathews, D. H. (2011) TurboFold: Iterative probabilistic estimation of secondary structures for multiple RNA sequences. *BMC Bioinf. 12*, 108.
(26) Fu, Y., Sharma, G., and Mathews, D. H. (2014) Dynalign II: Common secondary structure prediction for RNA homologs with domain insertions. *Nucleic Acids Res. 42*, 13939−13948.
(27) Brierley, I., Digard, P., and Inglis, S. C. (1989) Characterization of an efficient coronavirus ribosomal frameshifting signal: Requirement for an RNA pseudoknot. *Cell 57*, 537−547.
(28) Tuerk, C., MacDougal, S., and Gold, L. (1992) RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proc. Natl. Acad. Sci. U. S. A. 89*, 6988−6992.
(29) Matsufuji, S., Matsufuji, T., Miyazaki, Y., Murakami, Y., Atkins, J. F., Gesteland, R. F., and Hayashi, S. (1995) Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme. *Cell 80*, 51−60.
(30) Ferre-D'Amare, A. R., Zhou, K., and Doudna, J. A. (1998) Crystal structure of a hepatitis delta virus ribozyme. *Nature 395*, 567−574.
(31) Adams, P. L., Stahley, M. R., Kosek, A. B., Wang, J., and Strobel, S. A. (2004) Crystal structure of a self-splicing group I intron with both exons. *Nature 430*, 45−50.
(32) Ke, A., Zhou, K., Ding, F., Cate, J. H. D., and Doudna, J. A. (2004) A conformational switch controls hepatitis delta virus ribozyme catalysis. *Nature 429*, 201−205.
(33) Klein, D. J., and Ferré-D'Amaré, A. R. (2006) Structural basis of *glmS* ribozyme activation by glucosamine-6-phosphate. *Science 313*, 1752−1756.
(34) Rivas, E., and Eddy, S. R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol. 285*, 2053−2068.
(35) Liu, B., Shankar, N., and Turner, D. H. (2010) Fluorescence competition assay measurements of free energy changes for RNA pseudoknots. *Biochemistry 49*, 623−634.
(36) Deigan, K. E., Li, T. W., Mathews, D. H., and Weeks, K. M. (2009) Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U. S. A. 106*, 97−102.
(37) Cordero, P., Kladwang, W., VanLang, C. C., and Das, R. (2012) Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry 51*, 7037−7039.
(38) Hajdin, C. E., Bellaousov, S., Huggins, W., Leonard, C. W., Mathews, D. H., and Weeks, K. M. (2013) Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. U. S. A. 110*, 5498−5503.
(39) Hart, J. M., Kennedy, S. D., Mathews, D. H., and Turner, D. H. (2008) NMR-assisted prediction of RNA secondary structure: Identification of a probable pseudoknot in the coding region of an R2 retrotransposon. *J. Am. Chem. Soc. 130*, 10233−10239.
(40) Keel, A. Y., Rambo, R. P., Batey, R. T., and Kieft, J. S. (2007) A general strategy to solve the phase problem in RNA crystallography. *Structure 15*, 761−772.
(41) Zhang, J., and Ferré-D'Amaré, A. R. (2014) Dramatic improvement of crystals of large RNAs by cation replacement and dehydration. *Structure 22*, 1363−1371.
(42) Cromsigt, J. A. M. T. C., Hilbers, C. W., and Wijmenga, S. S. (2001) Prediction of proton chemical shifts in RNA − Their use in structure refinement and validation. *J. Biomol. NMR 21*, 11−29.
(43) Barton, S., Heng, X., Johnson, B. A., and Summers, M. F. (2013) Database proton NMR chemical shifts for RNA signal assignment and validation. *J. Biomol. NMR 55*, 33−46.
(44) Brown, J. D., Summers, M. F., and Johnson, B. A. (2015) Prediction of hydrogen and carbon chemical shifts from RNA using database mining and support vector regression. *J. Biomol. NMR 63*, 39−52.
(45) Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Kent Wenger, R., Yao, H., and Markley, J. L. (2007) BioMagResBank. *Nucleic Acids Res. 36*, D402−D408.
(46) Nielsen, P. E. (1999) Peptide nucleic acid. A molecule with two identities. *Acc. Chem. Res. 32*, 624−630.
(47) Kool, E. T. (2002) Replacing the nucleobases in DNA with designer molecules. *Acc. Chem. Res. 35*, 936−943.
(48) Henry, A. A., and Romesberg, F. E. (2003) Beyond A, C, G and T: Augmenting nature's alphabet. *Curr. Opin. Chem. Biol. 7*, 727−733.
(49) Rozners, E., Katkevica, D., Bizdena, E., and Strömberg, R. (2003) Synthesis and properties of RNA analogues having amides as interuridine linkages at selected positions. *J. Am. Chem. Soc. 125*, 12125−12136.
(50) Chen, J. L., Dishler, A. L., Kennedy, S. D., Yildirim, I., Liu, B., Turner, D. H., and Serra, M. J. (2012) Testing the nearest neighbor model for canonical RNA base pairs: Revision of GU parameters. *Biochemistry 51*, 3508−3522.
(51) Piotto, M., Saudek, V., and Sklenář, V. (1992) Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR 2*, 661−665.
(52) Grzesiek, S., and Bax, A. (1993) The importance of not saturating H_2O in protein NMR. Application to sensitivity enhancement and NOE measurements. *J. Am. Chem. Soc. 115*, 12593−12594.
(53) Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR 6*, 277−293.
(54) Goddard, T. D., and Kneller, D. G. (2004) *SPARKY, NMR Assignment and Integration Software*, version 3, University of California, San Francisco.
(55) Cavanagh, J., Fairbrother, W. J., Palmer, A. G. I., and Skelton, N. J. (1996) *Protein NMR Spectroscopy: Principles and Practice*, Academic Press, San Diego.
(56) Easton, L. E., Shibata, Y., and Lukavsky, P. J. (2010) Rapid, nondenaturing RNA purification using weak anion-exchange fast performance liquid chromatography. *RNA 16*, 647−653.
(57) Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res. 28*, 235−242.
(58) Leontis, N. B., and Zirbel, C. L. (2012) Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking.

DOI: 10.1021/acs.biochem.5b00833
Biochemistry 2015, 54, 6769−6782

In *RNA 3D Structure Analysis and Prediction* (Leontis, N. B., and Westhof, E., Eds.) pp 281−298, Springer, Berlin.
(59) Word, J. M., Lovell, S. C., Richardson, J. S., and Richardson, D. C. (1999) Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol. 285*, 1735−1747.
(60) Fürtig, B., Richter, C., Wohnert, J., and Schwalbe, H. (2003) NMR spectroscopy of RNA. *ChemBioChem 4*, 936−962.
(61) Stefl, R., and Allain, F. H. T. (2005) A novel RNA pentaloop fold involved in targeting ADAR2. *RNA 11*, 592−597.
(62) Korth, M. M. T., and Sigel, R. K. O. (2012) Unusually high-affinity Mg^{2+} binding at the AU-rich sequence within the antiterminator hairpin of a Mg^{2+} riboswitch. *Chem. Biodiversity 9*, 2035−2049.
(63) D'Souza, V., Dey, A., Habib, D., and Summers, M. F. (2004) NMR structure of the 101-nucleotide core encapsidation signal of the Moloney murine leukemia virus. *J. Mol. Biol. 337*, 427−442.
(64) Vallurupalli, P., and Moore, P. B. (2003) The solution structure of the loop E region of the 5 S rRNA from spinach chloroplasts. *J. Mol. Biol. 325*, 843−856.
(65) Zuo, X. B., Wang, J. B., Yu, P., Eyler, D., Xu, H., Starich, M. R., Tiede, D. M., Simon, A. E., Kasprzak, W., Schwieters, C. D., Shapiro, B. A., and Wang, Y. X. (2010) Solution structure of the cap-independent translational enhancer and ribosome-binding element in the 3′ UTR of turnip crinkle virus. *Proc. Natl. Acad. Sci. U. S. A. 107*, 1385−1390.
(66) Kim, N.-K., Zhang, Q., and Feigon, J. (2014) Structure and sequence elements of the CR4/5 domain of medaka telomerase RNA important for telomerase function. *Nucleic Acids Res. 42*, 3395−3408.
(67) Ren, J., Rastegari, B., Condon, A., and Hoos, H. H. (2005) HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots. *RNA 11*, 1494−1504.
(68) Xia, T. B., SantaLucia, J., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X. Q., Cox, C., and Turner, D. H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry 37*, 14719−14735.
(69) Turner, D. H., and Mathews, D. H. (2010) NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res. 38*, D280−D282.
(70) Smit, S., Rother, K., Heringa, J., and Knight, R. (2008) From knotted to nested RNA structures: A variety of computational methods for pseudoknot removal. *RNA 14*, 410−416.
(71) Aalberts, D. P., and Nandagopal, N. (2010) A two-length-scale polymer theory for RNA loop free energies and helix stacking. *RNA 16*, 1350−1355.
(72) Bellaousov, S., and Mathews, D. H. (2010) ProbKnot: Fast prediction of RNA secondary structure including pseudoknots. *RNA 16*, 1870−1880.
(73) Kierzek, E. (2009) Binding of short oligonucleotides to RNA: Studies of the binding of common RNA structural motifs to isoenergetic microarrays. *Biochemistry 48*, 11344−11356.
(74) Cash, D. D., Cohen-Zontag, O., Kim, N.-K., Shefer, K., Brown, Y., Ulyanov, N. B., Tzfati, Y., and Feigon, J. (2013) Pyrimidine motif triple helix in the *Kluyveromyces lactis* telomerase RNA pseudoknot is essential for function in vivo. *Proc. Natl. Acad. Sci. U. S. A. 110*, 10970−10975.
(75) Houck-Loomis, B., Durney, M. A., Salguero, C., Shankar, N., Nagle, J. M., Goff, S. P., and D'Souza, V. M. (2011) An equilibrium-dependent retroviral mRNA switch regulates translational recoding. *Nature 480*, 561−564.
(76) Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell 133*, 523−536.
(77) Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science 320*, 1344−1349.
(78) Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J., and Bähler, J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature 453*, 1239−1243.
(79) Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., and Segal, E. (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature 467*, 103−107.
(80) Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C., and Assmann, S. M. (2013) *In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature 505*, 696−700.
(81) Reuter, J. S., and Mathews, D. H. (2010) RNAstructure: Software for RNA secondary structure prediction and analysis. *BMC Bioinf. 11*, 129.
(82) Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol. 6*, 26−26.
(83) Hammond, N. B., Tolbert, B. S., Kierzek, R., Turner, D. H., and Kennedy, S. D. (2010) RNA internal loops with tandem AG pairs: The structure of the 5′GAGU/3′UGAG loop can be dramatically different from others, including 5′AAGU/3′UGAA. *Biochemistry 49*, 5817−5827.
(84) Kennedy, S. D., Kierzek, R., and Turner, D. H. (2012) Novel conformation of an RNA structural switch. *Biochemistry 51*, 9257−9259.
(85) Miller, S. B., Yildiz, F. Z., Lo, J. A., Wang, B., and D'Souza, V. M. (2014) A structure-based mechanism for tRNA and retroviral RNA remodelling during primer annealing. *Nature 515*, 591−595.
(86) Ferré-D'Amaré, A. R., Zhou, K., and Doudna, J. A. (1998) A general module for RNA crystallization. *J. Mol. Biol. 279*, 621−631.
(87) Nameki, N., Chattopadhyay, P., Himeno, H., Muto, A., and Kawai, G. (1999) An NMR and mutational analysis of an RNA pseudoknot of *Escherichia coli* tmRNA involved in trans-translation. *Nucleic Acids Res. 27*, 3667−3675.
(88) Tanaka, Y., Hori, T., Tagaya, M., Sakamoto, T., Kurihara, Y., Katahira, M., and Uesugi, S. (2002) Imino proton NMR analysis of HDV ribozymes: Nested double pseudoknot structure and Mg^{2+} ion-binding site close to the catalytic core in solution. *Nucleic Acids Res. 30*, 766−774.
(89) Kang, M., Eichhorn, C. D., and Feigon, J. (2014) Structural determinants for ligand capture by a class II $preQ_1$ riboswitch. *Proc. Natl. Acad. Sci. U. S. A. 111*, E663−E671.
(90) Shen, L. X., and Tinoco, I., Jr. (1995) The structure of an RNA pseudoknot that causes efficient frameshifting in mouse mammary-tumor virus. *J. Mol. Biol. 247*, 963−978.
(91) Du, Z. H., Holland, J. A., Hansen, M. R., Giedroc, D. P., and Hoffman, D. W. (1997) Base-pairings within the RNA pseudoknot associated with the simian retrovirus-1 gag-pro frameshift site. *J. Mol. Biol. 270*, 464−470.
(92) Delfosse, V., Bouchard, P., Bonneau, E., Dagenais, P., Lemay, J.-F., Lafontaine, D. l. A., and Legault, P. (2010) Riboswitch structure: An internal residue mimicking the purine ligand. *Nucleic Acids Res. 38*, 2057−2068.
(93) Yan, X., Xue, H., Liu, H., Hang, J., Wong, J. T.-F., and Zhu, G. (2000) NMR studies of *Bacillus subtilis* $tRNA^{Trp}$ hyperexpressed in *Escherichia coli*. *J. Biol. Chem. 275*, 6712−6716.
(94) Gong, Q., Guo, Q., Tong, K.-L., Zhu, G., Wong, J. T.-F., and Xue, H. (2002) NMR analysis of bovine $tRNA^{Trp}$. *J. Biol. Chem. 277*, 20694−20701.
(95) Jiang, T., Kennedy, S. D., Moss, W. N., Kierzek, E., and Turner, D. H. (2014) Secondary structure of a conserved domain in an intron of influenza A M1 mRNA. *Biochemistry 53*, 5236−5248.
(96) Donghi, D., Pechlaner, M., Finazzo, C., Knobloch, B., and Sigel, R. K. O. (2013) The structural stabilization of the κ three-way junction by Mg(II) represents the first step in the folding of a group II intron. *Nucleic Acids Res. 41*, 2489−2504.
(97) Buck, J., Li, Y.-L., Richter, C., Vergne, J., Maurel, M.-C., and Schwalbe, H. (2009) NMR spectroscopic characterization of the adenine-dependent hairpin ribozyme. *ChemBioChem 10*, 2100−2110.

DOI: 10.1021/acs.biochem.5b00833
Biochemistry 2015, 54, 6769−6782